Using Penalized Linear Discriminant Analysis and Normalized Difference Indices Derived from Landsat 8 Images to Classify Fruit-tree Crops in the Aconcagua Valley, Chile

by

Renfang Liao

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Master of Science
in
Geography

Waterloo, Ontario, Canada, 2016

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Accurate crop type maps are critical for yield estimation and agricultural practices in modern agriculture. A new approach is proposed in this thesis to improve the crop type classification accuracy, by creating a new feature set containing new spectral indices in addition to basic bands. Two types of penalized linear discriminant analysis classifiers are adopted to do the classification, and the cross-validated classification accuracies on the two different feature sets are compared to see whether the new feature set can improve the crop identification. The result shows with new indices in the feature set the classification mean error rates were decreased substantially for both classifiers (21.6% and 25.2%). Through analyzing the coefficients retrieved from the best model, the variable importance was assessed. The coefficients are summarized by different bands and images, and the result suggest that red and shortwave infrared are the two bands highly related to the fruit-trees type identification in the study area in Aconcagua valley, Chile. Also late winter to early spring may be the best time to do crop type mapping for these crop types.

Acknowledgement

I would like to thank my thesis advisor Prof. Alexander Brenning for his valuable academic guidance and support. Thank you for inspiring me and giving me directions when I got lost. This thesis would not have been possible without your help.

I would like to thank Marco Peña who was the first to bring up the idea and was involved in the data collection work for this research. Thank you for your great collaboration in this project.

I would like to thank Kralisch Sven in University of Jena who provided technical support in the computationally intensive tasks. Thank you for being supportive and patient when helping me to solve the technical problems.

Also, I thank the readers of my thesis, Prof. Jonathan Li, Prof. Richard Kelly, and Prof. Ellswoth LeDrew for positively contributing to an excellent learning environment in the Department of Geography and Environment Management, University of Waterloo, where I have been spending a wonderful time in the past two years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 INTRODUCTION

1.1 Motivation of the research

This study is expected to find a better way to classify different fruit trees to increase the classification accuracy. The idea is to explore the possibility of improving the accuracy by creating new spectral indices. These indices are generated from various bands not limited to vegetation related bands; thus, new indices may be identified in this process. Consequently, the research findings will contribute to general crop type classification and the comparison on these classifiers will provide a reference for future studies.

1.2 Overview

Information of regional-scale crop type distribution is critical for agriculture management, crop inventory and crop monitoring. Satellite remote sensing data have therefore been widely applied for the crop classification or crop type mapping (Mkhabela, Bullock, Raj, Wang, & Yang, 2011; Tatsumi, Yamashiki, Canales Torres, & Taipe, 2015). Crop-type classifications from satellite imagery are mostly based on the analysis of the distinguishing spectral signatures. However, different crop types may have very similar spectral responses during some stages in their growing season, especially when the spectral resolution is coarse (Esch, Metz, Marconcini, & Keil, 2014; Peña & Brenning, 2015). Due to different timing of crop growth stages, satellite images from multiple acquisition dates can be used to detect phenological behavior of different crops, thus these satellite images time series (SITS) can facilitate crop-type discrimination (Odenweller & Johnson, 1982; Schuster, Schmidt, Conrad, Kleinschmit, & Förster, 2015; Zhong, Hawkins, Biging, & Gong, 2011). In this research, a temporal sequence of Landsat 8 images acquired across the entire growing season of the crops of interest is therefore used in order to achieve improved classification accuracies.

Based on previous research, full-band satellite images times series (SITS) temporal profiles have more accurate output compared to NDVI temporal profiles (Peña & Brenning, 2015). In this study, an enhanced feature set including the full-band SITS and all possible normalized difference indices derived from any two different bands in the image stack is considered. It is hypothesized that the classification accuracy can be further improved with these new indices, which are not only from the typical vegetation related bands, i.e. green band and near infrared band, but also from all the available bands that could contribute to crop type

discrimination. For one thing, the bands related to the most widely used vegetation index NDVI seem not to be the most important bands in the previous crop classification project (Peña & Brenning, 2015). For another, the potential of indices derived from other bands and from different dates can be explored. In order to show the contribution of the new indices in crop classification, a feature set consisting full-bands SITS only (six bands per image with nine images) is considered as the baseline method to make sure feature set is the only difference.

As for the available classifiers, linear discriminant analysis (LDA) was found to be the best-performing classifier in previous research in a similar setting compared to Random Forest (RF), Support Vector Machine (SVM) and Artificial Neural Network (ANN), which are novel machine learning algorithms (Peña & Brenning, 2015). Besides, LDA is less time consuming than these computationally more complex classifiers. In this study, LDA is therefore chosen as the first classifier to be applied on both feature sets. However, LDA may not be able to deal with collinear high-dimensional data, therefore the new and enhanced feature set cannot be classified by ordinary LDA. LDA based classifiers such as penalized LDA, sparse LDA, regularized LDA are gaining increasing attention (Bandos, Bruzzone, & Camps-Valls, 2009a; Clemmensen, Hastie, Witten, & Ersbøll, 2011; Li, Zhu, & Ogihara, 2006; Merchante, Grandvalet, & Govaert, 2012; Z. Zhang, Dai, Xu, & Jordan, 2010). Hence, two types of penalized LDA are tested to solve the problem. The performance of these two modified versions of LDA are also compared by applying on same feature set. In addition, in order to reduce the uncertainty in result assessment brought by training set sampling, a spatial cross validation process is adopted in this research for classifier validation. Different from non-spatial cross validation, the spatial one

tries to account for the impact of the spatial autocorrelation in the spatial dataset (Brenning, 2012).

1.3 Goal and objectives

The purpose of this study is to investigate the potential of indices in improving the crop type classification accuracy of linear discriminant analysis. These indices are created, as normalized differences of every two different band-variables in the feature set, including both within-image and between-image (or within-date and between-date) normalized band-ratios. These new features may detect the change of bands ratios in different crop phenology states and contribute to the discrimination of different crops. As a use case for the proposed methodology, fruit-tree crop classification in the Aconcagua valley of central Chile is investigated in this study.

The goal of this research is to use LDA based classifiers to find out whether the new index features can improve the classification result or not. The steps to achieve the objective of this study are

- to create new indices from the preprocessed spectral bands
- to apply statistical classifiers to discriminate six crop types
- to compare the cross-validation based model estimation results

1.4 Structure of thesis

This study introduces a new method that aims at improving crop type classification accuracy by creating various spectral band-based indices, and it therefore focuses on the performance comparison with feature sets previously used in crop classification from time series of optimal satellite imagery. This thesis first presents a brief introduction to remote sensing applications in agriculture and a context for this study including important remote sensing based crop classification methods and main feature selection methods to address the research gaps (Chapter 2); then it gives a description of data and study area (Chapter 3); provides an explanation of the classifiers chosen in this study, and the spatial cross-validation method applied to assess the results for later comparison (Chapter 4); an illustration of the results of different feature sets and classifiers, and the interpretation of the results (Chapter 5); a discussion of the results, sources of uncertainty, implication and future directions of this study (Chapter 6); and the main conclusions (Chapter 7).

CHAPTER 2 RESEARCH CONTEXT

2.1 Role of remote sensing in modern agriculture

Remote sensing technology was first applied for agricultural purpose in the 1970s, and since then the technique has been advancing with the changing needs of modern society (Atzberger, 2013). The worldwide growing population and increasing consumption of protein require more agriculture products than ever before, and this trend is expected to last for the next three to four decades (Foley et al., 2011). Meanwhile, environmental issues have drawn more attention than before (Foley et al., 2011). Farming is therefore also required to increase the production while minimizing the pressure on environment. Thus, timely and accurate monitor on crops on a large space and temporal scale is preferred (Craig & Atkinson, 2013; FAO, 2011).

Characterized as being able to rapidly acquire information over large areas with high revisit frequency compared to traditional field survey, remote sensing is very suitable to improve agriculture practices (Atzberger, 2013; Foley et al., 2011). Remote sensing, GIS and GPS are tools that can help farmers to understand the health condition and the phenological stages of their crops, leading to efficiency gains in fertilizer application and irrigation. In the comprehensive review about remote sensing of agriculture by Atzberger (2013), five major applications of remote sensing techniques are chosen to show the importance of using the derived information for decision making: (1) biomass and yield estimation, (2) vegetation vigor and drought stress monitoring, (3) assessment of crop phenological development, (4) crop acreage estimation and cropland mapping, and (5) mapping of disturbances and land use/land

cover (LULC) changes. It is noteworthy that these applications need crop type classification or crop species mapping.

2.2 Crop type classification

Production prediction or yield estimation is of great interest for not only farmers but commodity brokers and public agriculture sectors, for it is highly related to price control and worldwide trading (Goswami, Aruna, & Bairagi, 2012). General production estimation approaches can be summarized as the product of two components: the harvested areas of a given crop and the yield per unit area. Thus, accurate crop-type maps are among the most important datasets, for they are used to calculate the cropping areas (Craig & Astkinson, 2013). The "propagation of error" theory agrees that the minor errors or uncertainties in the processing of raw data, will be amplified in the final extracted information (Heuvelink, Burrough, & Stein, 1989). In order to avoid misguided decisions or biased policies, which may cause loss in trade and serious food security issues, having accurate production prediction is necessary, which itself requires accurate crop type maps.

The spectral classification or identification of crops can be more difficult than mapping land use/land cover, for the status of crops are affected by many factors and showing strong seasonal patterns. In some stages in the life circle of vegetation in one area, their spectral signatures can be very similar and hard to identify, and this can be solved by acquiring the spectral-temporal profile through the crop life circle (Esch et al., 2014; Odenweller & Johnson, 1982; Ozdogan, 2010; Zheng, Myint, Thenkabail, & Aggarwal, 2015). However, it is mostly graminaceous crops that are investigated in these studies, and the graminaceous plants,

cosmopolitan herbaceous or woody plants with hollow jointed stems and long narrow leaves, usually have annual growth cycles. In addition, these crops may not be planted at the same time and have characteristic phenology stages, such as green-up, maturity and senescence which drives the bulk spectral behavior of a crop (Peña & Brenning, 2015; Zhang et al., 2003). The green-up, maturity and senescence stages are referred to the onset of photosynthetic activity, the peak of green leaf area, and the fast decreasing of photosynthetic activity, and the distinctive bulk spectral signatures can facilitate the identification of the crop. In addition, rotations or multiple cropping activities are common for graminaceous plants, and harvesting and regrowth will exhibit in the spectral behavior clearly which also help crop classification.

Compared to the graminaceous plants, the classification problem for fruit-tree crop types are less addressed (Simonneaux et al., 2008; Zhong et al., 2011). The fruit-trees are generally perennial summer crops, which means they can live for more than one year and relatively similar phenology timing, for example growing and blooming in spring and summer; thus, the discrimination between different crop types is not easy as graminaceous plants, especially when it is based simply on several NDVI images (Zhong et al., 2011). The data type for crop classification will be further discussed in the next section. This study is focused on classification problem of perennial crops, in particular fruit trees, and is expected to provide a different solution to fill the research gap.

2.2.1 Data for crop classification

Remote sensing data can be categorized in many different ways, such as by platform type (airborne vs satellite-based images), by sensor type (i.e. RADAR, LiDAR, imagery sensor), by dimensionality (hyperspectral vs multispectral images). Satellite multispectral images are the most widely used data type for crop classification or identification (Araya, Ostendorf, Lyle, & Lewis, 2013; Brenning, Kaden, & Itzerott, 2006; Chen, Son, Chang, & Chen, 2011; Dhumal, Kale, & Mehrotra, 2013; Odenweller & Johnson, 1982; Ozdogan, 2010; Peña & Brenning, 2015; Rodriguez-Galiano, Ghimire, Rogan, Chica-Olmo, & Rigol-Sanchez, 2012; Tatsumi et al., 2015; Zheng et al., 2015; Zhong et al., 2011), and they are the focus of this research.

There are also studies using high resolution optical satellite data such as SPOT, and hyperspectral remote sensed data such as AVIRIS for vegetation or crop classification. The high spatial resolution images have advantage in crop field segmentation or retrieving land parcels without using ancillary data, thus they can facilitate the crop mapping of small fields (Conrad, Fritsch, Zeidler, Rücker, & Dech, 2010; Zhong et al., 2011). Hyperspectral data have many narrow contiguous spectral bands, and it is usually used for classification purpose after useful bands being extracted with sophisticated feature selection methods (Archibald & Fann, 2007). However, hyperspectral data contain fine details about crops, which is of great help when discriminating crops with very similar spectral signatures (Dhumal et al., 2013).

The most widely used multispectral satellite data products for crop classification can be further grouped into two classes: full-band images and vegetation indices. Full bands satellite images can be used to retrieved vegetation indices like normalized difference vegetation index (NDVI) and enhanced vegetation index (EVI), and a subset of the bands can be used for

multiple-crop type classification as well. They each have advantages and disadvantages in data acquisition and processing stages.

Moderate Resolution Imaging Spectroradiometer (MODIS) and Landsat series satellite sensors are the two most common data source for vegetation monitoring purpose, and their features in revisit frequency, spatial resolution and free availability have impacts on the choice of researchers. Retrieved from coarse spatial resolution satellite images, vegetation indices as phenological indicators have been used in many studies to get distinctive features from different crops. For one reason, this type of coarse resolution data can cover a large area, so the research may be done on states or even continental level. For another, research about crop classifications or crop phenology monitoring based on NDVI temporal profiles has been greatly encouraged by the free availability of MODIS 16- and 32-day NDVI image composites, as well as 8-day surface reflectance products, which can be used to make 8-day NDVI composites (Araya et al., 2013; Chen et al., 2011; Zhong et al., 2011).

These freely available MODIS NDVI products reduce the preprocessing work such as geometric and radiometric correction for users (Peña & Brenning, 2015). The platforms with shorter revisiting frequency are likely to produce more cloud-free images, which is very important for satellite image analysis for vegetation, especially in short growing seasons (Chen et al., 2011; Fritz, Massart, Savin, & Leo, 2006; Mkhabela et al., 2011). This is a strong point of using MODIS data for crop classification or monitor; however, the high temporal resolution is traded for a poorer spatial resolution. The spatial resolution of MODIS is from 250m to 1000m, which may be too big to study and monitor on small fields level, but it is very useful for

continental or regional scale crop monitoring which requires shorter revisiting time (Conrad et al., 2010; Fritz et al., 2006).

The Landsat program is the longest running satellite imagery system for the earth observation ever since 1972, and the archived full bands stack images, including newest Landsat-8 series, are free to download from USGS website. Compared to MODIS images, Landsat series have a lower revisit frequency at about every two weeks, so there probably are fewer cloud-free images available. Nevertheless, Landsat's spatial resolution is 30 m × 30 m, much higher than that of MODIS. This spatial resolution is fine enough for field level crop monitoring and reduces the classification error caused by mixed pixels. NDVI and other vegetation indices can be generated from the full-band images too, but advanced geo- and radiometric correction and index calculation all need to be done by users.

The NDVI can represent the physiological and structural condition of vegetation, for it is constructed from a red band (R) and a near infrared band (NIR) as (NIR − R) / (NIR + R). Generally, healthy vegetation will absorb most of the visible light that falls on it, and reflects a large portion of the near-infrared light. The pigment in plant leaves, chlorophyll, strongly absorbs visible light (from 0.4 to 0.7 μm) for use in photosynthesis. The cell structure of the leaves, on the other hand, strongly reflects near-infrared light (from 0.7 to 1.1 μm). However, according to Tatsumi et al. (2015), single or very limited spectral signatures are not as good as multi-date, high-resolution and multi-spectral bands at spectral discrimination for crop types. It is found that NDVI temporal profile and the related red and NIR bands are not the most significant to classify fruit trees, and full-band SITS is another choice for crop classification and spectral bands exploration (Peña & Brenning, 2015; Tatsumi et al., 2015). The NDVI can reflect

the phenological change of vegetation to a certain extent, but it only contains relationship of the red and near infrared band without any information from other spectral range. The full-band image series have more abundant spectral bands than NDVI temporal profile. Also, the implication of the work by Peña & Brenning (2015) suggested that other spectral bands deserve further attention, for the classification accuracy of full-band feature set was significant higher than that of the NDVI.

Hence, for the crop-type mapping on field level, the high spatial resolution Landsat full-band SITS is more suitable than the MODIS NDVI temporal profiles. In this study, all the spectral bands are treated equally and normalized difference indices are derived from every two bands, thus by assessing the importance of variables the other bands' potential can be explored.

2.2.2 Classifiers for crop classification

There are multiple ways to categorize classifiers, for example by classification logic: parametric or nonparametric; by partition logic: hard or fuzzy; by classification algorithm: supervised or unsupervised versus hybrid involving artificial intelligence, and by image segmentation: pixel-based or object-oriented (Jensen, 2005). The important difference between supervised and unsupervised classification is that in an unsupervised classification, the classes are not predefined, while in supervised classification, classifiers learn from training data with predefined classes and are then applied to test data (Fielding, 2007b).

Unsupervised classification also called clustering, such as K-means and ISODATA, is an effective way to partition feature space into sub-space, and does not require as much knowledge about the data as supervised classification (Jensen, 2005). Clustering is more often used for data exploration and sub-structure finding in knowledge-poor environments. However, natural grouping may not solve classification problem with high spectral complexity, for the spectral clusters are likely to represent mixed surface types (Jensen, 2005).

Supervised classification or pattern recognition is the main type of classifiers in LULC and crop classification, for it is generally more accurate than unsupervised classification (Rozenstein & Karnieli, 2011). The supervised classification process contains training and testing steps. In short, from provided training data, which contains predictors and corresponding class labels, classifiers seek patterns and fit models, then they make predictions based on the fitted models on test datasets for an independent assessment. The class labels are usually obtained from field surveys or the interpretation of higher-resolution images. After prediction result

being assessed, if the accuracy is acceptable, the model can be used for the rest of the remotely-sensed data (Jensen, 2005).

Supervised machine learning is a form of artificial intelligence (AI) using many statistical models and algorithms to find the relationship between predictors and class labels from training data. The word "learning" in this context refers to the process that the classification error is gradually reduced through each iteration that the classifier is "learning" from each observation case (Fielding, 2007b). In the well-known text book Jensen's *"Introductory Digital Image Processing: A remote Sensing Perspective"* (2005), various classification methods are introduced systematically; however, the book is ten years from now and it does not cover state-of-art machine learning classifiers. For example, Random Forest (RF) and Support Vector Machine (SVM) are widely investigated in LULC classification including crop type classification in recent research (Archibald & Fann, 2007; Brenning et al., 2006; Mathur & Foody, 2008; Peña & Brenning, 2015; Rodriguez-Galiano, Ghimire, et al., 2012; Tatsumi et al., 2015; Wieland & Pittore, 2014; Zheng et al., 2015).

The basic idea of SVM is to maximize the margin between the two classes in the training data through determining the separating hyperplane, or decision surface as cited in (Archibald & Fann, 2007). According to Mathur and Foody (2008), SVM can derive very accurate classification with accuracy higher than 90% on different sizes of training sets. Brenning et al. (2006) compared SVM with five statistical classifiers including stabilized linear discriminant analysis (SLDA), classification trees (bagging and double-bagging), k-Nearest Neighbours (1-NN) and Logistic regression, and SVM produced overall better results than the other classifiers,

while SLDA was the best on field level classification. However, the training process of SVM is very slow.

Random forest (RF), a machine learning algorithm, is a combination of decision tree predictors providing another option for land cover / land use classification, and it is known for high efficiency and considerable accuracy; furthermore, it can estimate variable importance and is insensitive to outliers (Breiman, 2001; Gislason, Benediktsson, & Sveinsson, 2006; Rodriguez-Galiano, Chica-Olmo, Abarca-Hernandez, Atkinson, & Jeganathan, 2012; Tatsumi et al., 2015). Compared to SVM, RF is much faster in training, but may not have result as accurate as SVM does. In the study by Gislason et al., (2006), RF had the second best result compared to other tree based classifiers such as CART bagging and boosting. According to Rodriguez-Galiano, Chica-Olmo, et al., (2012), the RF had various results with different texture-variable window sizes.

Peña & Brenning (2015) compared three classifiers SVM, RF and linear discriminant analysis (LDA) on different size of training sets and different feature sets in a study area adjacent to the present one. LDA, a conventional statistical classifier, produced the best classification result followed by SVM, while RF was relatively less accurate. The study area and crop types in this study are similar to those of the study done by Peña & Brenning (2015), in which LDA performed the best. Therefore, LDA as a simple but powerful classifier, is investigated in detail to futher explore its potential in crop type classificatioin in this study.

2.3 Feature selection methods

According to the definition given by Guyon & Elisseeff (2003), feature selection or variable selection is a data mining process which minimizes the data dimensionality while achieving very good classification performance for a given set of classifiers by using part of the original variables. However, feature selection, in a broader context, can be seen as a type of method which can modify the feature space, trading off between the classification accuracy and problems cause by the high dimensionality. High dimensional data may contain more useful information, but it also brings trouble for data analysis. First, more features do not guarantee better classification result, while instead they may sometimes contain much noise and lead to randomness in discrimination (Fan & Lv, 2010; Guyon & Elisseeff, 2003; James, Witten, Hastie, & Tibshirani, 2013). Second, over-fitting, model misidentification and difficulties in variable assessment are possible problems brought by collinearity as a result of high dimensionality (Fan & Lv, 2010). Third, reducing the data dimensionality can decrease the complexity of computing to save time and computing power.

Subset selection, shrinkage and dimension reduction are the three important classes of methods that are widely chosen to exclude irrelevant variables from models (Fan & Lv, 2010). The word “selection” may be a bit ambiguous here because feature selection does not necessarily mean the number of features is reduced. For example, depending on the type of performed shrinkage, some of the coefficients may be shrunk to exactly zero, so shrinkage methods can, in some instances, also select out variables based on coefficients. Hence, shrinkage can be seen as a feature selection method in a general way. Besides, feature selection also includes projections of the variables and creating new variables based on the

combination of old variables e.g. PCA though in some context it is not viewed as a feature selection method. In this study penalized linear discriminant analysis, a combination of dimension reduction and shrinkage method, is examined, so these two feature selection methods are explained in more detail in Chapter 4.

2.3.1 Subset

In a previous project, whose study area and data are similar with this one, a stepwise selection approach of eight Landsat images was implemented to find the optimal balance between accuracy and dimensionality. The result shows that with only a few images involved in classification, the result may be as good as involving the full set of images (Peña & Brenning, 2015). In that project, feature selection was done on temporal domain manually through a wrapper approach (introduced latter in this Chapter), and then the classification result of the feature subset was evaluated. In this study, Penalized LDA can simultaneously select predictors and estimate the change of mean classification error rate, which is more efficient than doing wrapper based subset feature selection. Though it is not applied in this study, subset selection is an important method in the feature selection family.

By objective functions and by subsets search strategies are the two major ways to categorize the "subset" feature selection methods (Guyon & Elisseeff, 2003). Objective functions are the performance evaluation criteria for different subsets of variables, by which the subsets selection methods can be classified as filters and wrappers. The only difference between filters and wrapper is the way they evaluate the feature subsets. The wrapper approach uses the classification algorithm itself to assess the usefulness of a feature subset, so the selected features is dependent on the classification algorithm (Guyon & Elisseeff, 2003).

Subset search strategies include exhaustive search, sequential forward selection, sequential backward selection, bidirectional search and floating search, and for different data the best strategy may not be the same. The exhaustive search can provide the global optimal subset, but in many situations the dimensionality is too high to do exhaustive search, which tests every combination of the features.

The difference between filters and wrappers is the objective to select features. The filter approach based feature selections evaluate the variables based on some scoring function, such as Fisher criterion expressed as the ratio of inter-class variance to inner-class variance, thus a filter is independent of the classification algorithm (Karegowda, Manjunath, & Jayaram, 2010). Conversely, the wrapper approach uses the classification algorithm itself to assess the usefulness of a feature subset, so the selected features is dependent on the classification algorithm (Guyon & Elisseeff, 2003; Kohavi & Kohavi, 1997). With the wrapper approach, the machine learning algorithm can be viewed as a black box since the feature selection is only based on the results from classification. Generally, the filter approach is faster than the wrapper approach, while the latter usually produces better results (Guyon & Elisseeff, 2003; Karegowda et al., 2010). Though wrapper approach is time consuming when the dimensionality of data is high, it would still be a good choice if the computing complexity and time is affordable. Furthermore, with efficient search strategies wrappers can also be less time-consuming.

2.3.2 Shrinkage

By adding a regularization term to the loss function of linear models, the prediction accuracy and model interpretability may be improved. Different from the feature subset method, regularization or shrinkage retains all the predictors or features. This method constrains or regularizes the coefficients or weights of the features, reducing them towards zero. The reason for the improvement brought by the constraint on coefficients may not be immediately obvious, but it is proven that the shrinkage can significantly reduce the variance (Fan & Lv, 2010; Guyon & Elisseeff, 2003; James et al., 2013). According to James et al. (2013) ridge and the lasso are two most popular techniques to shrink the coefficients in regression, and they are chosen as the two feature selection methods in this study. These two feature selection methods are introduced in Section 4.3.2 in detail.

2.3.3 Dimension reduction

The dimension reduction is a family of approaches that includes projecting the original high dimensional data into a lower dimensional feature space (James et al., 2013). In this process, a small number of different linear combinations of the variables are chosen as new predictors for further analysis like regression or classification. The well-known Principal Component Analysis (PCA) is a representative of dimension reduction. Usually the new variables are required to be linearly uncorrelated with each other. In PCA, the new combinations of variables are calculated through orthogonal transformation on the covariance matrix. Instead of keeping all the projections of variables, it is more common to select a few combinations that can explain most of the variance in the old data (James et al., 2013). In Chapter 4.3.1 the introduction to Linear Discriminant Analysis (LDA), it is easy to get the conclusion that LDA its self is a dimension

reduction method. Whereas the orthogonal transformation is done on the between-classes scatter matrix instead of the overall variance, so that the combinations that best distinguish different classes can be selected out.

2.4 Chapter Summary

This chapter presents a literature review about the crop classification topic, and from three perspectives, data types, classifiers and feature selection methods. NDVI temporal profiles and full-band images are compared in a crop classification context. The three major feature selection methods subset, shrinkage and dimension reduction are generally introduced, for there are more than a thousand of features in the dataset.

The research gaps are identified including the fruit-tree classification compared to common graminaceous plants, and the potential of LDA in crop-type classification. The fruit-tree classification problem is less studies than staple food crops like wheat and rice, but it is also important for fruit industry which can contribute to local economy in some areas. The fruit-trees are also supposed to be more difficult to identify than graminaceous plants.

CHAPTER 3 STUDY AREA AND DATA

The data acquisition and preprocessing work, for both remote sensed and ground surveyed data, were performed by Peña in 2015. The research of the study area is also credited to him.

3.1 Introduction of the study area

The study area of this project is located in the Andean foothills of Aconcagua River basin, Fifth Region of Valparaiso, Chile. Aconcagua valley (7,340 km^2) has a large agricultural land and it is one of the biggest valleys in the country. In that area, Mediterranean climate, featured as hot dry summer and cool wet winter is the major climate type in the study area. In spring, large amount of snowmelt from the mountains flows into the Andean headwaters, while in winter it is precipitation that control the local hydrological regime. Agricultural land use takes the most area in the gentle terrains of this valley (862.37 km^2), with fruit-tree as the main crop and table grape, avocado, peach and walnut as the main cultivated species (INE, Instituto Nacional de Estadísticas, 2008).

The study area is not equal to the whole agriculture land use area in the Aconcagua valley. It is distributed in four counties: Petorca, Calle Larga, Los Andes and Zapallar, which together cover an area of 576 km^2, mostly cultivated with six tree-fruit species: table grape (*Vitis vinifera* L.), walnut (*Juglans regia* L.), peach (*Prunus persica* L.), avocado (*Persea americana* Mill.), nectarine (*Prunus persica* var. nectarine), mandarine (*Citrus reticulata* Blanco) (Figure 3-1 and Table 3-1). These crops are conformed by relatively small fields (*N* = 5,233; mean size 38,390m^2), which are heterogeneously arranged between the cities of Los Andes and San Felipe and around the Aconcagua River (Figure 3-1). The growing season of these species

generally starts from late September (green-up stage at southern spring) to early April (senescence stage at southern autumn), showing some variations in their inter- and intra-specific onset and offset dates according to the species variety or the specific management practices applied on the crop (Peña, written communication, January 28, 2016).

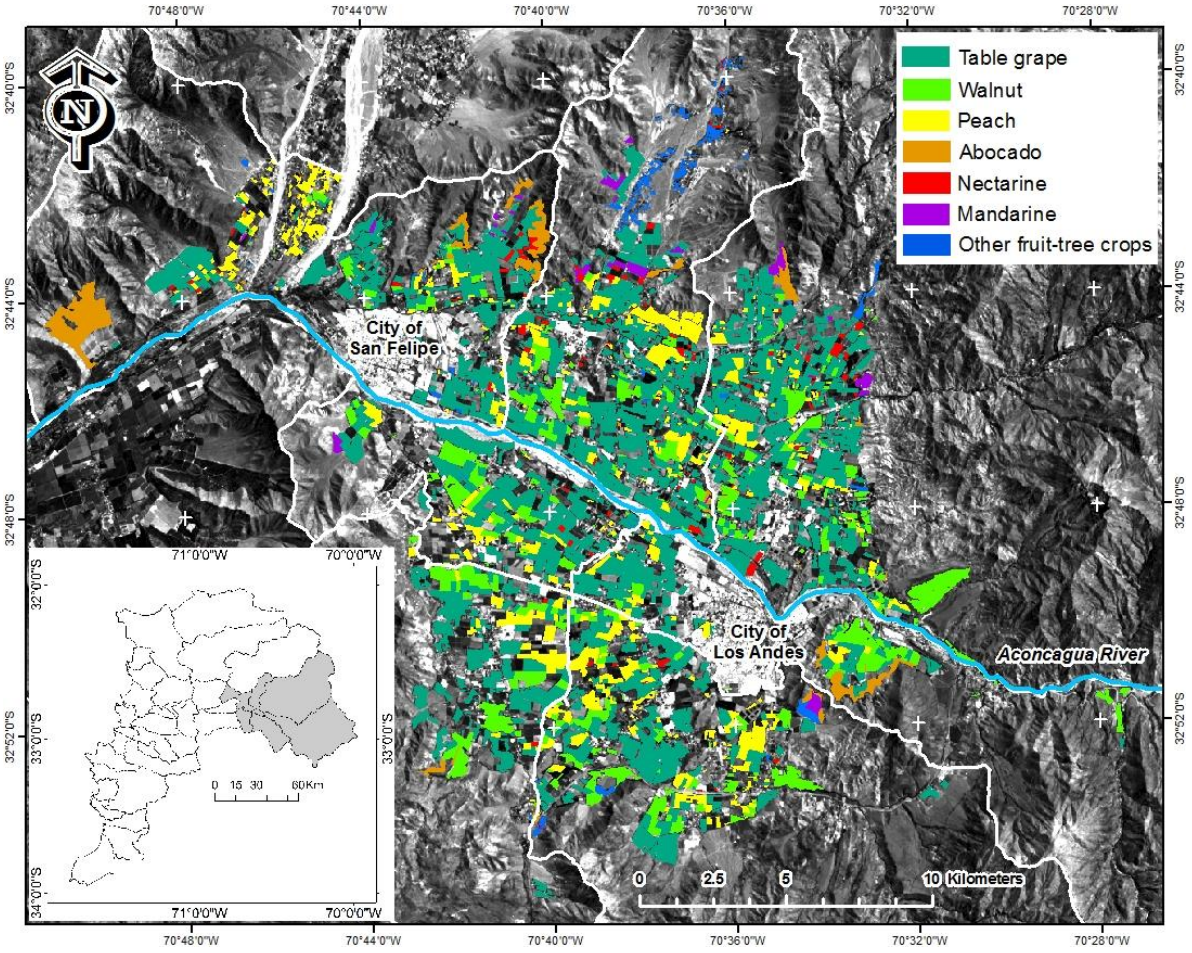


Figure 3-1: Distribution of the tree-fruit crops within the study area, located in the Aconcagua Valley, central Chile. White lines denote county boundaries. Map courtesy of Peña

Table 3-1: Areas of crop types involved in classification (Peña, written communication, January 28, 2016).

Tree-fruit crop	Area	
	km^2	%
Table grape	96.39	53.61
Walnut	42.21	23.48
Peach	25.93	14.42
Avocado	7.22	4.02
Nectarine	2.86	1.59
Mandarin	2.05	1.14
Others	3.12	1.74
Total	179.79	100

3.2 Remote sensing data

3.2.1 Landsat-8 images

A SITS was constructed with all the cloud-free Landsat-8 images collected during the 2014-15 growing season of the study area (Table 3-2). The images were provided by the USGS (United States Geological Survey), and were searched and downloaded from the Global Visualization Viewer (http://glovis.usgs.gov/).

Landsat-8 was launched on February 11th, 2013, as the eighth satellite in the United States' Landsat program since 1972 (USGS, 2015). The temporal frequency or revisit period of the Landsat-8 satellite is 16 days. The pushbroom sensor OLI (Operational Land Imager) is carried on the Landsat-8, which samples nine bands in the optical spectrum. In this study six out of the nine bands are used, except for the Bands 1 (0.43–0.45 µm), Bands 8 (0.5-0.68 µm) and 9 (1.36–1.39 µm). The panchromatic band 8 was not used because of its broad spectral width, while Bands 1 and 9 neither were used because they were designed for coastal water and atmospheric aerosol applications (Roy et al., 2014). Besides, thermal Bands 10 (10.60-11.19 µm) and 11 (11.50-12.51 µm) acquired by the TIRS (Thermal Infrared Sensor) instrument on board this satellite were not included in the SITS, mainly because their relatively coarse spatial

resolution of 100 m compared to the 30 m in band 2-7. Table 3-3 lists the main technical characteristics of the Landsat-8 OLI bands used in this study.

Table 3-2: Acquisition dates of the Landsat-8 images comprising the time series used in this study (Peña, written communication, January 28, 2016).

#	Acquisition date	
1	August 6, 2014	Winter
2	September 6, 2014	Spring
3	October 24, 2014	Spring
4	January 12, 2015	Summer
5	January 28, 2015	Summer
6	February 13, 2015	Summer
7	March 1, 2015	Summer
8	March 17, 2015	Summer
9	April 2, 2015	Autumn

Table 3-3: Main technical characteristics of the Landsat-8 OLI (Operational Land Imager) image bands used in this study (USGS, 2015).

BAND				
#	Spectral region	Spectral width (µm)	Spatial resolution (m)	Radiometric resolution (bits)
2	blue	0.45-0.51	30 m	12
3	green	0.53-0.59		
4	red	0.64-0.67		
5	near Infrared	0.85-0.88		
6	shortwave infrared	1.57-1.65		
7	shortwave infrared	2.11-2.19		

3.2.2 Data preprocessing

The downloaded Landsat-8 images are the Level 1 terrain corrected (L1T) products, and a geometric correction had already applied on them. The procedures include a systematic

geometric correction as well as the use of ground control chips and a digital elevation model (DEM) to further improve the image geometry, resulting in a circular geolocation error below 12 m at the 90% confidence level (Roy et al., 2014). The default image output coordinate system is Universal Transverse Mercator (UTM) with World Geodetic System 1984 (WGS84) datum. In addition to the consistent geometric correction procedures in L1T, spatial coregistration across the time series were also be verified by checking the spatial match between some randomly selected crop fields across different images. Only some negligible spatial shifts on the boundaries of the fields were detected which is less than the length of half pixel (Peña, written communication, January 28, 2016).

The radiometrical correction was done in two steps: first converted digital numbers to top-of-atmospheric (ToA) radiance; then converted into surface apparent reflectance. The radiometric calibration parameters for the first conversion and the FLAASH (Fast Line-of-Sight Atmospheric Analysis of Spectral Hypercubes) module for the second conversion are available in ENVI© (Environment for Visualizing Images) 5.0.3 software (Exelis Visual Information Solutions, Inc., Boulder, USA). This procedure is a MODTRAN (Moderate Resolution Atmospheric Transmission) based algorithm that allows modeling the at-surface irradiance and at-surface radiance of the image's pixels by taking into account a set of user-defined scene- and atmosphere- parameters (Peña, written communication, January 28, 2016).

3.3 Ground information

The Fruit Cadastre of the study area, carried out by the Chile's Agrarian Policies and Studies *Bureau* (Oficina de Estudios y Políticas Agrarias, ODEPA), was used to construct the field sample database to train the classifiers and to validate their results. This is an online database (http://odepa.cl) that shows the fruit-tree crop field boundaries of the Region of Valparaíso updated by field campaigns to the 2013-2014 growing season. The training/validation pixels were manually selected from full vegetation covered areas observed within the field boundaries, which were superimposed to one of the summer images of the time series. To avoid the inclusion of mixed edge pixels, fields below 22,500 m^2 (equivalent to a spatial window size of 5 × 5 pixels) or with a too narrow shape were discarded (Table 3-4) (Peña, written communication, January 28, 2016).

Table 3-4: Fields comprising each of the target crops. The original number of fields (second column) was somewhat reduced after a size-filtering procedure (third column) (Peña, written communication, January 28, 2016).

Crop type	**Original number of fields (polygons)**	**Fields used as samples**				
		#		Area (m^2)		
		Polygons	Pixels	Mean	Standard deviation	Total
Table grape	3,361	1,962	9,486	39557	15181	77,611,529
Walnut	613	359	2,846	64,694	65,850	23,225,170
Peach	925	348	2,408	45,489	30,241	15,830,244
Avocado	93	52	1,496	13,1146	155,495	6,819,633
Nectarine	193	50	544	34,450	14,241	1,722,521
Mandarin	48	37	464	51,335	35,571	1,899,402
Total	5,233	2,808	17,244	45,267	38,820	127,108,500

CHAPTER 4 METHODS

In this chapter, the methods applied to solve the research questions are introduced in detail. The first research question is whether and how much the indices, generated based on bands, are improving the classification result of the crop types of interest in the study area. To answer this question, two datasets should be prepared: one contains spectral bands only, the other has both bands and created indices. The form of indices needs to be decided first, then the indices can be calculated from the bands. Next, the classifier to conduct the classification should be implemented, which will be applied on both datasets so that the result can be compared. In this project, Linear Discriminant Analysis (LDA) and two types of penalized LDA, the ridge and the lasso, are chosen as classifiers, for LDA has been proven to have the best performance in a previous project with similar study area and crop types (Peña & Brenning, 2015). Spatial cross-validation resampling strategies are recommended for accuracy assessment in order to take the spatial autocorrelation into account (Brenning, 2012). With these methods and strategies, the classification result of the two dataset can be compared to estimate the effect of the indices and with cross-validation the randomness in sampling can be mostly compensated. In short, crop type classification was carried out for two feature sets, with and without normalized difference ratios, using three different classifiers LDA, ridge- and lasso-based penalized LDA, in order to identify the effect of the created indices on crop type classification and assess the performance of different classifiers, using a methodology similar to Peña & Brenning (2015). The work flow of the project is shown in Figure 4-1.

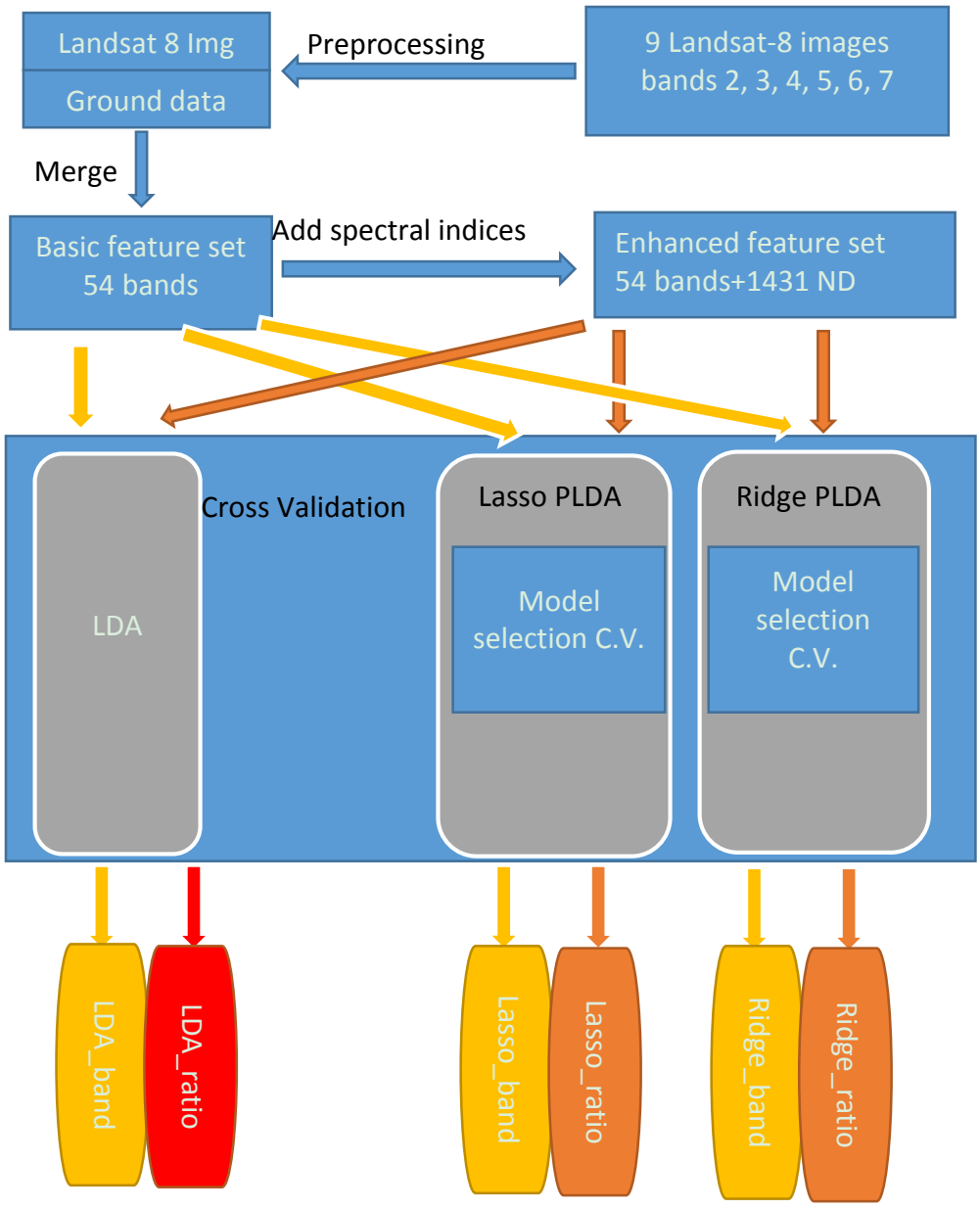


Figure 4-1: Work-flow. "LDA ratios" is red because it did not work on the enhanced feature set. In the output, XXX_ratio stands for the results on the enhanced feature set, while the XXX_band stands for the result derived from the old feature set.

4.1 Normalized difference index

Relying on the univariate characteristics of a predictor may cause the missing critical information about data, for the relationship between one predictor and another may be even more important than their seperate effect (Fielding, 2007a). Thus, index contains more than one predictors can provide extra information for data analysis. Normalized difference vegetation index (NDVI) is one of the most popular vegetation greeness indicators in the remote sensing context due to its simplicity and robutness, and NDVI timeseries have been widely used for crop-type classification, biomass estimation and crop pheonogy monotoring et al. (Araya et al., 2013; Chen et al., 2011; Ozdogan, 2010; Peña & Brenning, 2015; Zheng et al., 2015; Zhong et al., 2011). However, according to Peña (2015), NDVI timeseries based crop type classfication produced much less accurate results compared to full-band Landsat image timeseries; furthermore, based on variable importance assessment the two bands involved in NDVI (red and NIR) are not the most significant spectral bands for the discrimination between the crop types in that project.

In order to explore the potential of other indices, whose composing bands are not limited to red and NIR, ratios of every two bands in the satellite image stack are added into the full-bands time series dataset. In addition, ratios of the bands from different images may be able to detect “jumps” or “deltas” in the phenological curve of the crops of interests.

The final form of new features is the normalized difference of two different bands. Simple ratios of two bands would have large range of values including $\pm\infty$ because there is a considerable number of zeros in the dataset after preprocessing. The simple ratios’ value may

change dramatically with small disturbance in one of the spectral bands, so they are replaced by normalized difference ratios, which have the form: $\frac{b1-b2}{b1+b2}$. The “b1” and “b2” are two different bands, which can from either the same or different images. In addition, this normalized form can compensate illumination differences within one scene as well as the illumination difference between images (Peña, Brenning, & Sagredo, 2012). After normalized formulation, most of the values of the new indices are within the range from minus one to one. In case both b1 and b2 are zero, zeros in the dataset were set to ten power minus seven. For n bands, every two bands in the dataset can generate one new index, but only n(n-1)/2 number of indices are necessary instead of n(n-1), because the two indices using the same two bands b1 and b2 but in reverse order are additive inverses of each other. By keeping one of every two ratio indices that only differ in their signs, 54*53/2 = 1431 new features are added into the 54-band stack creating a high dimensional dataset.

4.2 Cross-validation and spatial aggregation

Theoretically at least one training set and one test set are required to build a supervised classification model, but in practice resampling methods are essential for model selection and evaluation, which involves repeatedly drawing training set and re-training the model (James et al., 2013). In the train/test approach the data is partially used either as training set or as test set. The resampling approach can help to gain more information about the model by fully use all the data instead of sampling fit and test the model once (James et al., 2013). Cross-validation is one of the resampling-based estimation procedures with different flavors such as leave-one-out cross-validation (LOOCV) and *k*-fold cross-validation. In LOOCV, a dataset is split into two parts as training set and test set with only one observation in the test set; in other words, in each iteration the model is fitted on the training set and tested on the left-out observation. LOOCV has advantages including little bias and zero randomness in dataset division; however, it is obviously very time consuming (James et al., 2013). The k-fold is much more efficient, for it has the dataset divided into k parts of equal size, and in each iteration leaves one of the k parts for validation and the rest for training.

The k-fold approach is more widely used and has been applied in crop type classification studies in order to produce unbiased error estimates (Brenning, Kaden, & Itzerott, 2006; Mathur & Foody, 2008; Peña & Brenning, 2015; Tatsumi, Yamashiki, Canales Torres, & Taipe, 2015; Zheng et al., 2015). In Mathur & Foody (2008), a 5-fold cross-validation was used for model selection, and Zheng et al. (2015) used a 10-fold cross validation instead; however, selection while the models are evaluated on fixed validation datasets in these two study. in Tatsumi et al. (2015) a 4-fold cross-validation was chosen for model selection, and a 10-

repeated 10-fold cross-validation for model robustness assessment; and similarly, Peña & Brenning (2015) applied a 100-repeated 10-fold cross validation for performance estimation (r-repeated means the k-fold cross validation process will repeat for r times with different data partitioning).

A 5-fold cross-validation is in the inner loop for selecting the penalty coefficient λ of the penalized LDA classifiers that produce the minimum mean classification error rates in the predefined range. A 100-repeated 10-fold cross-validation is applied in outer CV loop: at the beginning of every repetition, the data is equally divided into 10 parts. Only one of ten part is used as test data, while the remaining nine are parts used as training data. This procedure will repeat 100 times to calculate a mean error rate (cvMER) and a standard deviation (SD) for each classifier. This process makes sure the model is tested on data independent from training set in every model fitting iteration, also the average performance and the robustness of the classifiers can be estimated from the 100-repeated 10-fold cross validation.

Using cross-validation, the training and test sets are expected to be independent; however in a classification project on pixel based remote sensed data, the spatial autocorrelation related problem cannot be ignored in statistical analysis (Brenning et al., 2006; Brenning, 2012; Ruß & Brenning, 2010). In the original dataset after preprocessing, each row or records represents one pixel's spectral profile in the image stack, but the pixels within one field are very similar in spectral signature regarded as strong dependent observations. If the nearly identical observations from one field can enter into both training and test sets, the test set will not be independent from the training set; as a consequence, the possible overfitting of classifiers to the training set cannot be detected on the training set and the prediction accuracy

will probably be overestimated (Brenning, 2012; Peña & Brenning, 2015; Ruß & Brenning, 2010). On the other hand, data from different fields can be seen as independent observations, for the crop variety and cultivation patterns, which are generally managed on field levels, are the main drives for the spectral behaviors (Brenning et al., 2006). Thus, pixels in the same field should be assigned to the same fold and be selected in to the training set as a pack to reduce the impact of spatial autocorrelation on the model. This process is also called spatial clustering or spatial aggregation for data sampling.

Figure 4-2 shows the nested process of the cross-validated model estimation and model selection with spatial aggregation. First, when the dataset is divided into 10 folds at the beginning of the outer loop iteration the division is done on field level, and so is the following resampling process. Second, after predictions on test set are made according to the fitted model, the predicted results are aggregated to the field level by assigning the type of the majority records of that field to the rest of the records of the same field.

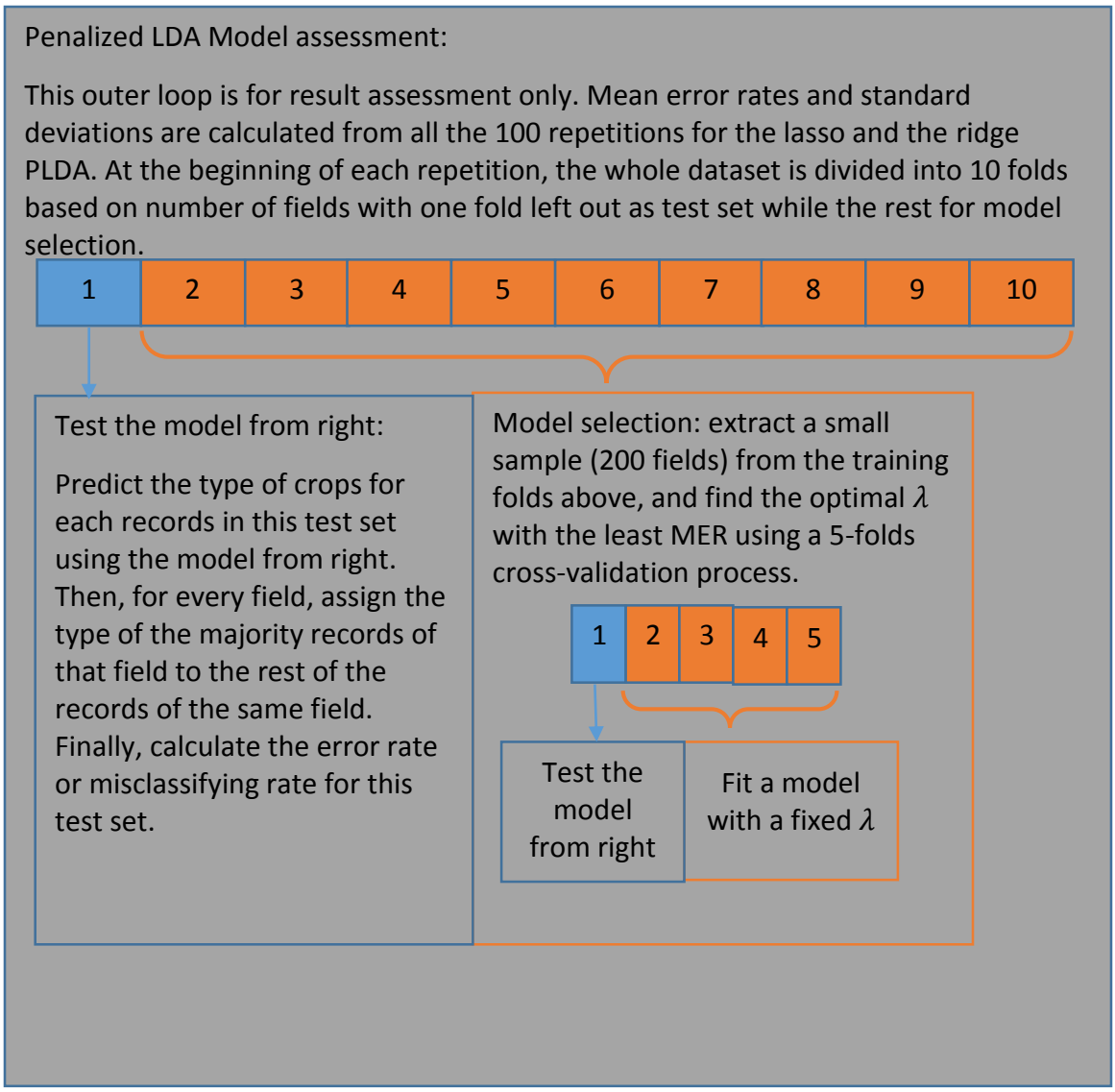


Figure 4-2: Description of the cross-validation and resampling process

4.3 Classifier

In this project, Fisher's linear discriminant analysis (LDA) method was chosen to classify the six crops in the study area, the Andean foothills of the Aconcagua River basin in Chile. In spite of the trend of new advancing machine learning algorithms applied in remote sensed data analysis, linear model has distinct advantages over them in respect to inference on real-world problem and often surprisingly outperforms the new methods (James et al., 2013). According to Fielding (2007), despite its simple model and relatively strong assumptions on the data distribution, LDA's performance is among the best in some empirical tests. The performance of LDA is evaluated in the study by Li, Zhu, & Ogihara (2006), their conclusion is that LDA has accurate output especially on high dimensional data, and the mean error rate is comparable to SVM in that study. This is also proved by the previous project in the adjacent Maipo basin, Chile, in which LDA outperformed Support Vector Machine (SVM) and Random Forest in addition to being much faster (Peña & Brenning, 2015). The classical LDA has been used for conifer species and rainforest tree species mapping, and canonical LDA has been applied for LULC classification (Clark, Roberts, & Clark, 2005; Guo, Pu, & Bin, 1997; Lobo, 1997).

However, LDA is not suitable for ill-posed problems, i.e. more features than training samples, and penalized LDA is not as stable as SVM (Bandos, Bruzzone, & Camps-Valls, 2009). The high accuracy and fast speed are the two main reasons for LDA being chosen as the classifier in this project, while the possible collinearity in high dimensional data may cause singularity problem for LDA. In this situation, PLDA can be applied to solve the ill-posed problem.

4.3.1 Introduction to Linear Discriminant Analysis

The LDA classifier is not a simple concept, for it is derived through multiple ways such as the normal model, the optimal scoring problem and Fisher's discriminant problem (Witten & Tibshirani, 2011). Fisher's linear discriminant analysis was first introduced in 1936 aiming to find a set of uncorrelated linear combinations of the variables which can maximize the ratio of the between-class variance and within-class variance (Venables & Ripley, 2002). LDA solves an optimization problem and has been a simple but popular method widely used for various multiple-class classification problems (Fielding, 2007a; James et al., 2013; Li et al., 2006). As introduced in Chapter 2, LDA can be seen as a dimension reduction method, for the original feature space is projected on to a new lower-rank feature space by choosing a few linear combinations of the variables as coordinates. For instance, with two of the linear combinations, the original high dimensional feature space can be projected to a two dimensional feature space, then one can use linear boundaries to partition the lowered feature space into regions labeled with classes (Hastie, Tibshirani, & Buja, 1994).

Multiple versions of the mathematical derivation of LDA can be found in many text books and introduction papers (Duda, Hart, & Stork, 2001; Fielding, 2007a; Fukunaga, 1990; James et al., 2013; Venables & Ripley, 2002). However, there are several closely related concepts and techniques including discriminant function analysis (DFA), discriminant analysis (DA) and Fisher's discriminant analysis. In this section, multiple sources of mathematical derivation of LDA are reviewed and combined in order to provide a general overview.

- **Defining the optimization problem**

Fisher's LDA seeks a set of project directions which can maximize the ratio of the separation of the class means, calculated as the between-class scatter matrix denoted by B, and sum of within-class covariance, denoted by W (James et al., 2013; Welling, 2005). Assume there are *n* total training observations *x* of *p* variables grouped into *K* predefined classes, and there are n_k observations belonging to the kth class. In other words, if vector a can maximize the objective function

$$J(a) = \frac{a^T B a}{a^T W a} \tag{1}$$

then $a^T x$ is the linear combination of the variables, which can explain most of between-class variance. To estimate the real covariance matrix, scatter matrices B and W are calculated from training dataset as:

$$W = \sum_{k=1}^{K} \sum_{i \in k} (x_i - \mu_k)(x_i - \mu_k)^T \tag{2}$$

$$B = \sum_{k=1}^{K} (\mu_k - \bar{x})(\mu_k - \bar{x})^T \tag{3}$$

$\bar{x}$ is the overall mean of the training dataset, and μ_k is the expected value of the kth class, estimated by the mean of each class in the training dataset (Welling, 2005).

- **Solving the constrained optimization problem**

To simplify the computation of maximizing the objective function (1), we can choose a to make $a^T W a$ = 1 since it is a scalar itself (Welling, 2005). Then the problem of maximizing $J(a)$ can be transformed to following constrained optimization problem:

$$min_a \quad -\frac{1}{2}a^T Ba \tag{4}$$

$$s.t. \quad a^T Wa = 1 \tag{5}$$

(The coefficient $-\frac{1}{2}$ in (4) is added for convenience)

Then we can use the method of Lagrange multipliers to solve the optimal a. According to the Lagrangian:

$$L_p = -\frac{1}{2}a^T Ba + \frac{1}{2}\lambda(a^T Wa - 1) \tag{6}$$

a stationary point is

$$Ba = \lambda Wa \quad => \quad W^{-1}Ba = \lambda a \tag{7}$$

Then, it is becoming a generalized eigen-problem. It is ease to prove that the eigenvector corresponding to the maximum Eigen value is the one wanted (Welling, 2005). In addition, the eigenvalues are the proportions of the between classes variance explained by the linear combinations. To classify *k* groups, at most *k-1* canonical coordinates (eigenvectors) can be found to project the original feature space. The percent between-group variance explained by each dimension can reveal their importance in LDA. With the optimal set of discriminant functions, for each new observation a discriminant score will be calculated and class membership will be assigned following the application of some threshold to the discriminant score (Fielding, 2007a; Hastie et al., 1994).

- **From a probability classifier perspective**

The maximum likelihood algorithm with Bayes' decision rules is the most widely used parametric classifier (Jensen, 2005). Another way to understand LDA is through probability

models based on Bayes' Theorem (James et al., 2013; Venables & Ripley, 2002). In the Chapter 4 of the book "An introduction of statistical learning: with Application in R", James et al. provide a step by step explanation of how to use Bayes' Theorem for classification, as well as LDA's relationship with logistic regression and Quadratic Discriminant Analysis (QDA). Here is a brief review of related contents of the chapter, with some important point of view to understand LDA from Vernables and Ripley (2002). Continue with the p variables n observation case mentioned before, it is natural to assign an observation x to the class having the maximum probability among the K classes, and according to Bayes' Theorem:

$$Pr(Y = k|X = x) = \frac{Pr(Y = y) * P\,r(X = x|Y = k)}{P\,r(X = x)} \tag{8}$$

$$P_k(x) = \frac{\pi_k * f_k(x)}{\sum_{i=1}^{K} \pi_i * f_i(x)} \tag{9}$$

$\Pr(\mathrm{Y} = \mathrm{y})$ is called prior probability of the kth class, denoted by π_k. $\Pr(\mathrm{Y} = \mathrm{k}|\mathrm{X} = \mathrm{x})$can also be rewritten as density function of X for an observation that comes from the kth class, denoted by $f_k(x)$. $Pr(Y = k|X = x)$, abbreviated as $P_k(x)$, is the posterior probability of one given observation x belongs to the kth class. For one observation, $Pr(X = x)$ is the same when $P_k(x)$ is calculated for different k (James et al., 2013). Then $Pr(Y = k|X = x)$ or $P_k(x)$ is proportional to the product of $Pr(Y = y)$ and $Pr(X = x|Y = k)$ as expressed in the (9). If π_k and $f_k(x)$ are estimated, a classifier can be developed to approximate the Bayes classifier.

Bayes classifier is a family of simple probabilistic classifier based on Bayes' theorem. π_k are estimated by $\frac{n_k}{n}$, but $f_k(x)$ is more difficult to estimate. Thus, a few assumptions need to be introduced into the model. First, assume the observation X has a p-dimensional Gaussian or Normal distribution with μ being the mean of the whole dataset and Σ being the covariance

matrix written as $X \sim N(\mu, \Sigma)$ (James et al., 2013; Venables & Ripley, 2002). The second assumption is that the observations in the kth class are drawn from a multivariate Gaussian distribution: $X_k \sim N(\mu_k, \Sigma_k)$ where μ_k and Σ_k the class specific mean vector and covariance matrix. A third assumption is only made by LDA that all the K classes share the same covariance matrix, written in mathematical expression: $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k = \Sigma$. Without this assumption, the decision boundary is not linear and LDA becomes Quadratic Discriminant Analysis (James et al., 2013; Venables & Ripley, 2002). Then, the density function of the kth class is:

$$f_k(x) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma (x-\mu_k)\right) \quad (10)$$

With $f_k(x)$ plugged into (9) and some algebra, including transforming to the logarithm of $P_k(x)$, the kth class which can maximize $L_k$is the one to assign to the observation x. In reality, these assumptions may not be satisfied, but LDA has reasonable robustness to non-normal distribution and even to lightly different class covariance (Hastie, Buja, & Tibshirani, 1995).

$$L_k = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k{}^T \Sigma^{-1} \mu_k + log\, \boldsymbol{\pi_k} \quad (11)$$

To calculate L_k, μ_k and Σ need be estimated from sample mean and scatter matrix within each class. $\mu_k{}^T \Sigma^{-1} \mu_k$ is called Mahalanobis distance, on which the discriminant is operated (Venables & Ripley, 2002; Welling, 2005). The relationship between this maximum probability classification method and the optimization problem is that by solving the latter new predictors are created, those linear discriminants or coordinate functions, with unit within-class variance. Thus, $||\Sigma|| = 1$, and $L_k(x)$ can be simplified as:

$$L_k = x^T \mu_k - \frac{1}{2}\|\mu_k\|^2 + log\, \pi_k \quad (12)$$

Besides, Fisher's LDA also reduces the dimension to $r = min(p, K-1)$. For example,

$$L_2 - L_1 = x^T(\mu_2 - \mu_1) + const$$

By calculating the differences r times for one observation, the maximum L_k can be found and the observation x will be assign to the corresponding class (Venables & Ripley, 2002).

4.3.2 Shrinkage Methods

With 1431 newly created normalized differences ratios added, the enhanced feature set has 1485 variables. The new dataset may contain redundant or highly correlated variables, and LDA becomes unstable suffering from high variance (James et al., 2013). Then penalized versions of LDA are proposed to solve the high dimensional discriminant problem, since it maintains the advantages of LDA while adding shrinkage to the discriminant vectors (Witten & Tibshirani, 2011). One reason for the failure of LDA in classifying high-dimensional dataset is the singularity problem of the within-class covariance matrix, caused by multicollinearity of variables, or caused by too more variables or features than observations (Fielding, 2007a; Witten & Tibshirani, 2011). In the context of multiple linear regression, the two best-known techniques for shrinking the regression coefficients towards zero to reduce the model variance are ridge regression and the lasso (James et al., 2013). In this research, lasso and ridge based LDA are utilized to solve the singularity problem that LDA faced when new normalized band ratios adding to the feature set, and their performances are evaluated based on mean error from cross-validation.

It would be easier to introduce the ridge and the lasso in a linear regression context to see how the penalization or regularization works. In a p-variable multivariable regression the response, assume that response variable Y can be predicted by the *p* different predictors (X_j) as:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{13}$$

The ordinary fitting process is the procedure to find the set of β_i that can minimize the residual sum of squares (RSS), which is the most common approach called the least squares fit.

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \quad (14)$$

$\hat{y}_i$ is a prediction of Y when X equals to x_i, and (x_i, y_i) is the observed pair. Ridge regression and the lasso are very similar to least squares fit except that the loss functions are different:

$$F_R = RSS + \lambda \sum_{j=1}^{p} {\beta_j}^2 \quad (15)$$

$$F_L = RSS + \lambda \sum_{j=1}^{p} |\beta_j| \quad (16)$$

(15) is for the ridge regression and (16) is for the lasso. These two loss functions indicate that small values of coefficients are preferred in ridge regression and the lasso. They both have shrinking effects on the coefficients' absolute values, and the tuning parameter λ should be greater than zero; however, the Lasso tends to penalize some of the coefficients to exactly zero faster than the ridge regression when λ is large enough (James et al., 2013). Due to the choice of norm distance, the ridge regression has bigger regularization effects on the large coefficients while for the lasso the effects are the same for large and small coefficients. Thus, for the lasso penalization, small coefficients can be tuned to exactly zero while the large ones remain positive. Nevertheless, when the tuning parameter λ is too big, all of the coefficients can be turned to zero making a null regression. It is obvious that a too large λ or too harsh penalty can result in an under-fit model, while a really small λ may cause a high variance over-fit model. Hence, adjusting the λ is to balance between model variance and bias.

4.4 Investigation of model selection process and coefficients

In order to have insight of the ridge and penalization processes, the relationship between the MER and shrinkage penalty coefficient λ is investigated. The dataset is divided into two parts as training and test set, and each contains the same number of fields. A series of ridge and lasso based penalized LDA models are fitted and tested on these two sets with a sequence of λ whose range was predefined as 10^{-5} to 10^{-25}.

After the optimal λ being located, the canonical coefficients are retrieved from the fitted model, and five variables corresponding to the five quantiles of the coefficients: the maximum, the third quantile, the median, the first quantile and the minimum are investigated. The coefficients of these five variables are selected out as examples to show their changes over λ.

In order to assess the importance of different variables, the canonical coefficients in the first discriminant vector of the optimal model is retrieved and analyzed. The first discriminant vector covers almost 50% of the between-group variance, and the combination of the second and the first discriminant vector is almost 78% percent of the total between-group variance. The remaining three discriminant vectors also contribute a small portion, but without doubt the first discriminant vector plays the most critical role in the classification of six types of crop plants. Thus, the coefficients in the first discriminant vector are analyzed in particular in this chapter. In addition, to standardize the coefficients for different variables, they are multiplied to the corresponding standard deviation. If not otherwise specified, the word “coefficients” in the following context means the adjusted canonical coefficients in the first discriminant vector.

4.5 Chapter Summary

In this chapter, the creation of the new feature set is introduced, followed by the mathematical formulation of LDA and penalized LDA models are introduced in detail. Then it presents as well as the methods to interpret the results.

The high-performance computing facility of the Chair of Geographic Information Science at the University of Jena, Germany was used for computationally intensive tasks. In Witten and Tibshirani's work (2011), they proposed a penalized version of LDA with lasso through Fisher's discriminant framework. Daniela Witten implemented the method in the R package "penalizedLDA", and it is the function with the same name used in this project to realize the lasso PLDA (2015). The function "fda" in the R package "mda" (Hasty, 2015) is applied for the ridge based penalized LDA. The package adopted for spatial aggregation and spatial cross validation framework is "sperroreset", which created and maintained by Dr. Alexander Brenning since 2012.

CHAPTER 5 RESULT

5.1 Comparison of different classifiers and different feature sets

Table 5-1: Cross-validated results

	cvMER train	SD train	cvMER test	SD test
lasso PLDA on basic feature set	0.269	0.009	0.304	0.028
lasso PLDA on Enhanced feature set	0.210	0.007	0.238	0.022
ridge PLDA on basic feature set	0.044	0.003	0.137	0.018
ridge PLDA on Enhanced feature set	0.005	0.002	0.103	0.020
classic LDA on basic feature set	0.044	0.003	0.132	0.019

The cross-validated results of the five combinations of classifiers and features, including the cross-validated mean error rates (cvMER) on training and test sets, as well as standard deviation (SD) of the cvMERs. The basic feature set, or baseline feature set contains bands only, while the enhanced feature set has indices in addition to the spectral bands. If not otherwise specified, the cvMER means cvMER test.

Except for lasso-based classifiers, all the other classifiers' cross-validated mean error rates (cvMER test) are below 0.15 (Table 5-1). With the enhanced feature set, the classification results of both lasso and ridge are improved notably. There are around 6.6 percentage points (p.p.). increase in accuracy using lasso penalized LDA on the enhanced feature set compared to the basic feature set, while for LDA with ridge penalty the increase is less but still as high as 3.5 p.p. In other words, the new feature set reduce the cvMER by 21.6% (lasso) and 25.2% (ridge). The lowest test cvMER is from ridge-based penalized LDA with indices, and the highest is from the lasso on feature set without indices. Lasso penalty is not suitable in classifying these trees with Landsat images, for it increases the cvMER in classification compared to LDA. Ridge works much better than lasso in this case.

On the same feature set without indices, Lasso penalized LDA produces substantially inferior classification results. The cvMER of LDA_bands is 17.2 p.p. lower than Lasso_bands and around 0.5 p.p lower than ridge_bands. The ridge penalty does not affect LDA's result considerably, for a 0.5 p.p. difference in cvMER is negligible.

Training errors of all the combinations are lower than the corresponding test errors as expected, and standard deviations of the cvMERs are also lower on training data. Ridge penalized LDA clearly overfits the training data resulting in an extraordinarily low mean error rate compared to the cross-validated test error on the feature set with new indices (cvMER train= 0.005, cvMER test= 0.103). Whereas, the cvMER of ridge_bands (0.137) is not so different from its mean training error (0.044). In spite of the fact of overfitting the training sets, the ridge penalized LDA is also generalizable for the enhanced feature set and produces more accurate result.

5.2 Model selection process of ridge PLDA and lasso PLDA

5.2.1 Ridge penalized LDA

In order to have insight of the ridge penalization processes, the relationship between the coefficients of discriminant vectors and shrinkage penalty coefficient λ is investigated with the change of classification error rate. The method is introduced in Section 4.4. Probably because of the large size of training set including 817 fields of data, the error rate is notably lower than the cross validated results whose training sets are smaller than 200 fields.

As shown in the Figure 5-1 below, these coefficients change their signs twice before gradually getting closer to zero, but their absolute values shrink all the way towards zero when λ gets bigger. First, these coefficients dramatically shrink towards zero and this trend prolongs a little longer after crossing zero, then they keep approaching to zero smoothly. However, when λ at some point between $10^{-1.75}$and $10^{-1.5}$, these coefficients suddenly flip their signs.

In the Figure 5-1, the dotted line shows the lowest classification error rate: it reaches its minimum at 0.032 when λ equals $10^{-0.5}$(≈0.316). Seen from the plot in Figure 5-2 for ridge MER: it has more than one local minimum, and the second lowest error rate appears at 10^{-4}. This is a very small penalty on the discriminant vectors, for which the penalized LDA is almost identical to an LDA. The error rate also fluctuates, inferring that the penalization algorithm is not stable (Bandos, Bruzzone, & Camps-Valls, 2009b).

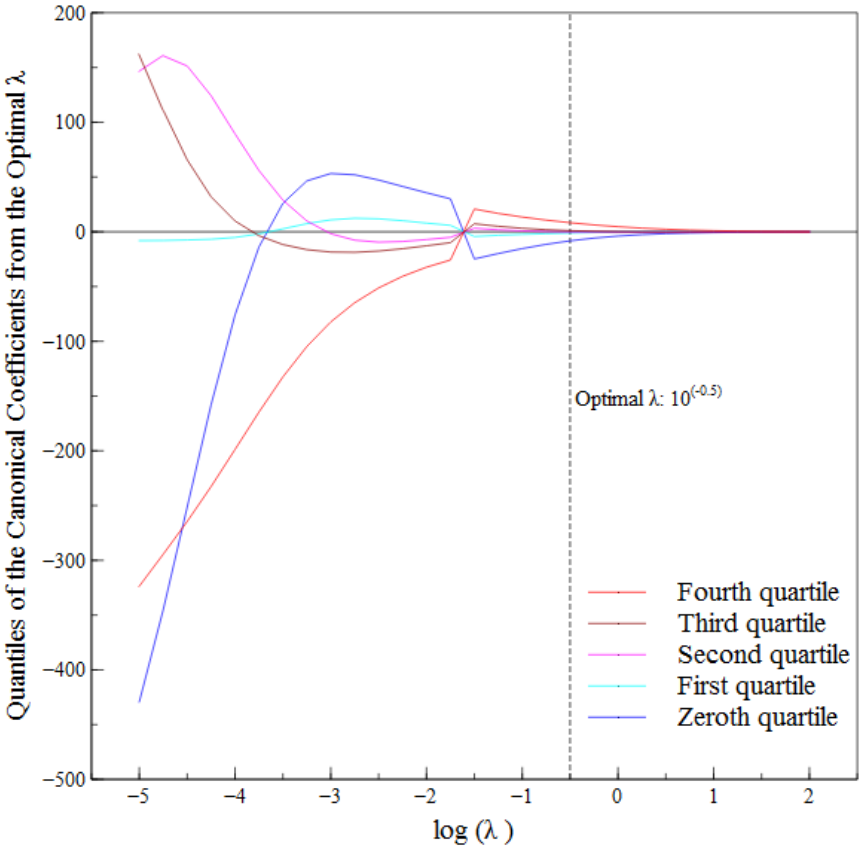


Figure 5-1: Five variables' canonical coefficients change over λ in ridge penalized LDA. The five variables are corresponding to the five quantiles of the coefficients: the maximum, the third quantile, the median, the first quantile and the minimum, in the model with the lowest MER.

Table 5-2: Confusion matrix of the optimal Ridge PLDA where λ = $10^{-0.5}$).

Predicted class	Reference						
	Table Grape	Walnut	Peach	Avocado	Nectarine	Mandarin	PPV
Table Grape	249	6	0	0	15	0	0.922
Walnut	7	244	0	0	0	0	0.972
Peach	20	0	972	0	0	0	0.980
Avocado	0	0	0	46	0	0	1.000
Nectarine	0	0	0	6	23	0	0.793
Mandarin	0	0	0	0	0	110	1.000
TPR	0.902	0.976	1.000	0.885	0.605	1.000	0.968

The TPR (true positive rate), or the producer's accuracy of a class is defined as the ratio of true positive to the true number of the observations belonged to the class. The positive predictive value (PPV), or the user's accuracy of a class is defined as the ratio of true positive to the number of observations labeled as this class. The result refers to a fixed test set without cross-validation as introduced in Section 4.4.

Peach and mandarin have the highest true positive rate, which means all the fields of these two crop types are correctly labeled. Walnut, table grape and avocado also have decent TPRs above 0.85; whereas only around 60% nectarine are classified correctly. Two out of five fields cultivating nectarine are misclassified as table grape, which is a serious confusion between these two crops. Seen from the user's accuracy or PPV, all the classes except for nectarine have very high precision, which means only 79.3% of all the fields of fruit-trees predicted to be Nectarine trees are truly Nectarine. This result suggests that even when the overall accuracy is high even close to one, the unbalanced sizes of classes may cover serious misclassification on small classes.

5.2.2 Lasso penalized LDA process

Lasso and Ridge penalized LDA show completely different patterns in the change of accuracy and coefficients shrinkage strength, while Lasso is even more unpredictable in classification accuracy (Figure 5-2). Lasso has significant higher error rates compared to the ridge at all λ levels. Seen from the grey plot representing the lasso, the lowest error rate is 0.164 occurred at $10^{-4.25}$and $10^{-3.5}$. However, the confusion matrix of the optimal λ (Table 5-2) shows that only 12% of the predicted nectarine trees are true and 52.6% nectarine trees are classified correctly. Mandarin trees are easily classified as avocado, while all the avocado trees are labeled correctly.

Actually, the plot of the lasso PLDA does not show any obvious global minimum error rate, for it fluctuates lightly when λ is smaller than $10^{-1.5}$, then dramatically goes up at $10^{-1.25}$. When λ equals to $10^{-1.25}$ the overall accuracy is as low as 0.6. Seen from the producer's accuracy or TPR in the confusion matrix (Table 5-3), only avocado is perfectly labeled but the TPRs of all the other fruit-trees are low. When λ is larger than 10^{-1}, error rate is stable around 0.4. It can be inferred that the discrimination is almost random when λ is too large, which is also proved in the corresponding confusion matrix (Table5-5). All the fields in the test set are predicted as peach, indicating the model is underfitting as a result of over-penalization of the model with large λ.

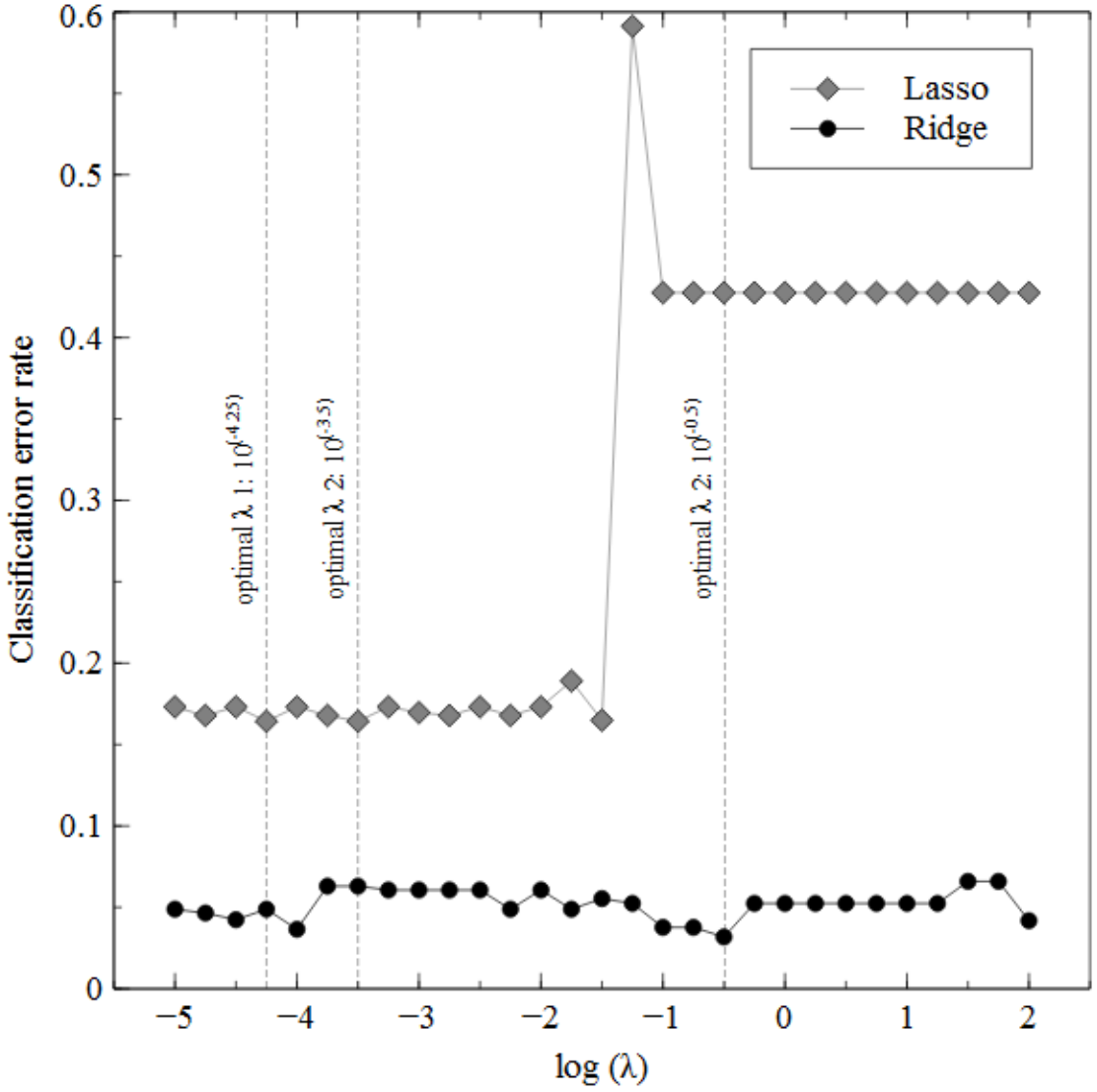


Figure 5-2: The relationship of classification error rates and penalty parameter λ with lasso and ridge classifiers.

Table 5-3: Confusion matrix of the lasso PLDA model where $\lambda = 10^{-3.5}$ or $10^{-4.25}$. The meanings of PPV and TPR are the same with Table 5-2. The result refers to a fixed test set without cross-validation as introduced in Section 4.4.

Predicted class	Reference						
	Table Grape	Walnut	Peach	Avocado	Nectarine	Mandarin	PPV
Table Grape	184	0	11	0	18	0	0.864
Walnut	0	208	46	0	0	0	0.819
Peach	17	0	886	0	0	0	0.981
Avocado	0	0	0	52	0	41	0.559
Nectarine	75	42	29	0	20	0	0.120
Mandarin	0	0	0	0	0	69	1.000
TPR	0.667	0.832	0.912	1.000	0.526	0.627	0.836

Table 5-4: Confusion matrix of the lasso PLDA model where $\lambda = 10^{-1.25}$.

Predicted class	Reference						
	Table Grape	Walnut	Peach	Avocado	Nectarine	Mandarin	PPV
Table Grape	80	19	91	0	0	0	0.421
Walnut	47	170	412	0	0	0	0.270
Peach	54	61	328	0	18	0	0.711
Avocado	0	0	0	52	0	66	0.441
Nectarine	95	0	141	0	20	0	0.078
Mandarin	0	0	0	0	0	44	1.000
TPR	0.290	0.680	0.337	1.000	0.526	0.400	0.409

Table 5-5: Confusion matrix of the lasso PLDA model where $\lambda >= 10^{-1.25}$.

Predicted class	Reference						
	Table Grape	Walnut	Peach	Avocado	Nectarine	Mandarin	PPV
Table Grape	0	0	0	0	0	0	0
Walnut	0	0	0	0	0	0	0
Peach	276	250	972	52	38	110	0
Avocado	0	0	0	0	0	0	0
Nectarine	0	0	0	0	0	0	0
Mandarin	0	0	0	0	0	0	0
TPR	0.000	0.000	1.000	0.000	0.000	0.000	0.572

5.3 Interpretation of the canonical coefficients

The canonical coefficients of the variables in the model with least mean error rate are retrieved as introduced in Chapter 4.4. To show the importance of the variables in the classification, the absolute values of their canonical coefficients are graphed in Figure 5-3. Generally speaking, the lower left part shows brighter colors than the rest parts (Figure 5-3), which may indicate that the first four Landsat images play an important role in classification than the rest five. The band features from first two images seem to be dominant predictors with very large canonical coefficient values as very bright pixels on the diagonal elements. Seen from the band level, the third band, which is the band 4 (red) in Landsat 8 image, seem to have larger values, and many outstanding predictors are related to the third bands of each image. Actually, the top ten predictors that carry the most weights are all band variables, they are the band 4, 5, 6 and 7 in image 1, the band 3, 4, 5 and 7 in image 2, and the band 7 in image 8.

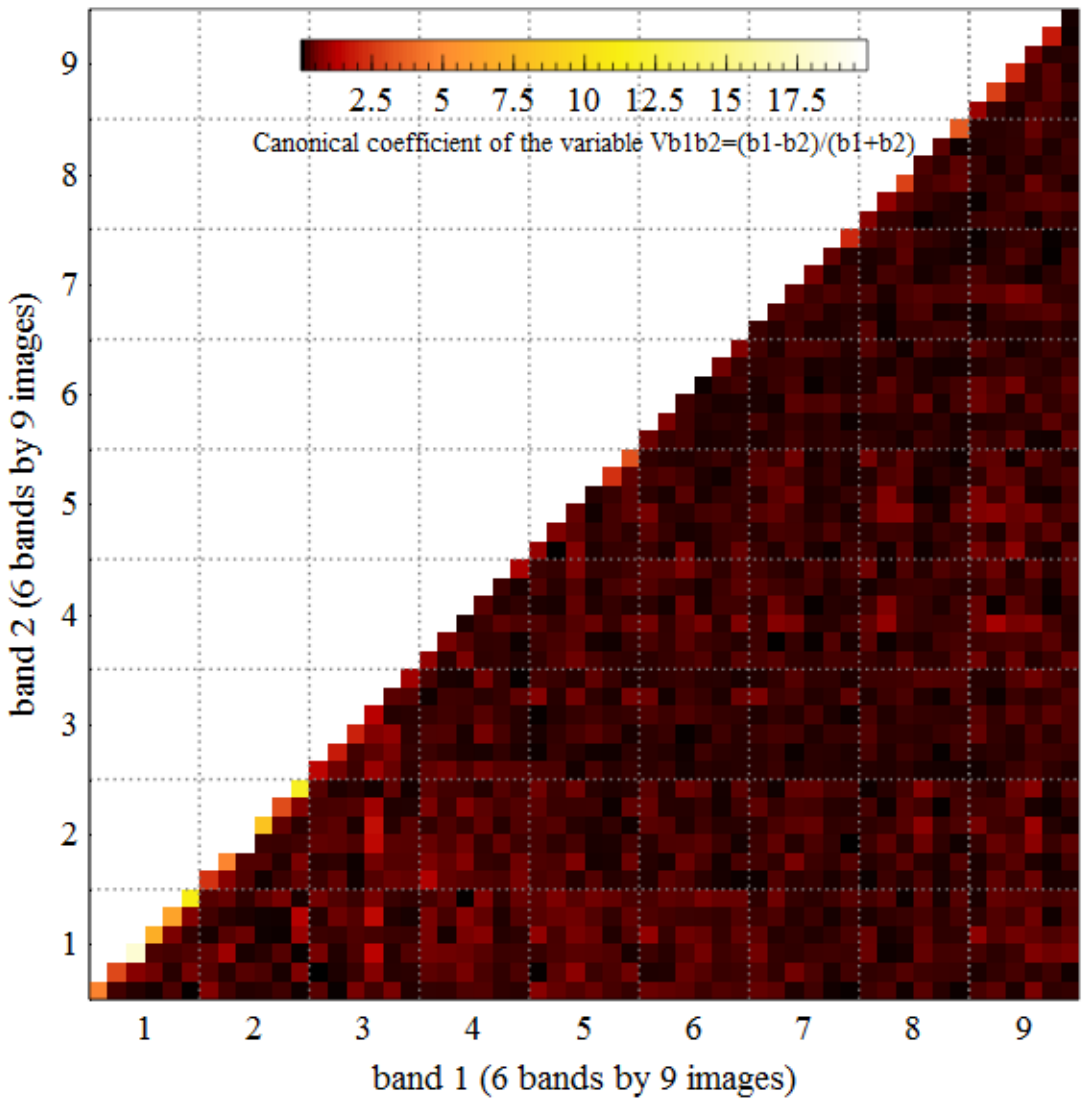


Figure 5-3: Canonical coefficients of 1485 predictors derived from the ridge PLDA model with lowest cvMER.

Figure 5-3 is a 2D graph designed to show the importance of variables and its link with bands and dates in a comprehensive way. The absolute value of a coefficient multiplied by the standard deviation of that variable is used as a measure of importance of it (abbreviated as new coefficients). The 2D graph is made of 1485 small blocks representing the new coefficients in the first discriminant vector. There are nine Landsat images, and only the bands from band 2 to band 7 for each image are included in this study. For example, the new coefficient of the created variable ind_b23_b16 is located at the cross of eighth column from left and fifth row from bottom. In addition, the diagonal elements are representing the bands themselves.

5.3.1 Statistical summary

Table 5-6: The mean values of the canonical coefficients by band and image

Image	Band 2 (blue)	Band 3 (green)	Band 4 (red)	Band 5 (nir)	Band 6 (shortwave)	Band 7 (shortwave)	Avg_by_image
1	2.474	1.557	9.299	3.634	3.298	5.787	4.342
2	1.254	2.538	10.185	4.245	1.604	6.054	4.313
3	0.751	0.944	1.048	0.768	0.243	0.541	0.716
4	0.491	0.335	0.163	0.231	0.129	0.621	0.328
5	0.500	0.369	0.319	0.107	1.211	1.687	0.699
6	0.287	0.328	0.165	0.078	0.268	0.377	0.251
7	0.125	0.182	0.253	0.308	0.223	1.083	0.362
8	0.380	0.464	1.445	0.178	0.190	1.782	0.740
9	0.521	1.428	1.122	0.369	0.941	0.079	0.743
AVG_BY_BAND	0.754	0.905	2.667	1.102	0.901	2.001	1.388

The mean value of the adjusted coefficients for each band is calculated from all the indices related to that band as well as the band itself. The average canonical coefficients of the 54 bands are listed in the table above. Avg_by_img is the average canonical coefficient of a Landsat image, while the Avg_by_Band is the mean value of the canonical coefficients of the bands with same wavelength in all the nine images.

The information contributed to correct classification is strongly related a few bands or images (Figure 5-4 and Figure 5-5). The first and second images play dominant roles in the crop classification, for they have much higher average coefficients than the other images (Table 5-6, Figure 5-4). According to the Table 3-2 in Section 3.2.2 about the seasons and acquired dates of every images, it can be inferred that the early spring and early summer may be more suitable for the classification among the six tree-fruit crops in the study area. From the bar chart in Figure 5-4, the highest two bars represent the third bands in their images (red), and the shortwave infrared bands of the first two images have the third and the fourth highest average canonical coefficients. After calculating the mean canonical coefficients of the variables related to spectra, from the highest to the lowest the order is red (0.64 µm -0.67 µm), shortwave

infrared (1.57 µm - 1.65 µm), near infrared (0.85 µm -0.88 µm), green (0.53 µm -0.59 µm), blue (0.45 µm - 0.51 µm) (Figure 5-5).

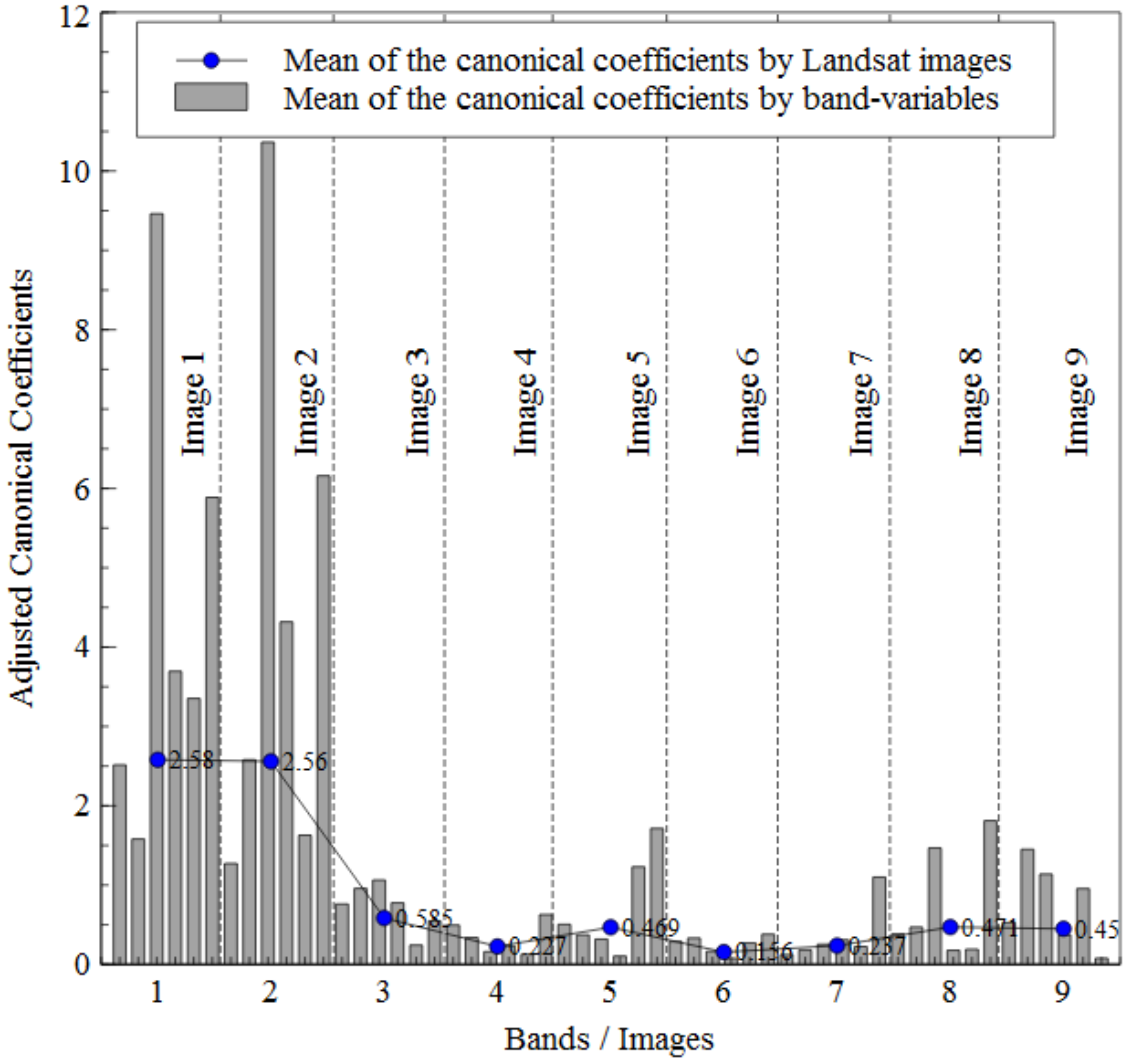


Figure 5-4: The mean of canonical coefficients of each bands grouped by different images.

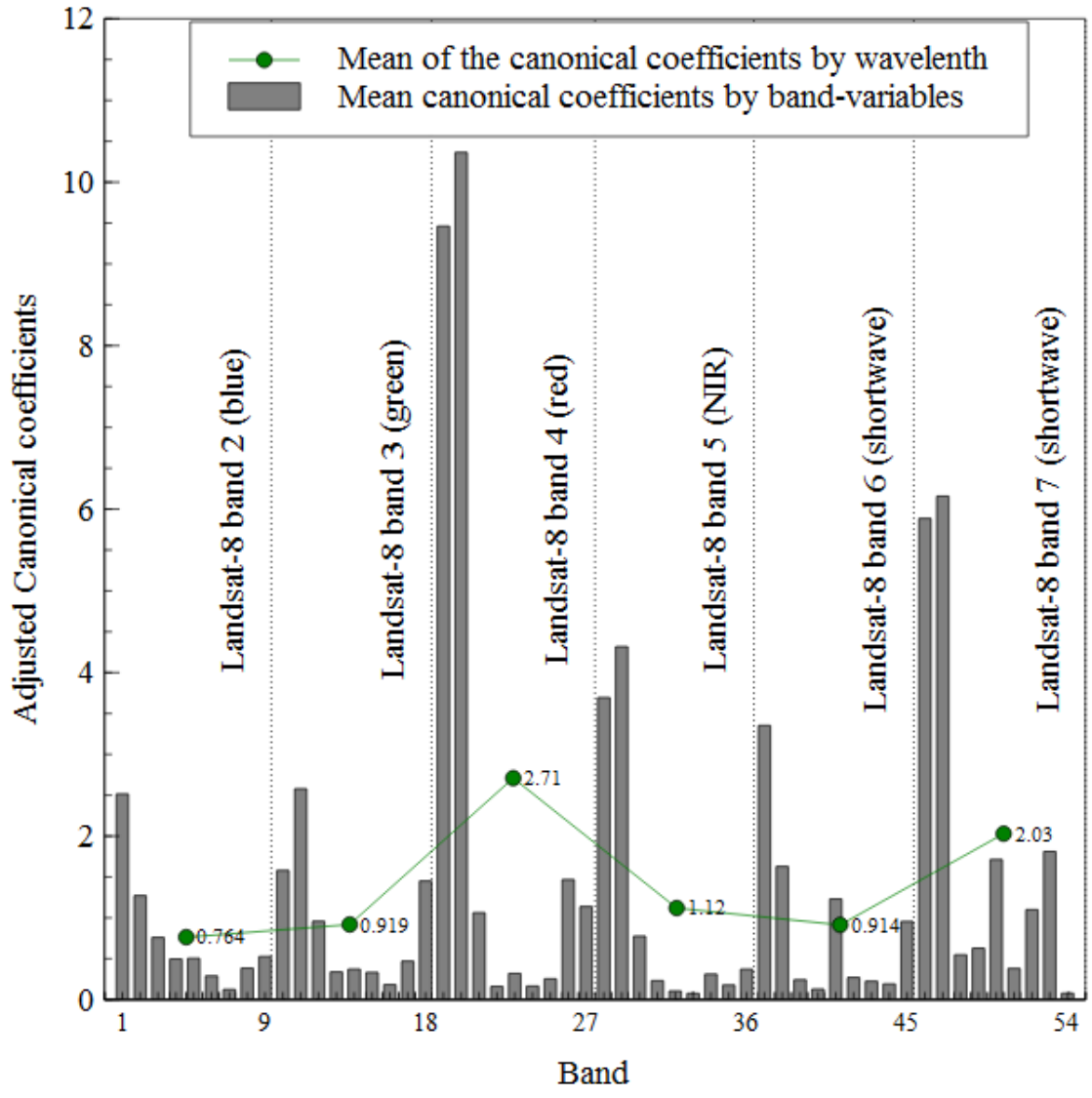


Figure 5-5: The mean of canonical coefficients of each bands grouped by different wavelength.

CHAPTER 6 DISCUSSION

6.1 Discussion of the results

Classification accuracy has been improved on the enhanced feature set with thousands of indices created from existing bands, and the overall accuracy of ridge PLDA is good (around 90%). With same size of training set (200 fields) best results from the previous project is around 94% using LDA as classifier on full-band SITS or full feature set (Peña & Brenning, 2015). However, there were only four crop types involved in that research, the increase of spectral complexity may decrease the correctness of crop identification. In an older research done by Brenning (2006), six classifiers are compared for crop identification also based on Landsat multispectral images; it was found that stabilized linear discriminant analysis (SLDA) and SVM outperformed the other four classifiers, including classification trees (bagging and double-bagging), k-Nearest Neighbours (1-NN) and Logistic regression, with mean error rate at 0.201 and 0.2. This may seem not very impressive; however, there were 9 crop types used in the study and the classifiers were trained on fewer fields, which may increase the difficulty of the training of classifiers and the mean error rate. When SVM and RF are applied for LULC classification or crop type mapping, SVM has an overall accuracy from 80% to 90% (more or less) while RF around 80% (Brenning et al., 2006; Gislason et al., 2006; Peña & Brenning, 2015; Rodriguez-Galiano, Chica-Olmo, et al., 2012; Tatsumi et al., 2015). Thus, the 90% accuracy of ridge penalized LDA produced in this project is a good result.

The outcome that the lasso-based penalized LDA being the least accurate clssifier for this dataset may suggest that each feature contains a small portion of useful information in distinguish different types of trees. Meanwhile, many variables' coefficients are penalized to zero in the lasso penalization process, while the ridge tends to only reduce the coefficients gradually but will not turn the coefficients to exactly to zero (James et al., 2013). This feature does not affect the accuracy but the model may not be easy for interpretation with too many variables left. However, in this case, preserving more features may be the critical reason for the ridge to produce the better results than the lasso, for in a situation where all the predictors are related to the response non of the coefficients should equal to zero (James et al., 2013).

According to James et al. (2013) ridge regression can balance bias and variance and performs best in situations where the least squares estimates have high variance. Ridge regression with LDA is also regarded as efficient and effective by Zhang, Dai, Xu, & Jordan (2010). In a study by Bandos et al., (2009) penalized LDA is compared to a proposed regularized LDA on clssification with hyperspectral data; PLDA is considered as unstable and less accurate than RLDA though on some dataset it outperformed the rest types of LDA solving ill-posed probelem. Neverthless, the authors also admitted that the unstable was probably due to the choice of regularizer, which is not the ridge regression nor the lasso.

6.2 Sources of uncertainty

Field survey data about fruit types was conducted by the Chile's Agrarian Policies and Studies *Bureau* (Oficina de Estudios y Políticas Agrarias, ODEPA), they are viewed as ground truth data. The uncertainty of the ground data is not the focal point in this study. Instead, the uncertainty from remote sensing data should be focused on, for errors and bias can be introduced in to the system through several preprocessing steps. First, the Landsat-8 images are Level 1 terrain corrected (LIT) product downloaded from USGS website has geometric correction done previously, including systematic correction and geometric correction with DEM and ground control chips. According to Roy et.al (2014) after this geometric correction the product has a circular geolocation error below 12m at the 90% confidence level. Second, the spatial coregistration has verified by Marco Pena, a collaborative researcher from Chile, for every Landsat-8 images by checking the spatial match between some randomly selected crops fields across different images. Only a few slight spatial shifts, below half the size of a pixel (<15m) on the boundaries, have been found. Considering the size of the fields, it is safe to say that these shifts do not affect the classification result significantly.

The radiometric correction process has been explained in detail in Section 3.3 in detail. First, raw digital numbers in the geometrically corrected products are converted into radiance images. This process is standard and implemented by ENVI 5.0.3; furthermore, the coefficients used are from the metadata provided by USGS. If systematic errors exist in this process, the uncertainties should still be acceptable.

Next, the output radiance is experienced atmospheric adjustment trying to counteract the effect of atmosphere. There is undeniable uncertainties and possible errors in atmospheric correction. In this step, extra uncertainties can only enter the system from a set of user-defined scene- and atmosphere- parameters, which are relatively more subjective. If the operator wrongly set the parameters, the atmospheric influence cannot be estimated accurately and neither can be the offsetting work. Besides, this atmospheric process may overcorrect some pixels causing negative values in the output whether the user-defined parameters are correct or not, for these parameters can only roughly describe a large area. The negative values of radiance obviously do not have any physical meaning. This may affect the classification result, for those negative values are all turned to zero and the formal variance in these negative values are gone.

However, this study is mainly focused on the question “whether the adding of normalized difference index can improve the classification result”. This change may increase the difficult of classification, but the influence on difference between two kinds of data or the difference between two penalized LDA can hardly be measured. I would argue that even though the surface apparent reflectance may not be accurate due to subjective coefficients and over correction, the differences between the five combination combinations are not notably affected.

6.3 Limitations and implications

6.3.1 Limitations

One limitation in this study is that the training sets, as well as test sets, contain six types of crops of equal number of fields in both the model selection and model estimation steps in the cross-validation section. The reason is that the six crops in the dataset are unequally distributed as introduced in Section 3.1. The numbers of fields and pixels of each crop type after necessary data preparation are presented in Table 3-4, from which we can see how serious the unbalance is, especially between the first three crop types and the last three. The difference is even larger after sampling objects are aggregated from pixels to fields. For example, the ratio of the number of pixels between table grape and Avocado is around five, but that of the fields is around eight.

200 fields were used to be selected into training set randomly, but the result is that it is highly possible that one or more crop types are missing from the sample. Then the absence of a few crop types leads to critical malfunction in classification. A stratified sampling strategy was tried after random sampling failed, but stratified sampling also faced the same problem of missing some crop types. Finally, equal number of fields are randomly selected by crop types into training and test sets. The plan that at most 200 fields are used to train the classifiers does not change, for the final training set can have at most 186 fields. Because the model is trained and selected on data with classes of the same size, a new test set is created by retrieving an equal number of fields of each crop type from the left-out fold in cross-validation estimation. How robust an algorithm is towards different distribution of data and number of classes is also noteworthy, which is not addressed in this study.

In the coefficients investigation section, the whole dataset is equally divided into two parts as training set and test set, which means within training set and test set the sizes of the six classes are different. The purpose of this section is not comparing different classifiers but to analyze the change of coefficients and error rate along the change of λ. Then the training set and test set should be kept the same. By using half of the dataset, there is no need to concern the absent of crop types. However, the effect of difference in balanced or unbalanced dataset on ridge and lasso PLDA is not considered. Whether the inconsistence of training/test set design in these two sections will affect the conclusions is a question to answer.

6.3.2 Possible implications

One of the main findings of this research is that stable and obvious enhancement on classification accuracy can be made by adding normalized difference indices derived from all two-band indices calculated across the 9 images. This new approach to reduce mean error rate does not require any new images or ancillary data. For future researchers interested in land cover classification, this method can be helpful to improve the accuracy with barely additional computer work. Second, three flavors of LDA have been tested and compared in a crop classification context and ridge penalized LDA produced the overall best results. LDA as a simple but powerful classifier has not been paid much attention in this field, and result of this study addresses its ability in distinguishing different fruit trees. LDA series classifiers are very fast considering their high accuracy in crop type identification, which is comparable to or even

better than some the state of art algorithms. For future projects and work about crop type classification, ridge penalized LDA can be a good choice. Also, the application of the ridge and the lasso PLDA may encourage studies in other fields, such as bioinformatics which also encounter with high dimensional data, to pay attention to these two PLDA methods.

In addition, through the analysis on canonical coefficients in the optimal model, the most important spectral bands and images are found out. Red and shortwave infrared (1.57 µm - 1.65 µm) carry dominant higher weights than other spectral bands. These two bands deserve further attention in vegetation indices related researches.

There are not found studies using ridge and lasso PLDA for LULC. The canonical coefficients are not investigated to explore the variables importance in many studies related to LDA based classifiers. In Bandos et al (2009), the research was focused on the comparison of performance of LDA based classifiers. Several different flavors of lasso PLDA are compared on simulated data in (Merchante et al., 2012) and on bioinformatics data (Ma & Huang, 2008), while canonical coefficients were not interested either. In the studies by Brenning (2006) and by Peña & Brenning (2015) LDA is treated as a black box. This study provides a new approach to analysis the variable importance with LDA based classifiers in a LULC context.

6.4 Future directions

One of the future research directions can be the generalization of the conclusions, “adding new indices features can decrease the classification mean error rate, on other classifiers. LDA series classifiers may perform very well in this case, but for other crop types or land covers types, whether the created normalized difference index can decrease the error rate is needed for exploration. The classifiers to be tested next should also be able to deal with large amount of features, such as SVM which is the second best classifier in the previous project (Peña & Brenning, 2015). However, SVM was extremely time consuming even without adding new indices into the feature set. Another option may be other flavours of LDA, which should be much faster than complicated machine learning algorithms.

The result of variable importance assessment shows that red and shortwave infrared (band 4 and 7) are the most important spectral bands in Landsat-8 images in the crop classification, and these two bands are not the traditional vegetation related spectral bands green and near infrared. More research is needed to have insight into the biophysical structures and phenomena of the interaction between fruit-trees and solar radiation. Also, the normalized difference indices based on these two bands would better be paid more attention or compared to NDVI specifically in future studies.

CHAPTER 7 SUMMARY AND CONCLUSIONS

The purpose of this study is to seek a new approach to reduce the crop type misclassification rate by creating new spectral indices based on available bands. The study area is located in a valley in Chile, with six main fruit tree crop types interested in this research. After collecting and preprocessing all the available cloud-free Landsat-8 images, the ground truth data are combined with remote sensed data in a table format for batch processing purpose. The indices are generated through calculating the normalized difference ratios of every two bands (in a form like NDVI) in the old dataset which only contains band variables from the satellite images. This approach is tested on two types of penalized LDA classifiers and the results are assessed by spatial aggregated cross-validation. The LDA series of classifiers has been recognized as efficient and accurate models in crop type classification problems, and in this study the ridge penalized LDA and the lasso penalized LDA are chosen to solve the ill-posed problem which Fisher's LDA cannot solve. The results clearly show that, either on the ridge or the lasso based LDA, the mean classification error rates are decreased by adding new indices into the old feature set, which means the spectral indices can provide additional valuable information in crop type identification.

However, the ridge LDA significantly outperformed the lasso on both feature sets, which probably is caused by the different ways of penalization in the classifiers. The lasso tends to shrink the coefficients of small variables to exactly zero, while the ridge tends to keep all the variables in the feature set in the shrinking process. Thus, in a dataset which all the variables are related to the response, the ridge would have advantages over the lasso in accuracy. To assess the importance of the variables, the canonical coefficients are analyzed by different

bands. The first two images, acquired in late winter and early spring, have dominant higher average coefficients over the rest images mainly because of the Bands 4 and 7 in the two images carry the most weights. This is another interesting finding, for red and shortwave infrared are not the traditional vegetation related bands.

In conclusion, additional spectral indices like NDVI but generated from other spectral bands can also contribute to the crop classification. Second, the ridge based LDA is more suitable to classifying the fruit-tree crops in this study than the lasso. Third, the period from late winter and early spring may be the best period to recognize the crop types in this study. The last, the red and the second shortwave infrared bands contribute more to correct classification than the other spectral bands in this study.

APPENDIX A

List of Acronyms

LDA- Linear discriminant analysis

LULC- Landuse/landcover

NDVI- Normalized difference vegetation index

NIR- Near infrared

PLDA- Penalized linear discriminant analysis

RF- Random Forest

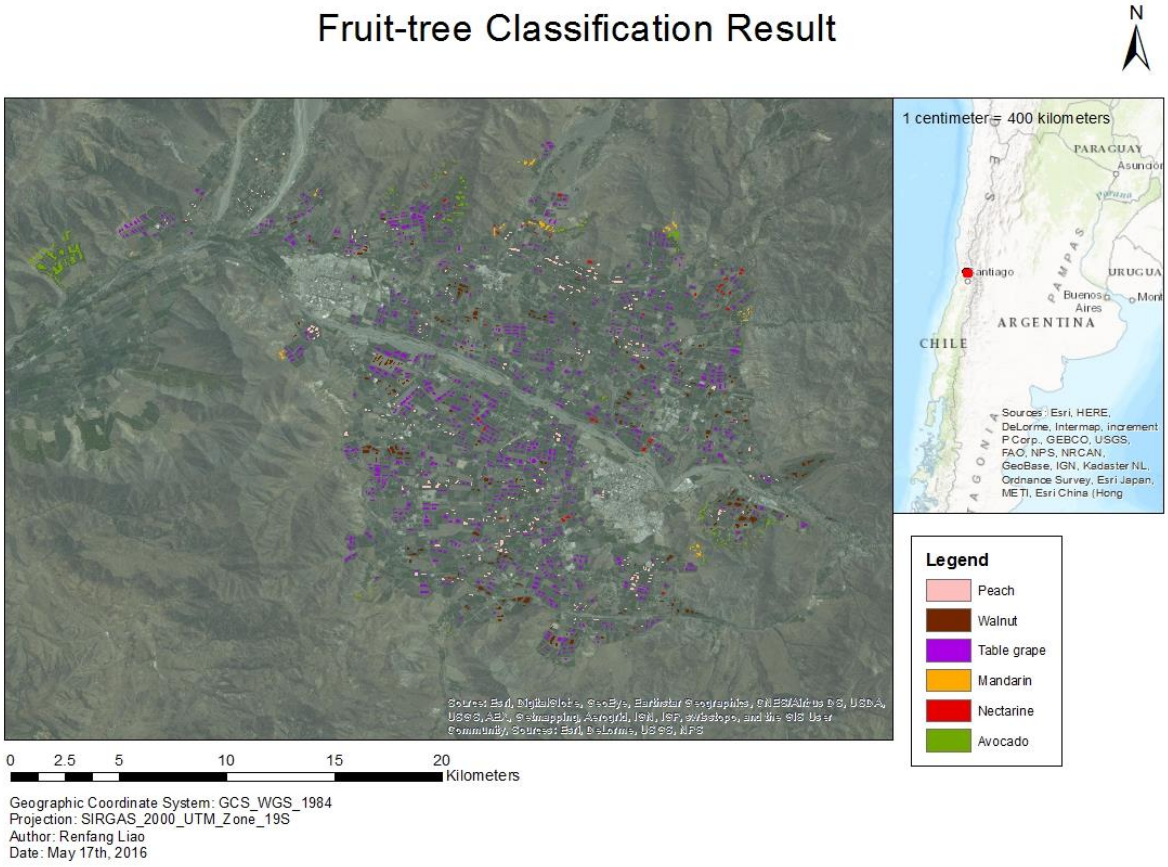
SD- Standard deviation

SITS- Satellite images times series

SVM- Support vector machine

APPENDIX B

Map of best classification result



REFERENCES

Araya, S., Ostendorf, B., Lyle, G., & Lewis, M. (2013). *Crop phenology based on MODIS satellite imagery as an indicator of plant available water content*.

Archibald, R., & Fann, G. (2007). Feature selection and classification of hyperspectral images with support vector machines. *Geoscience and Remote Sensing Letters …*, *4*(4), 674–677. http://doi.org/10.1109/LGRS.2007.905116

Atzberger, C. (2013). Advances in remote sensing of agriculture: Context description, existing operational monitoring systems and major information needs. *Remote Sensing*, *5*(2), 949–981. http://doi.org/10.3390/rs5020949

Bandos, T. V., Bruzzone, L., & Camps-Valls, G. (2009a). Classification of Hyperspectral Images With Regularized Linear Discriminant Analysis. *IEEE Transactions on Geoscience and Remote Sensing*, *47*(3), 862–873. http://doi.org/10.1109/TGRS.2008.2005729

Bandos, T. V., Bruzzone, L., & Camps-Valls, G. (2009b). Classification of Hyperspectral Images With Regularized Linear Discriminant Analysis. *IEEE Transactions on Geoscience and Remote Sensing*, *47*(3), 862–873. http://doi.org/10.1109/TGRS.2008.2005729

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

Brenning, A. (2012). Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. *International Geoscience and Remote Sensing Symposium (IGARSS)*, 5372–5375. http://doi.org/10.1109/IGARSS.2012.6352393

Brenning, A., Kaden, K., & Itzerott, S. (2006). Comparing classifiers for crop identification based on multitemporal landsat TM/ETM data. *Proceedings of the 2nd Workshop of the EARSeL SIG on Land Use and Land Cover*, (viii), 64–71. Retrieved from http://www.zfl.uni-bonn.de/earsel/papers/64-71_brenning.pdf

Chen, C. F., Son, N. T., Chang, L. Y., & Chen, C. R. (2011). Classification of rice cropping systems by empirical mode decomposition and linear mixture model for time-series MODIS 250 m NDVI data in the Mekong Delta, Vietnam. *International Journal of Remote Sensing*, *32*(18), 5115–5134. http://doi.org/10.1080/01431161.2010.494639

Clark, M. L., Roberts, D. A., & Clark, D. B. (2005). Hyperspectral discrimination of tropical rain forest tree species at leaf to crown scales. *Remote Sensing of Environment*, *96*(3-4), 375–398. http://doi.org/10.1016/j.rse.2005.03.009

Clemmensen, L., Hastie, T., Witten, D., & Ersbøll, B. (2011). Sparse Discriminant Analysis. *Technometrics*, *53*(4), 406–413. http://doi.org/10.1198/TECH.2011.08118

Conrad, C., Fritsch, S., Zeidler, J., Rücker, G., & Dech, S. (2010). Per-field irrigated crop classification in arid Central Asia using SPOT and ASTER data. *Remote Sensing*, *2*(4), 1035–1056. http://doi.org/10.3390/rs2041035

Craig, M., & Atkinson, D. (2013). *A Literature Review of Crop Area Estimation*. Retrieved from http://www.fao.org/fileadmin/templates/ess/documents/meetings_and_workshops/GS_SAC_2013/Improving_methods_for_crops_estimates/Crop_Area_Estimation_Lit_review.pdf

Dhumal, R. K., Kale, K. V, & Mehrotra, S. C. (2013). Classification of Crops from remotely sensed Images : AnOverview. *International Journal of Engineering Research and Applications (IJERA)*, *3*(3), 758–761.

Duda, O. R., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification* (2nd ed.). John Wiley & Sons, Inc.

Esch, T., Metz, A., Marconcini, M., & Keil, M. (2014). Combined use of multi-seasonal high and medium resolution satellite imagery for parcel-related mapping of cropland and grassland. *International Journal of Applied Earth Observation and Geoinformation*, *28*(1), 230–237. http://doi.org/10.1016/j.jag.2013.12.007

Fan, J. Q., & Lv, J. C. (2010). A Selective Overview of Variable Selection in High Dimensional Feature Space. *Statistica Sinica*, *20*(1), 101–148. http://doi.org/10.1016/j.micinf.2011.07.011.Innate

FAO. (2011). *Global Strategy to Improve Agricultural and Rural Statistics*. Rome, Italy.

Fielding, A. H. (2007a). Classification Algorithms 1. In *Cluster and Classification Techniques for the Bioscience* (1st ed., pp. 97–136). New York: Cambridge University Press.

Fielding, A. H. (2007b). Introduction. In *Cluster and Classification Techniques for the Bioscience* (1st ed., pp. 1–10). New York: Cambridge University Press.

Foley, J. a., Ramankutty, N., Brauman, K. a., Cassidy, E. S., Gerber, J. S., Johnston, M., … Zaks, D. P. M. (2011). Solutions for a cultivated planet. *Nature*, *478*(7369), 337–342. http://doi.org/10.1038/nature10452

Fritz, S., Massart, M., Savin, I., & Leo, O. (2006). The use of modis data in southern russia for crop acreage Estimations and inter-comparison of results from various crop acreage estimation methods. *ISPRS Archives XXXVI-8/W48 Workshop Proceedings: Remote Sensing Support to Crop Yield Forecast and Area Estimates*, (2005), 59–64. Retrieved from http://www.isprs.org/proceedings/XXXVI/8-W48/59_XXXVI-8-W48.pdf

Fukunaga, K. (1990). *Introduction to statistical pattern recognition* (2nd ed.). Academic Press Professional, Inc.

Gislason, P. O., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random forests for land cover

classification. *Pattern Recognition Letters*, *27*(4), 294–300. http://doi.org/10.1016/j.patrec.2005.08.011

Goswami, S. B., Aruna, M. S., & Bairagi, S. G. D. (2012). A Review : The application of Remote Sensing , GIS and GPS in Precision Agriculture. *Internarional Journal of Advanced Technology & Enginering Research (IJATER)*, *2*(1), 50–54.

Guo, P., Pu, R., & Bin, Y. (1997). Conifer Species Recognition: An Exploratory Analysis of In Situ Hyperspectral Data. *Remote Sensing of Environment*, *62*(2), 189–200. http://doi.org/10.1016/S0034-4257(97)00094-1

Guyon, I., & Elisseeff, A. e. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Hastie, T., Buja, A., & Tibshirani, R. (1995). Penalized Discriminant Analysis. *The Annals of Statistics*.

Hastie, T., Tibshirani, R., & Buja, A. (1994). Flexible discriminant-analysis by optimal scoring. *Journal of the American Statistical Association*, *89*(428), 1255–1270. Retrieved from <Go to ISI>://WOS:A1994PU33000012

Heuvelink, G. B. M., Burrough, P. a., & Stein, A. (1989). Propagation of errors in spatial modelling with GIS. *International Journal of Geographical Information Systems*, *3*(4), 303–322. http://doi.org/10.1080/02693798908941518

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. New York: Springer Science+Business Media. http://doi.org/10.1007/978-1-4614-7138-7

Jensen, J. R. (2005). Thematic Information Extraction: Pattern Recognition. In *Introductory Digital Image Processing: A Remote Sensing Perspective* (3rd ed., pp. 337–401). Toronto: Person Education Canada, Inc.

Karegowda, A. G., Manjunath, A. S., & Jayaram, M. a. (2010). Feature Subset Selection Problem using Wrapper Approach in Supervised Learning. *International Journal of Computer Applications*, *1*(7), 13–17. http://doi.org/10.5120/169-295

Kohavi, R., & Kohavi, R. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1-2), 273–324. http://doi.org/10.1016/S0004-3702(97)00043-X

Li, T., Zhu, S., & Ogihara, M. (2006). Using discriminant analysis for multi-class classification: an experimental investigation. *Knowledge and Information Systems*, *10*(4), 453–472. http://doi.org/10.1007/s10115-006-0013-y

Lobo, A. (1997). Image segmentation and disciminant analysis for the identification of land cover units in ecology. *IEEE Transactions on Geoscience and Remote Sensing*, *35*(5), 1136–1145.

Ma, S., & Huang, J. (2008). Penalized feature selection and classification in bioinformatics. *Briefings in Bioinformatics*, *9*(5), 392–403. http://doi.org/10.1093/bib/bbn027

Mathur, A., & Foody, G. M. (2008). Crop classification by support vector machine with intelligently selected training data for an operational application. *International Journal of Remote Sensing*, *29*(8), 2227–2240. http://doi.org/10.1080/01431160701395203

Merchante, L. F. S., Grandvalet, Y., & Govaert, G. (2012). An Efficient Approach to Sparse Linear Discriminant Analysis. *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 1167–1174.

Mkhabela, M. S., Bullock, P., Raj, S., Wang, S., & Yang, Y. (2011). Crop yield forecasting on the Canadian Prairies using MODIS NDVI data. *Agricultural and Forest Meteorology*, *151*(3), 385–393. http://doi.org/10.1016/j.agrformet.2010.11.012

Odenweller, J. B., & Johnson, K. I. (1982). Crop identification using Landsat temporal-spectral profiles. *Remote Sensing of Environment*, *54*, 39–54. http://doi.org/10.1016/0034-4257(84)90006-3

Ozdogan, M. (2010). The spatial distribution of crop types from MODIS data: Temporal unmixing using Independent Component Analysis. *Remote Sensing of Environment*, *114*(6), 1190–1204. http://doi.org/10.1016/j.rse.2010.01.006

Peña, M., & Brenning, A. (2015). Assessing fruit-tree crop classification from Landsat-8 time series for the Maipo Valley, Chile. *Remote Sensing of Environment*, *171*, 234–244. http://doi.org/10.1016/j.rse.2015.10.029

Peña, M., Brenning, A., & Sagredo, A. (2012). Constructing satellite-derived hyperspectral indices sensitive to canopy structure variables of a Cordilleran Cypress (Austrocedrus chilensis) forest. *ISPRS Journal of Photogrammetry and Remote Sensing*, *74*, 1–10. http://doi.org/10.1016/j.isprsjprs.2012.06.010

Rodriguez-Galiano, V. F., Chica-Olmo, M., Abarca-Hernandez, F., Atkinson, P. M., & Jeganathan, C. (2012). Random Forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. *Remote Sensing of Environment*, *121*, 93–107. http://doi.org/10.1016/j.rse.2011.12.003

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, *67*(1), 93–104. http://doi.org/10.1016/j.isprsjprs.2011.11.002

Roy, D. P., Wulder, M. A., Loveland, T. R., C.E., W., Allen, R. G., Anderson, M. C., … Zhu, Z. (2014). Landsat-8: Science and product vision for terrestrial global change research. *Remote Sensing of Environment*, *145*, 154–172. http://doi.org/10.1016/j.rse.2014.02.001

Rozenstein, O., & Karnieli, A. (2011). Comparison of methods for land-use classi fi cation incorporating remote sensing and GIS inputs. *Applied Geography*, *31*(2), 533–544. http://doi.org/10.1016/j.apgeog.2010.11.006

Ruß, G., & Brenning, A. (2010). LNAI 6178 - Data Mining in Precision Agriculture: Management of Spatial Information. *Management*, 350–359.

Schuster, C., Schmidt, T., Conrad, C., Kleinschmit, B., & F??rster, M. (2015). Grassland habitat mapping by intra-annual time series analysis -Comparison of RapidEye and TerraSAR-X satellite data. *International Journal of Applied Earth Observation and Geoinformation*, *34*(1), 25–34. http://doi.org/10.1016/j.jag.2014.06.004

Simonneaux, V., Duchemin, B., Helson, D., Er-Raki, S., Olioso, a., & Chehbouni, a. G. (2008). The use of high-resolution image time series for crop classification and evapotranspiration estimate over an irrigated area in central Morocco. *International Journal of Remote Sensing*, *29*(1), 95–116. http://doi.org/10.1080/01431160701250390

Tatsumi, K., Yamashiki, Y., Canales Torres, M. A., & Taipe, C. L. R. (2015). Crop classification of upland fields using Random forest of time-series Landsat 7 ETM+ data. *Computers and Electronics in Agriculture*, *115*, 171–179. http://doi.org/10.1016/j.compag.2015.05.001

USGS. (2015). Frequently Asked Questions about the Landsat Missions. Retrieved April 15, 2016, from http://landsat.usgs.gov/best_spectral_bands_to_use.php

Venables, W. N., & Ripley, B. D. (2002). Classification. In *Modern Applied Statistics with S* (4th ed., pp. 331–351). New York: Springer.

Welling, M. (2005). *Fisher Linear Discriminant Analysis*. Retrieved from http://www.cs.huji.ac.il/~csip/Fisher-LDA.pdf

Wieland, M., & Pittore, M. (2014). Performance evaluation of machine learning algorithms for urban pattern recognition from multi-spectral satellite images. *Remote Sensing*, *6*(4), 2912–2939. http://doi.org/10.3390/rs6042912

Witten, D. (2015). Penalized Classification using Fisher's Linear Discriminant. Retrieved from http://cran.r-project.org/package=penalizedLDA

Witten, D., & Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *73*(5), 753–772. http://doi.org/10.1111/j.1467-9868.2011.00783.x

Zhang, X., Friedl, M. A., Schaaf, C. B., Strahler, A. H., Hodges, J. C. F., Gao, F., … Huete, A. (2003). Monitoring vegetation phenology using MODIS. *Remote Sensing of Environment*, *84*(3), 471–475. http://doi.org/10.1016/S0034-4257(02)00135-9

Zhang, Z., Dai, G., Xu, C., & Jordan, M. I. (2010). Regularized Discriminant Analysis, Ridge

Regression and Beyond. *Journal ofMachine Learning Research*, *11*, 2199–2228.

Zheng, B., Myint, S. W., Thenkabail, P. S., & Aggarwal, R. M. (2015). A support vector machine to identify irrigated crop types using time-series Landsat NDVI data. *International Journal of Applied Earth Observation and Geoinformation*, *34*, 103–112. http://doi.org/10.1016/j.jag.2014.07.002

Zhong, L., Hawkins, T., Biging, G., & Gong, P. (2011). A phenology-based approach to map crop types in the San Joaquin Valley, California. *International Journal of Remote Sensing*, *32*(22), 7777–7804. http://doi.org/10.1080/01431161.2010.527397