EARLY YIELD ESTIMATION BY PHOTOSYNTHETIC PIGMENT ABUNDANCES USING LANDSAT 8 IMAGE SERIES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

AYŞENUR ÖZCAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
GEODETIC AND GEOGRAPHIC INFORMATION TECHNOLOGIES

OCTOBER 2020

Approval of the thesis:

EARLY WHEAT YIELD ESTIMATION AT FIELD-LEVEL BY PHOTOSYNTHETIC PIGMENT ABUNDANCES USING LANDSAT 8 IMAGE SERIES

submitted by **AYŞENUR ÖZCAN** in partial fulfilment of the requirements for the degree of **Doctor of Philosophy in Geodetic and Geographic Information Technologies, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, **Graduate School of Natural and Applied Sciences** _____

Prof. Dr. Zuhal Akyürek
Head of the Department, **Geodetic and Geographic Information Technologies, METU** _____

Prof. Dr. M. Lütfi Süzen
Supervisor, **Geodetic and Geographic Information Technologies, METU** _____

Assoc. Prof. Dr. Uğur Murat Leloğlu
Co-Supervisor, **Geodetic and Geographic Information Technologies, METU** _____

Examining Committee Members:

Prof. Dr. Zuhal Akyürek
Civil Engineering Department, METU _____

Prof. Dr. M. Lütfi Süzen
Geological Engineering Department, METU _____

Assoc. Prof. Dr. Sinan Kalkan
Computer Engineering Department, METU _____

Assoc. Prof. Dr. Ali Özgün Ok
Geomatic Engineering Department, Hacettepe University _____

Asst. Prof. Dr. Emre Sümer
Computer Engineering Department, Başkent University _____

Date: 13.10.2020

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Ayşenur Özcan

Signature :

ABSTRACT

EARLY YIELD ESTIMATION BY PHOTOSYNTHETIC PIGMENT ABUNDANCES USING LANDSAT 8 IMAGE SERIES

Özcan, Ayşenur
Doctor of Philosophy, Geodetic and Geographic Information Technologies
Supervisor : Prof. Dr. Lütfi Süzen
Co-Supervisor: Assoc. Prof. Dr. Uğur Murat Leloğlu

October 2020, 88 pages

Timely estimation of crop yields is critical for monitoring global food production by international organizations as well as governments, farmers and the private sector dealing with storage, import and export of crops and associated products. Satellite remote sensing has the capability to provide near real-time information on a global scale. Combining satellite data and soft computing techniques to predict crop yields is a very effective strategy for continually forecasting crop yields. This thesis presents a novel approach for accurate and sustainable estimation of crop yields based on estimated abundances of endmembers that may be attributed to photosynthetic pigments. Landsat 8 images acquired during the time of the phenological cycle when plants have maximum greenness are the inputs to find endmembers and abundances within the pure wheat crop pixels using Robust Collaborative Nonnegative Matrix Factorization (R-CoNMF) unmixing algorithm. The endmembers are optimized to maximize the predictive power of the abundances for the yields. Wheat yields were then estimated with the four abundances, their relevant interactions, ten important agrometeorological

parameters, including parameters proposed in this thesis for the first time, and four different vegetation indices using three different machine learning algorithms (Generalized Linear Model (GLM), Artificial Neural Network (ANN) and Random Forest (RF)). Harvester records from 142 wheat fields distributed in 31 provinces of Turkey were used as the ground truth for testing the algorithm. In the literature, the coefficient of determination (R^2) is used as a proxy to show how good the relationship is between the estimated and real figures. According to these calculations, the yields were estimated with 64% accuracy when only the abundances were used in the GLM algorithm, 78% accuracy when ANN was used for yield estimation and 82% accuracy was reached when applying RF to all of the parameters. The similarity of the endmembers to photosynthetic pigment spectral signatures along with their predictive power suggested their relevance to the pigments. Although the R-CoNMF algorithm performs a linear unmixing of the intimate mixture of the photosynthetic pigments, the interactions of the abundances used in the endmember optimization and in classifications partially handle the non-linearity using the bilinear model. These results can be considered as a great success when using multispectral satellite data only and are recognized as a clear indication that much better results would be achieved while using images from future hyperspectral space missions like HyspIRI.

Keywords: Landsat 8; time series, yield estimation; random forest; artificial neural network; Generalized Linear Model; photosynthetic pigments; unmixing; R-CoNMF; endmember optimization, endmember extraction

ÖZ

LANDSAT 8 GÖRÜNTÜ SERİSİ KULLANILARAK FOTOSENTETİK PİGMENT BOLLUKLARI İLE ERKEN VERİM TAHMİNİ

Özcan, Ayşenur
Doktora, Jeodezi ve Coğrafi Bilgi Teknolojileri
Tez Yöneticisi: Prof. Dr. M. Lütfi Süzen
Ortak Tez Yöneticisi: Doç. Dr. Uğur Murat Leloğlu

Ekim 2020, 88 sayfa

Mahsul verimlerinin zamanında tahmin edilmesi, uluslararası kuruluşların yanı sıra hükümetler, çiftçiler ve mahsullerle beraber ilgili ürünlerin depolanması, ithalat ve ihracat ile ilgilenen özel sektör tarafından küresel gıda üretiminin izlenmesi için kritik öneme sahiptir. Uzaktan algılama, küresel ölçekte gerçek zamana yakın bilgi sağlama yeteneğine sahiptir. Mahsul verimlerini tahmin etmek için uydu verilerini ve bilgisayar programlarıyla hesaplama tekniklerini birleştirmek, mahsul verimlerini sürekli olarak tahmin etmek için oldukça etkili bir stratejidir. Bu tez, fotosentetik pigmentlerle ilişkili olabilecek tahmini son üye bolluklarına dayalı olarak mahsul verimlerinin doğru ve sürdürülebilir bir şekilde tahmin edilmesi için yeni bir yaklaşım sunmaktadır. Fenolojik döngüde bitkilerin maksimum yeşillikte oldukları sırasında çekilen Landsat 8 görüntüleri, sağlam işbirlikçi negatif olmayan matris çarpanlarına ayırma (R-CoNMF) karıştırma algoritması kullanarak saf buğday mahsul pikselleri içindeki son üyeleri ve bollukları bulmak için girdi olarak kullanılmışlardır. Son üyeler, verim için bolluğun tahmin gücünü en üst düzeye çıkarmak için optimize edilmiştir. Daha sonra buğday verimleri, dört bolluk değeri, bunların ilgili etkileşimleri, bu tezde ilk kez önerilen parametreler dahil on önemli agrometeorolojik parametre ve üç farklı makine öğrenme algoritması, yani

Genelleştirilmiş Doğrusal Model (GLM), Yapay Sinir Ağı (YSA) ve Rastgele Ormanlar (RF) kullanılarak tahmin edilmiştir. Türkiye'nin 31 iline dağılmış 142 buğday tarlasından hasat kayıtları, algoritmanın testinde yer kontrolü olarak kullanılmıştır. Literatürde determinasyon katsayısı (R^2), tahmin edilen ve gerçek rakamlar arasındaki ilişkinin ne kadar iyi olduğunu göstermek için bir temsilci olarak kullanılmaktadır. Buna göre, GLM algoritmasında sadece bolluklar kullanıldığında verimler %64 doğrulukla tahmin edilmiş, YSA ile tahmin yapıldığında %78 tahmin oranına ulaşılmış, ve ilgili tüm parametrelere RF uygulanırken ise %82 doğruluk seviyesine ulaşılmıştır. Son üyelerin kestirim güçleri ile birlikte fotosentetik pigment spektral imzalara benzerliği, pigmentlerle ilişkilerini göstermiştir. Her ne kadar R-CoNMF algoritması, fotosentetik pigmentlerin derin karışımının lineer çözülmesini gerçekleştirse de, son üye optimizasyonu ve sınıflandırmalarda kullanılan bollukların etkileşimleri bilineer model kullanılarak doğrusal olmama durumunu kısmen ele almaktadır. Bu sonuçlar sadece çok-bantlı uydu verileri kullanıldığında büyük bir başarı olarak değerlendirilebilir ve HyspIRI gibi gelecekteki hiperspektral uzay görevleri görüntüleri kullanılırken çok daha iyi sonuçların elde edilebileceğinin bir göstergesi olarak kabul edilebilir.

Anahtar Kelimeler: Landsat 8; zaman serileri, verim tahmini; rastgele orman; yapay sinir ağı; Genelleştirilmiş Doğrusal Model; fotosentetik pigmentler; ayrıştırma; R-CoNMF; son üye optimizasyonu, son üye çıkarma

Dedication

I dedicate this thesis to my son Alper, my mother Fatma, my father Nurullah and my sister Zehra, who have always been there for me no matter what. I cannot thank you enough for your support and patience.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my supervisor Prof. Dr. M. Lütfi Süzen and my co-supervisor Assoc. Prof. Dr. Uğur Murat Leloğlu for their guidance, advice, criticism, encouragements and insight throughout the research.

I also wish to thank Prof. Dr. Zuhal Akyürek and Assoc. Prof. Dr. Sinan Kalkan for their suggestions and comments.

I would like to thank the Ministry of Agriculture and Forestry for providing the vitally important yield data that made this thesis happen.

TABLE OF CONTENTS

ABSTRACT v

ÖZ vii

ACKNOWLEDGMENTS x

TABLE OF CONTENTS xi

LIST OF TABLES xiii

LIST OF FIGURES xiv

LIST OF ABBREVIATIONS xvi

CHAPTERS

1 INTRODUCTION 1

1.1 Description of the problem 2

1.2 The Approach of this thesis 3

1.3 Contributions of this thesis 3

1.4 Organization of this thesis 4

2 BACKGROUND AND LITERATURE SURVEY 5

2.1 Use of Remote Sensing in Agricultural Applications 6

2.2 The photosynthetic pigments 9

2.3 Unmixing of satellite image pixels 13

2.4 Yield estimation by remote sensing 15

2.5 Soft computing for yield estimation 18

2.6 Photosynthetic pigments in yield estimation 23

3 MATERIALS AND METHODS 27

3.1 The data and the study area 27

3.1.1 Yield data 27
3.1.2 Meteorological data 30
3.1.3 Satellite data 31
3.2 Methodology 32
3.2.1 Preparation of the data 34
3.2.2 Spectral Unmixing 34
3.2.3 Linear and non-linear regression 36
4 RESULTS AND DISCUSSION 39
4.1 The endmembers 39
4.2 The abundances 48
4.3 Parameter selection and interactions 50
4.4 Yield estimation using three different machine learning approaches 60
4.4.1 The Generalized Linear Model (GLM) approach 60
4.4.2 The neural network approach 61
4.4.3 The random forests approach 62
5 CONCLUSION 69
REFERENCES 71
CURRICULUM VITAE 87

LIST OF TABLES

Table 3.1 Landsat 8 satellite bands used in this study and their properties 32
Table 4.1 Parameters and their values used to perform R-CoNMF..... 40
Table 4.2 Estimated absorbance and reflectance values of Virginia creeper leaf reported as graphs in Gitelson and Solovchenko 2018..... 45
Table 4.3 Names and abbreviations of all the parameters and interactions..... 51
Table 4.4 Names and abbreviations of all the selected parameters and interactions 56
Table 4.5 The real vs. predicted yield accuracies and RMSE, of the applied methods: GLM, ANN and RF according to different parameter combinations for the training sets. 64
Table 4.6 The real vs. predicted yield accuracies and RMSE, of the applied methods: GLM, ANN and RF according to different parameter combinations for the test sets. 64

LIST OF FIGURES

Figure 2.1. Absorption spectra of the most important plant pigments (Blackburn, 2006). 11
Figure 2.2. Absorption spectra of Chlorophyll A, Chlorophyll B and Carotenoids, interpolated from data samples from plots published in Lichtenthaler, 1987 13
Figure 3.1. GPS data of the harvester forming yield grids overlaid on Google earth base image. The image on the right is the enlarged image of the left one to visualize the speed information of the harvester. 28
Figure 3.2. Locations of the 142 fields spread around 31 provinces of Turkey. The blue circles represent the meteorology stations and the red circles stand for the fields. 29
Figure 3.3 The flowchart of the proposed algorithm. 33
Figure 4.1 Mean square error, projection error and noise power used for the implementation of R-CoNMF as a function of number of endmembers (k) 41
Figure 4.2 Projection of the spectral vectors on the endmembers shown on the first two principal components. Projection of the spectral vectors $\bar{\boldsymbol{y}}_i$, for i = 1, . . . ,n (blue), (where $\bar{\boldsymbol{y}}$ is the sample mean vector); of the endmember signatures $\boldsymbol{m}_i$, for i = 1, . . . ,p (magenta); and of the columns of $\boldsymbol{A}$, which are not endmembers (green). The spectral mean value is shown in black. 42
Figure 4.3 The transformation of the absorbance spectra of photosynthetic pigments to reflectance spectra. 44
Figure 4.4 The reflection spectra of the major pigments obtained from the absorption spectra in Figure 2.2. 46
Figure 4.5 (a) The calculated pigment reflectance of the first four bands of Landsat 8; (b) The endmembers found from R-CoNMF algorithm and optimization. 47
Figure 4.6 Real yields vs. abundances with R^2s. Each dot represents one agricultural field. 49
Figure 4.7 Out-of-bag importance of all parameters. 53
Figure 4.8 Predictor importance estimation comparison. 54

Figure 4.9 Predictor association estimates of all parameters. 55
Figure 4.10 Out-of-bag importance of selected parameters. 57
Figure 4.11 Predictor importance estimation comparison. 58
Figure 4.12 Predictor association estimates of all parameters. 59
Figure 4.13 Relationship between real and predicted yields found by using all the parameters of Table 5.3 in GLM algorithm (R^2 = 0.67). 61
Figure 4.14 Relationship between real and predicted yields found by using all the parameters of Table 4.5 in ANN algorithm (R^2 = 0.78). 62
Figure 4.15 The relations between real and predicted yields using abundances and their interactions and selected agrometeorological parameters in RF (R^2=0.82). .. 63

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
GLM	Generalized Linear Model
NDVI	Normalized Difference Vegetation Index
MSE	Mean Squared Error
OOB	Out-of-bag Error
R-CoNMF	Robust Collaborative Nonnegative Matrix Factorization
RF	Random Forests
VIs	Vegetation Indices

CHAPTER 1

INTRODUCTION

Yield estimation of crops is one of the most popular subjects in the literature of many fields such as remote sensing, agriculture, geographic information systems (GIS), economics and maybe many more. Researchers are willing to find accurate estimates of crops so that countries can plan ahead their economies by taking into account these productions, i.e. the money to be reserved for import of goods that cannot be produced within the country or how much they can export products such as wheat, rice, tea, sugar, etc. according to the agreements between them and other countries, as well as the employment opportunities it provides to the population. Especially in developing countries, agriculture is the main source of livelihoods and even if there cannot be agricultural work all year, it is a source of income that contributes to the economy of the households throughout the whole year, even if they only work seasonally.

Agriculture is vital for the economic and social well-being of countries, regardless of their level of development. It is essential to estimate the yield before harvest across large areas because governments can implement their agricultural policies and plans based on these data. Agriculture also has contributions to the transportation sector as the goods are usually transported by road or railways, as well as to marketable surplus. A stable agriculture of a nation leads to the stable food security. So that malnourishment is prevented and people are healthy. A timely yield estimate also helps the private sector to plan for the storage, import and export of crops and related goods, international organizations to monitor the world's food production and the farmers to plan their next crop and order the appropriate seed quantity in advance.

Humanitarian response relies heavily on agriculture and its products. Therefore, humanitarian organisations invest in GIS and remote sensing in terms of technology and human power just to have a close enough idea on the situation of crops and production information in order to serve the people in need in a timely manner. Time is of the essence in life and in all sectors that have a direct impact on human lives.

Therefore, it is very important to estimate the yield as early as possible to help the involved parties take all the measures of precaution and decide on their roadmap for the planning of their near future.

1.1 Description of the problem

Accurate and timely estimates of yield increase the overall efficiency of the agricultural system. Since the satellite images have been put to the use of researchers, many researches have been done and many models have been presented to predict the crop yields accurately and before the harvest. However, there has not been a significant success in this field so far. The main reasons for this unfortunate result may be listed as the lack of information from satellite images due to the high percentage of cloud cover, the low temporal resolution of most of the freely available images and the technologies to be used in these studies not having been developed fast enough. These issues reduce the efficacy of the established crop yield estimation techniques and make them counterproductive.

To solve these problems, there have been attempts to use high temporal resolution and low spatial resolution images, using agrometeorological parameters along with various indices derived from remote sensing instruments, but researchers still could not achieve sufficient precision in yield estimates. In parallel, soft computer algorithms such as Artificial Neural Networks (ANN), Fuzzy Logic (FL), Genetic Algorithm (GA) and Random Forests (RF) have been used to achieve more accurate estimation results in order to increase the success of conventional

computing techniques. Although these algorithms helped to move a big step forward, a new point of view seemed like a necessity to overcome the problem of timeliness and accuracy when estimating the yields.

1.2 The Approach of this thesis

As a new point of view to seek a solution to the accurate yield estimation problem, photosynthetic pigments, namely, chlorophylls, xanthophylls and anthocyanins were investigated. This thesis describes an algorithm applied to satellite images to extract endmembers that possibly correspond to or at least relate to photosynthetic pigments in plants at the maximum NDVI value when the canopy closure of wheat crops is assumed to be 100%. The abundances of these endmembers within the crops and some indices derived from them were calculated and used as inputs for Generalized Linear Model (GLM), ANN and RF to estimate yields at least one month before harvest. Agrometeorological parameters, including new parameters proposed for the first time in this thesis, and/or vegetation indices (VIs), are also used as inputs of GLM, ANN and RF along with the abundances to determine their contribution to the yield estimation. The performance of the algorithms is tested using ground truth data obtained by harvesters. It is shown that, although the mixing of pigments is most presumably non-linear, the abundances of the endmembers, which are probably related to photosynthetic pigments, are useful in yield estimation. Some interactions of abundances are also proven to be good predictors that probably handle the non-linearity inherent in intimate mixtures partially.

1.3 Contributions of this thesis

This thesis contributes to the expansive research literature of yield estimation of crops via remote sensing, in a way that

- improves the timeline of estimation of the yield, as the yield can be estimated at least a month before the harvest;
- introduces a novel point of view in approaching the yield, i.e. by taking into account the three photosynthetic pigments, namely chlorophyll a, chlorophyll b and carotenoids all together in the methodology of the estimation, and
- reduces the number of sources and parameters that are used and focuses on the ones that are easily accessible or can be calculated with no additional cost. For example, Landsat 8 data, which are free of charge, an unmixing code and agrometeorological data that are also found free of charge, were used to conduct all the study.

1.4 Organization of this thesis

This thesis is organized as follows: In Chapter 2, a background of all the technologies used in this thesis are given. The subjects that are mentioned in that chapter are the basic history of satellite remote sensing, the photosynthetic pigments that exist in a crop and the brief history of unmixing algorithms. A literature review occupies Chapter 3 to create a more detailed point of view on how yield estimation has been performed using remote sensing until recently, the part of soft computing in these calculations and a brief explanation of the role of photosynthetic pigments in yield estimation by referencing the researchers that have contributed to the literature regarding these subjects. In Chapter 4, the materials and methods are introduced, starting with the data used and the study area, followed by the methodologies implemented in the study. Chapter 5 is designed to give the results and discuss the outcomes. The thesis is finalized with Chapter 6, conclusion and future work suggestions.

CHAPTER 2

BACKGROUND AND LITERATURE SURVEY

Agriculture is the key in the development of human civilization, as farming of domesticated species created food surpluses that enabled people to live in cities. The history of agriculture began thousands of years ago. Agriculture is still vital in the economic and social beings of countries, regardless of their level of development. There seems to be no substituent source available for this sector that produces raw materials and food that are necessary for human nutrition. Therefore, it is normal for the sector to have a big share in the total employment. With the rising of the incomes and growth of trade in the world, consumption per capita increases. According to these data, agricultural production is expected to be able to increase slowly but steadily in the following years.

It has become very important to make estimations on the growth process of crops and therefore the yield before harvest, or at least to be able to make accurate predictions of yield at harvest. In order to make these predictions, the area of cultivated crop should be monitored. The estimations on crop yields over large areas are important as the governments implement their agricultural policies and plans according to these data. The timely estimations of yields also help the planning of the private sector dealing with storage, import and export of goods, etc. The farmers can plan their next cultivation and order the appropriate amount of seeds beforehand. If done correctly, it will also be a good input for international organizations dealing with monitoring of the world food production, while the adjustment of storage, import, export, etc. according to yield increases the total efficiency of the system of the companies in agricultural sector. It will undoubtedly be beneficial to farmers as they can know their yield, manage their income

beforehand. All these facts emphasize the importance of early estimation of yields for each and every stakeholder.

2.1 Use of Remote Sensing in Agricultural Applications

Satellite monitoring was the starting point of crop monitoring, which started in the 1970s in developed countries. The United States (US) started monitoring its own wheat production, and then extended its studies to monitoring many other countries main crop production at the end of 1980s. Following the US, European Union built its own prediction and monitoring system at the Joint Research Center (JRC) (Monitoring Agricultural ResourceS (MARS) | EU Science Hub) with the name of MARS (Monitoring Agriculture with Remote Sensing). Other countries like France, Germany, Russia, Canada, Japan, India etc. pursued the US and EU-JRC in building their own monitoring and forecasting systems. Usually the NOAA/AVHRR and afterwards Landsat satellite systems were used. A substantial amount of accurate data collection from the field (ground truth) had to be done due to the low ground resolution of these satellites at the times.

Remote sensing in agriculture has usually been about the plant reflecting the radiation coming from the sun, measured by passive sensors. However, there are also studies on the transmittance, absorbance or emittance of the plants. Plants emitting energy for both photosynthetic function and biochemical constituent is known as fluorescence sensing (Apostol et al., 2003). Thermal remote sensing is about variations in the evaporation rate based on the response of the temperature of the plant to the emission of radiation, which leads to the information on water stress (Cohen et al., 2005). Absorption of plants is the opposite of reflection of plants and therefore also varies with the incident wavelength. It was found that the chlorophyll of plants absorb radiation at 400-700nm of the Electromagnetic spectrum (EMS), namely the visible region while reflectance is high in the near infrared region (700-1300nm) (Pinter et al., 2003). The sharp contrast between red and NIR parts of the spectrum was the motivation for the development of some

spectral indices (Mulla, 2013). Simple and complex spectral indices are able to detect variations in leaf area index (LAI) and variations in crop status such as chlorophyll and nitrogen content (Wang et al., 2014).

The indices are usually used for the estimation of yields of various crops. Yield estimation of crops, especially before harvest, plays an important role in agricultural policies and decision making. In the literature, yield estimation is usually done by using different models. The first type of these models, the Statistical Models are usually used when there is information on large areas and they can estimate in wide ranges. Statistical models are not recommended for accurate or near-accurate estimations. In crop yield forecasts with statistical regression, which is considered a common and easy method (Lobell et al., 2010), the basic principal is that a simple matrix is formed with some agrometeorological parameters (not too many) and previous yields, then a regression equation is derived between yields and the parameters. Usually the regression model is selected as multiple linear regression (MLR), however this model gives inaccurate and unstable solutions especially when large number of parameters are used and if there are correlations between these variables (Lobell et al., 2005). Magney et al. (2017) aimed to evaluate the usefulness of RapidEye spectral VIs to predict cumulative Nitrogen (N) uptake in wheat and to examine the usefulness of remotely sensed N uptake maps for precision agriculture (PA) applications. It was concluded that the top performing Vegetation Index (VI) was the Normalized Difference Red-Edge index (NDRE). They used seventeen commonly used spectral VIs to report that VIs from RapidEye imagery can be used for estimating wheat N uptake. Polynomial fit showed maximum R^2 of 0.81.

The second widely used model is the Mechanistic Model. These models are much more detailed than statistical models. They use fundamental mechanisms of soil and plant processes (Dourado-Neto et al., 1998). Third model is Functional Model, which is a more complex model and it is able to simulate models on data that are updated daily. Its functionality comes from simplifying the complex processes.

However, if not developed correctly, it may give less accurate results than mechanistic models (Watt, 2013).

In a more general way, models can be classified as deterministic and stochastic. Deterministic models make the assumption that all plants and soil are uniform throughout the space. Stochastic models have a more realistic approach, knowing that the parameters are changeable and the results may also produce some uncertainties due to soil properties, weather conditions, biotic and abiotic factors. These properties cause the model to be valid in small areas rather than large areas. Also the crop growth system is more stochastic than deterministic, because most parts of the agro-ecosystem are heterogeneous (Basso et al., 2014).

Remote sensing is a very efficient way to sample large number of plants at the plant scale to examine for example, plant breeding and to identify some specific physiological characteristics of varieties (Jones & Vaughan, 2010). It can also be noted that studies have been done on remote sensing for precision agriculture to analyze variations of parameters within fields (Plant, 2001), however there seem to be not many on between-field variations across the landscape. (Lobell et al., 2005) have pointed out the three advantages of yield remote sensing over ground based approaches as:

1- The sample sizes can increase once the field measurements of yield are bypassed.
2- Field measurements are collected via sensors and a small number of plots within fields which leads to sampling errors of only within-field variability, while with remote sensing yield estimates of a much wider range of spatial scales can be done.
3- There is a huge archive of remote sensing images which can help analysis of past surveys that may not have measured yield.

When dealing with complex systems, using conventional methods may not be cost-effective, analytical or provide complete solution. Thankfully, some ‘inexact’ methods have been developed to model and analyze the complex problems of these

complex systems. These ‘inexact’ computing techniques are referred to as ‘soft computing’ (Huang, et al. ,2010).

In the past years, the agrometeorological parameters along with various indices derived from remote sensing instruments have been used as predictors to perform yield estimation. As the greenness and soil driven indices could not achieve sufficient precision in yield estimates, photosynthetic pigments, chlorophylls, xanthophylls and to some extent anthocyanins were investigated. In parallel, soft computing algorithms such as Artificial Neural Networks (ANN), Fuzzy Logic (FL), Genetic Algorithm (GA) and Random Forests (RF) have been used to achieve more accurate estimation results in order to increase the success of conventional computing techniques. The developments in these fields so far will be examined in detail in the literature survey chapter.

2.2 The photosynthetic pigments

Many pigments exist in the structure of vegetation. The main pigments that exist in all types of vegetation are chlorophylls, carotenoids and flavonoids (mainly anthocyanins) (Lachman et al., 2017; Blackburn, 2006), which are also considered as photosynthetic pigments (accessory pigment or antenna pigment).

Chlorophylls are the most important pigments for the life cycle of all living things as they play the most important role in photosynthesis. Chlorophyll concentrations are play an active role in primary production of crops due to their control upon the solar radiation that the leaves absorb, leading to photosynthetic potential. Besides that, chlorophylls assemble a big portion of the leaf nitrogen content which is a measure of the plant nutrient status. Chlorophylls are light-dependent pigments and their amount decreases in low or no light environments, under stress and during senescence.

Carotenoids are one of the other important pigments that exist in the chloroplast of the plants. About 600 different carotenoids that are discriminated as xanthophylls,

most important of which is lutein (containing oxygen) and carotenes (containing hydrocarbons and no oxygen) exist. The carotenoid content determines the quality of the durum wheat, by giving it the yellow colour of the pasta (Lachman et al. 2017). Carotenoids absorb blue light strongly which leads to the conclusion that they have a dual function of absorbing photosynthetic energy (a.k.a. incident radiation) and contribute it to help chlorophylls and photo-protection of chlorophyll when exposed to excess light (Bartley and Scolnik, 1995). The first function is due to carotenoids (just as chlorophyll b is) being an accessory pigment for chlorophyll a, which takes energy from the antenna (accessory pigments). if the concentration of chlorophyll a is high, it will take more energy from chlorophyll b and carotenoids, which will result in higher photosynthetic activity, thus primary production (Chappelle et al., 1992). In case of excess radiation, carotenoids disperse the energy in xanthophyll cycle, making the xanthophyll pigment directly linked to the photosynthetic light use efficiency (LUE).

Anthocyanins are the third group of important pigments in plants. They belong to the group of flavonoids. They are the least examined pigment group in the field of remote sensing, therefore there is an uncertainty regarding their functions. Various roles of anthocyanins have been reported, such as being an antioxidant (Yamasaki et al. 1997), in case of oxidative stress, they reduce the excitation pressure and prevent oxidative damage (Field et al., 2001). Besides these, Steyn et al. (2012) found the photo-protective light screen potential of anthocyanins. It is seen that anthocyanins can alter the light environment within a leaf and adjust photosynthesis in a way that they limit photo-inhibition and photo-bleaching (Barker et al., 1997).

The spectral, spatial and temporal dynamics of these very important pigments in vegetation can not only provide scientific knowledge, but also significant help in agricultural and/or environmental management. According to the literature however, the spectral discrimination of these pigments are possible through multiple regression, stepwise regression, nonlinear approaches, PCA or ANN approaches and the combination of some of these, but not solely through linear

approaches as they have a structure of observably overlapping spectral signatures within a plant (Figure 2.1).

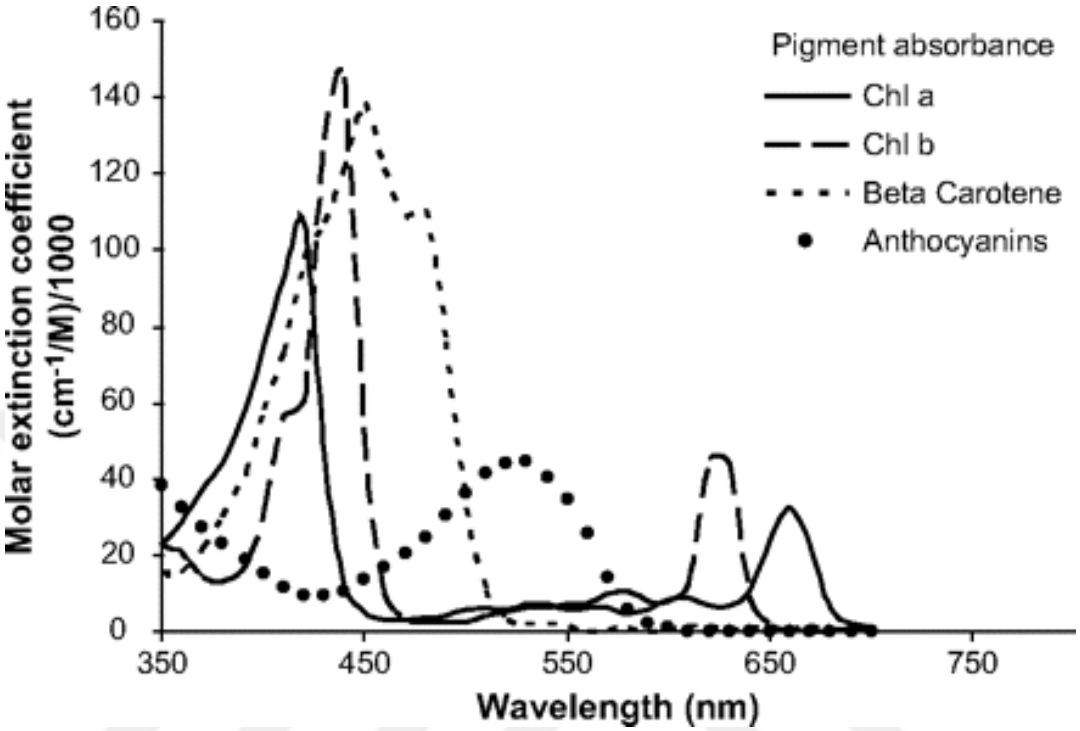

Figure 2.1. Absorption spectra of the most important plant pigments (Blackburn, 2006).

The most well-known carotenoid index is Photochemical Reflectance Index (PRI) (Gamon et al., 1992). It is a narrow-band index which gives the changes in the epoxidation of xanthophyll pigments, which can also be expressed as the changes in the efficiency of the photosynthetic light reactions (Jones and Vaughan, 2010). In other words, it shows the photosynthetic light-use efficiency (LUE) and works as an indicator of stress (Gamon J., 2010). PRI is formulated as:

$$PRI = (R_{531} - R_{ref})/(R_{531} + R_{ref})$$

(1)

where R_{531} and R_{ref} are leaf reflectance values at 531 nm and the reference wavelength. The 531 nm, although mostly taken literally in the scientific community, there were times when it was taken in the range between 505 nm and

535 nm (Bilger et al. 1989), 531 nm and 535 nm (Moraleset al., 1990; Ruban et al., 1993). The reference reflectance is usually taken as 570 nm, however it can be 550, 560 nm etc. It was found that the xanthophyll pigments absorb minimum at 531 nm and the reference wavelength can be chosen as where they make a peak. The index is aimed to minimize the effect of the diurnal sun angle changes (Gamon et al., 1992), meaning that PRI is more sensitive to environmental changes, parallel to xanthophyll epoxidation state and its effects to the crop in the shortest time scale than NDVI (Penuelas et al., 1994).

The estimation of leaf carotenoid content from reflectance was also investigated by Lichtenthaler (1987) and Sims and Gamon (2002), which is much more difficult than estimation of chlorophyll due to the overlap between the chlorophyll and carotenoid absorption peaks (Figure 2.2) and because of the higher concentration of chlorophyll than carotenoid in most leaves.

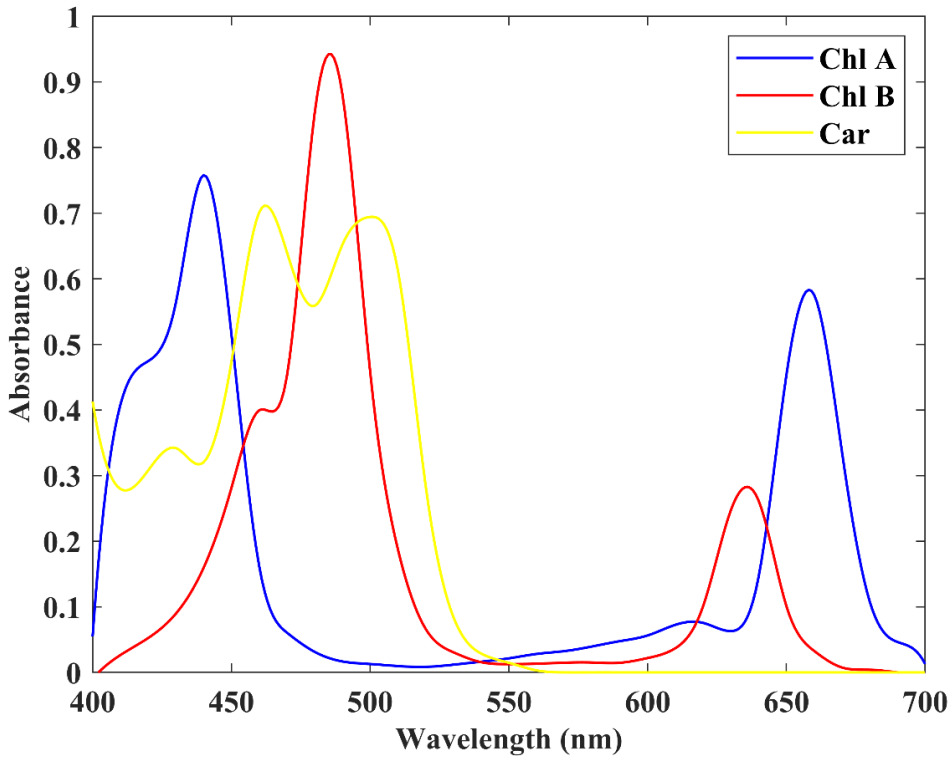

Figure 2.2. Absorption spectra of Chlorophyll A, Chlorophyll B and Carotenoids, interpolated from data samples from plots published in Lichtenthaler, 1987

Consequently, reflectance indices have proved more successful for the estimation of the ratio of carotenoid to chlorophyll, than in the estimation of the absolute carotenoid content (Penuelas et al., 1995; Merzlyak et al., 1999).

Multiple indices have been developed using band ratios near (not on) the absorption peak wavelengths of a certain pigment, usually chlorophyll (Zhang, 2011). Numerical inversion of leaf-level Radiative Transfer (RT) models, such as PROSPECT and LEAFMOD, has demonstrated success for predicting leaf chlorophyll content (Jacquemoud and Baret, 1990; Ganapol et al., 1998; Maier et al., 1999; Demarez, 1999; Renzullo et al., 2006). Numerical inversion techniques offer the potential of a generically superior approach to estimate leaf chlorophyll content from hyperspectral data than spectral indices and other approaches that are based on empirical calibrations.

Not much research has been done on separating anthocyanins from the total spectra of plant pigments. (Gamon and Surfus, 1999) examined the red region where the anthocyanins were absorbing light and created a red/green index. However, the tests of this index showed no relationship with anthocyanins. Similar studies to Gamon's, (Neill and Gould, 2000) have pointed out the problem as the existence of chlorophyll obscuring with the reflectance spectra discrimination of anthocyanins. Gitelson overcame this problem by creating a narrowband index at 550nm and 700nm (Gitelson et al., 2001). This new index proved to be successful over different types of plants.

2.3 Unmixing of satellite image pixels

The pixel sizes of multispectral sensors are big enough to contain varying materials in them, the extraction of the ratios of desired materials became important for the purposes of research. The standard extraction algorithm is called Spectral

Unmixing Algorithm (Keshava and Mustard, 2002). In linear spectral unmixing, basically the mixed pixel is assumed to be consisting of a set of constituent spectral signatures (endmembers) weighted by the subpixel fractional cover (abundances). The unmixing algorithms are designed and applied to hyperspectral images in the literature, and it should be noted that all the literature of the researchers explained below have applied their algorithms to hyperspectral imagery.

Crop fields ideally consist of only photosynthetic pigments, which are intimately mixed within the vegetation. Normally, endmember extraction, when not done at large scenes of satellite images, but with intimate mixtures, requires the use of complex non-linear techniques. Therefore, extracting endmembers from an intimate mixture of photosynthetic pigments requires techniques other than linear ones, whereas it should also be easily implemented by users, which may not be the case for non-linear techniques. The endmembers are the inputs for the next step of the process, generating the abundances, which are the inputs of the method for predicting the yield.

The standard unmixing algorithms in the literature, perform the unmixing in a considerably big area of an image. A new approach, which is also linear but different than the classical linear unmixing, called PCOMMEND was introduced by (Zare et al., 2013). The method was programmed to find multiple sets of endmembers which has shown it to be a better representative of hyperspectral imagery. Different from the standard models, PCOMMEND has the ability to execute iterative fuzzy clustering process while conducting spectral unmixing at the same time in order to partition a mixture (pixel) into multiple regions of the space defined by the endmembers. In each of these regions, a distinct set of endmembers that define a simplex occurs. This makes all the pixels in the image be represented by a union of all the simplices. Therefore, it could be possible for the pigments to be linearly separated within these small regions. Despite running a complex algorithm in the background, PCOMMEND could be conducted very easily with satisfying results. The endmember signatures were extracted as two sets of three distinct endmembers giving a total of six endmembers. They also showed

that the algorithm was proven useful when used on Landsat TM image, having limited number of wavebands.

A robust collaborative nonnegative matrix factorization (R-CoNMF) (Li et al., 2016) was used as an alternative to the other methods as it had the ability to find the actual number of endmembers instead of the user having to guess and then find the abundances accordingly.

In a multivariate system, the first assumption is that the variables are linearly related. However, in some situations it could be theoretically possible that a second predictor variable Z is itself moderating the influence of a predictor variable X on a criterion variable Y. Apart from the linear effects β_1 and β_2 an interaction effect β_3 becomes a part of the model structure here. To evaluate the interaction effect in combination with the linear effects in the regression equation, a new variable must be formed, i.e. the product XZ between the predictors X and Z, to be included in the multiple regression equation as third term (Dimitruk et al., 2007).

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon \quad (2)$$

Here, in Equation 2 that was first presented by Kenny and Judd in 1984 (Kenny and Judd, 1984), Y is the criterion variable, X and Z are the predictor variables, XZ is the interaction term, β_0 is the intercept, β_1 and β_2 are the linear effect terms, β_3 is the interaction effect, and ε is the error.

Interaction terms proved to be very helpful in solving multivariate non-linear problems. In the mixing phase nonlinear interactions are rarely controlled by material distortions, but rather by the non-linear interactions between them (Klein and Moosbrugger, 2000; Suzuki et al., 2009).

2.4 Yield estimation by remote sensing

Remote sensing in agriculture has usually been about the plant reflecting the radiation coming from the sun, measured by passive sensors. Radar signals also

interact with vegetation, but in a rather complicated way, and the signal penetration depends on the strength and water content of the vegetation. Return radar signals can involve trunks, stems and leaves as well as ground surfaces. Because such kind of landscape description details can only be obtained with this sort of instrument, SAR data is also used for plant studies. It is well-known that the chlorophyll of plants absorb radiation at 400-700nm of the electromagnetic spectrum, namely the visible region (Pinter et al., 2003) while reflectance is high in the near infrared region (700-1300nm). The development of spectral indices was due to the sharp contrast between red and NIR parts of the spectrum (Mulla, 2013). Simple and complex spectral indices are able to detect variations in Leaf Area Index (LAI) and variations in crop status such as chlorophyll and nitrogen content (Wang et al., 2014).

The Normalized Difference Vegetation Index (NDVI) was found useful in predicting yield forecasts (Benedetti and Rossini, 1993; Groten, 1993). Although NDVI is used very widely in vegetation studies, it has some limitations such as the intervention of soil at low crop densities and saturation in mature crops with LAI greater than two or three, since the red light absorption peak of the leaves is reached at these LAI values (Thenkabail et al., 2000). Some new indices such as Soil Adjusted Vegetation Index (SAVI), Modified Soil Adjusted Vegetation Index (MSAVI) and Modified Triangular Vegetation Index (MTVI) have been proposed to overcome these problems. However, indices that can be useful with fewer constraints are always needed (Mulla, 2013).

The indices are often used to estimate yields of various crops, which plays an important role in agricultural policy and decision-making, especially if they are available well before harvesting. In remote sensing, statistical models are frequently used to estimate yields when information is available in large areas. Statistical crop yield forecast regression, which is considered to be a common and easy method (Lobell et al., 2010), the basic principle is that a simple matrix is formed with some relevant agrometeorological parameters and previous yields, and the relation between yields and parameters is derived from regression. The

regression model is usually selected as Multiple Linear Regression (MLR), but this model provides inaccurate and unstable solutions, especially when a large number of parameters are used and if these variables are correlated (Lobell et al., 2005). The effects of correlation and important yield factors in GLM are studied extensively (Kravchenko and Bullock, 2000; Park et al., 2005; Gutiérrez et al., 2008; Huang, et al., 2010).

Remote sensing is a very efficient way to classify large number of plants at a large scale in order to examine plant breeding and to identify certain specific physiological characteristics of varieties (Jones and Vaughan, 2010). (Lobell et al., 2005) have highlighted some advantages of remote sensing yield estimation over ground-based approaches and highlighted that there is an extensive archive of remote sensing images that can help to analyse past surveys that may not have measured yield.

(Lobell et al., 2003; Jiang, P., Thelen, 2004; Fortin et al., 2010) estimated wheat yield by using only Landsat ETM+ images. They integrated their knowledge of crop phenology with multi-temporal imagery and used instantaneous estimates of canopy light absorption (fraction of Absorbed Photosynthetically Active Radiation - fAPAR) from satellite images to adjust a wheat growth model calibrated locally, which leads to an estimate of the yield at each pixel.

(Franch et al., 2019) proposed a new crop yield model based on Differential Vegetation Index (DVI). They used Landsat 8 and MODIS time series data to perform wheat signal unmixing from the signal of other surfaces. After the analysis of the unmixed wheat time-series, regression equations were used as the yield estimation models for each administrative unit and they found an R^2 of 0.86 at the national level and 0.70 at the subnational level in the US and Ukraine.

There are also some studies that enhance and use predefined yield estimation models. (Wang Y. et al., 2019) improved CASA NPP estimation model, which was based on the absorbed photosynthetically active radiation (APAR) and the light use

efficiency (ε) absorbed by vegetation, with HJ-1A/B and MODIS data to find 56% accuracy in estimating the yield in selected places of China.

2.5 Soft computing for yield estimation

Conventional methods may not be cost-effective, analytical or provide a complete solution when dealing with complex systems. Fortunately, some 'inexact' methods have been developed to model and analyse these complex problems. These inexact computing techniques are called ‘soft computing’ (Huang et al., 2010).

Fuzzy Logic (FL), Artificial Neural Networks (ANN), Genetic Algorithms (GA), Bayesian Inference (BI) and Decision Trees (DT) are some of the most important soft computing technologies. The conventional methods of hard computing are stochastic and statistical. Soft computing techniques refer to nature and are therefore flexible and open to inaccuracy, uncertainty, partial truth and approximation. To improve the system and results, these techniques can be used separately or can be combined. In addition to these classic methods, Random Forests (RF) has gained considerable attention in recent years.

The soft computing methods used in this study are Random Forests (RF), Neural Networks (NN) and Generalized Linear Model (GLM). RF is a supervised learning algorithm that uses the ensemble learning approach for classification and regression. An ensemble approach combines the estimations from many multiple machine learning algorithms to improve the predictions. At training time, the RF, being a meta-estimator, builds several decision trees and generates the mean prediction (regression) of the individual trees. Individual decision trees tend to overfit. However, bagged decision trees combine the results of many decision trees, which reduces the effects of overfitting and improves generalization. Therefore, RF is prone to overfitting.

Neural Network is the most famous type of machine learning algorithms and it models itself after the human brain, allowing the computer to learn by incorporating new data.

GLM is a particular class of nonlinear models that describe a relationship between a response and predictors that is nonlinear. The model's structural form defines the patterns of interactions and associations. The model parameters include measurements of association intensity.

To estimate crop yield from satellite data, linear and non-linear models were developed and evaluated by (Sayago and Bocco, 2018). These models were proposed and applied using Landsat and SPOT images to obtain soybean and maize yield in the central region of Córdoba (Argentina). This study concluded that images of Landsat 8 and SPOT 5 can be used effectively to predict the yield of maize and soybean early to mid-season crop growth. The pixel size from Landsat 8 was adjusted to SPOT 5 in order to make Landsat 8 and SPOT 5 spatial resolutions comparable (each Landsat 8 pixel was divided into nine parts with the same attribute value). They used ANN and MLR methods to determine yields. The MLR results were almost as high as the ANN results (Soybean: R^2NN=0.9 R^2MLR=0.82, Corn: R^2NN=0.92, R^2MLR=0.88).

Integrated yield models combine agricultural meteorological and remote sensing data (Basso et al., 2001; Basso et al., 2007; Dorigo et al., 2007). The use of Principle Component Analysis (PCA) and Factor Analysis (FA) along with multiple regressions is an example of the integrated technique. Using integrated models, various soil nutrients were attempted to be predicted.

A wheat yield prediction model was developed and evaluated by (Pantazi et al., 2014). Fusion vectors, consisting of the values of eight soil parameters and historical yield data from past two years, collected with an online soil sensor and NDVI values computed from satellite imagery, were used as input to three ANNs for yield prediction. They used Self-Organizing Map Models (SOMs), namely, Counter-Propagation Artificial Neural Networks (CPANN), XY-Fused Networks

(XY-Fs) and Supervised Kohonen Networks (SKNs), incorporating the factor that limit the yield in a multi-layer fusion model in the presented approach. In order to estimate the LAI of a temperate meadow steppe in China, Wu, et al., (2015) developed two inversion models and compared them using the regression model and the Back-Propagation Neural Network (BPNN) model. The comparison results showed that the BPNN method (accuracy: 82.2%) outperformed Statistical Regression model (accuracy: 78.8%). The development of ANN models was described as a factual technique for predicting the yield of maize and soybean in nutrient management planning in Maryland by (Kaul et al., 2005). The results showed that the prediction of ANN yield was more accurate than the yield model based on MLR. The accuracies of ANN-based corn prediction varied between 77% and 90% while MLR results only showed 42% accuracy. (Kang and Özdoğan, 2019) disaggregated one country-level (US) maize yield data into 30m Landsat resolution yield map by using Ensemble Kalman Filter using LAI time series data estimated from Landsat images, EVI, meteorological and soil texture. They compared their results to farmer-reported yield data to find the correlation coefficient (R) to be 0.82 as a maximum value.

Due to its resistance to overfitting problems and the noise in the dataset, RF has gained well-deserved attention in recent years. It is almost unaffected by the multi-collinearity problem because it has the ability to ignore spatial autocorrelation. It can also be used to improve the performance of other methods, such as regression and kriging. As can be seen from the studies listed below, RFs actually outperform ANN in many cases and are not affected by the size of the dataset.

(Cai et al., 2019) combined climate, satellite (MODIS EVI) and chlorophyll fluorescence (SIF from GOME-2 and SCIAMACHY) data to compare LASSO regression model, SVM and RF performances. SVM outperformed the others with an R^2 reaching to 0.80. (Leroux et al., 2019) combined MODIS NDVI, MODIS LST and SMOS SSM with outputs of SARRA-O crop model to estimate the maize yield in their study area. Performances of 10-fold cross-validated GLM and RF

were compared to find that RF outperformed GLM and estimated 46% of the observed end-of-season yields two months prior to harvest.

In order to assess the accuracy of winter wheat yield, (Heremans et al., 2015) compared two regression tree methods, namely Boosted Regression Trees (BRT) and RF, using NDVI obtained from the SPOT-VEGETATION sensor along with meteorological variables and fertilization levels. The results showed that for both methods' R^2 was over 0.80 and that BRT was sensitive to noise, inclined to overfitting and considerably slower than bagging. RFs were comparable to boosting in terms of accuracy, but did not have the above limitations. It was also noted that the computational cost of RF was much lower than boosting. Li et al. (2016), produced accurate and timely predictions of grassland LAI, using various regression approaches and hybrid geostatistical methods. The results showed that the RF model provided the most accurate predictions for regression models such as Partial Least Squares Regression (PLSR), ANNs, RFs and Regression Kriging (RK). In Li, et al., 2016, all the positive features of RF have been shown. The R^2 was calculated for different methods as 0.77 for PLSR, 0.81 for ANN, 0.89 for RF, 0.92 for RK and 0.91 for RFRK. Guo, et al. 2015 compared two different approaches, namely Stepwise Linear Regression (SLR) and Random Forest Residual Kriging (RFRK), to predict and map the spatial distribution of soil organic matter for the rubber plantation. It was observed that the RFRK model did not require any assumptions concerning the correlations between the target and the predictor variables. These relationships could be either nonlinear or hierarchical. The R^2s were found to be 0.43 for SLR, 0.65 for RF and 0.86 for RFRK respectively. In Yue et al., (2018), Above-Ground Biomass (AGB) is estimated by using 54 vegetation indices and eight statistical regression techniques. Their results showed that, out of the investigated eight techniques, PLSR and GLM perform the best concerning stability and are most suitable when high-accuracy and stable estimates are requisite from relatively few samples. Furthermore, RF has been shown to be extremely resistant against noise and was ideally suited for dealing with repetitive observations involving remote-sensing data. The results showed that

GLM performed poorly in the case of multi-collinearity data to estimate AGB. ANN, BBRT and RF were the models most unaffected by the problem of multi-collinearity. Their experimental results showed that PLSR, GLM and RF can be appropriate for work requiring high-precision estimation models.

Hunt et al., (2019) estimated the within-field yield variability with Sentinel-2 and environmental data such as meteorological and soil parameters in 39 wheat fields in the UK using RF regression. They used harvester yield monitors data as ground truth and found out that Sentinel-2 data is capable of estimating within-field yield variability; however, combining satellite data with environmental data increased the accuracy. They also noted that RF outperforms the VI-based simple linear regression.

Mulla (2013) pointed out several future needs in the field of remote sensing with soft and hard computing methods. He indicated that more work is needed on chemometric or spectral decomposition/derivative methods of analysis in precision agriculture applications, while the development of new sensors is necessary to estimate nutrient deficiencies without the use of reference strips directly. He also stressed the need to develop more spectral indices to assess multiple crop characteristics (e.g. LAI, biomass etc.) and stresses (e.g. water and N; weeds and insects etc.). In order to improve decision-making in precision agriculture, historical collections of satellite remote sensing data with moderate to high spatial resolution and conventional spectral resolution should be combined with high spatial and spectral resolution real-time remote sensing data. Studies in the literature have also used soft computing to determine yields. In their paper, however, Huang et al. 2010 stated that there were no applications for the fusion of soft computing and hard computing techniques and that this could be a good research problem. The system proposed in the study of Huang et al., (2010) meets this need and essentially monitors crop yield throughout the growth process and warns the producer or decision-maker before harvesting, so that cautions can be taken to improve yield or price adjustment.

The study presented in this paper aims to meet the first component of the future needs of Mulla (2013) and the usage of soft computing mentioned by Huang et al., (2010) for early warnings.

2.6 Photosynthetic pigments in yield estimation

All vegetation contains pigments. The main pigments are chlorophylls, carotenoids and flavonoids (mainly anthocyanins) (Blackburn, 2006; Lachman et al., 2017). Photosynthetic pigments are chlorophylls, carotenoids and partially anthocyanins. There have been studies in the literature to estimate the yield by working on the structure of the photosynthetic pigments at narrow band scale and taking each pigment into account individually. The examples of these studies are presented below.

The most important pigments for the life cycle of all living beings are chlorophylls because they play the most important role in photosynthesis. Chlorophyll concentrations play an active role in primary crop production due to their control of the solar radiation absorbed by the leaves, leading to photosynthetic potential. Finally, chlorophylls are light-dependent pigments and decrease their quantity in low or no light environments, under stress and during senescence.

As for carotenoid-based indices, xanthophylls are carotenoids which are the accessory pigments of the chlorophyll a, that capture the energy that chlorophyll misses and also turns this energy into chlorophyll to make photosynthesis occur, increasing the efficiency. Therefore, xanthophyll plays a major role in the chlorophyll content of the crop (Patel et al., 2013). Chlorophyll absorbs the blue and the red light during photosynthesis and reflects the green light. The energy from the absorption of blue and red light enables photosynthesis.

The most well-known carotenoid index is Photochemical Reflectance Index (PRI) (Gamon et al., 1992). It is a commonly used index correlating with the xanthophyll process pigment’s epoxidation state. The Carotenoid Index (CARI), which is

proposed as the basis for non-destructive estimation of the leaf carotenoid content with remote sensing techniques (Huang et al. 2018). Recently a Carotenoid / Chlorophyll (car/chl) Ratio Index (CCRI) was proposed in the form of CARI / Red-Edge Chlorophyll Index ($CI_{red\text{-}edge}$). Calibration and validation results on winter wheat leaf level data showed that CCRI estimated car/chl content with 54% accuracy (Zhou X. et al., 2019).

There are about 600 different carotenoids discriminated as xanthophylls (with oxygen and most importantly lutein in wheat) and carotenes (with hydrocarbons and no oxygen). If the chlorophyll concentration is high, more energy from chlorophyll b and carotenoids will be needed, resulting in higher photosynthetic activity, thus primary production (Chappelle et al., 1992).

Anthocyanins are in the flavonoid group. Various roles of anthocyanins have been reported, such as being an antioxidant (Yamasaki et al., 1997) and in case of oxidative stress, they reduce the excitation pressure and prevent oxidative damage (Field et al., 2001). In addition, Steyn et al., (2002) found the photo-protective light screen potential of anthocyanins.

The spectral, spatial and temporal dynamics of these very important vegetation pigments can not only provide scientific knowledge, but also contribute significantly to the management of agriculture or the environment. However, the spectral discrimination of these pigments is not possible with simple linear unmixing, since they have a structure of observably non-linear, overlapping spectral signatures in a plant (Blackburn, 2006).

When examined at canopy level, the reflectance spectrum of plants is affected by leaf layers (LAI), percentage of the plant covering the ground, areas under shadow etc. Various researchers have studied the discrimination of pigments using remote sensing techniques. Using hyperspectral data, single pigments have been attempted to decompose at certain wavelengths in which clear spectral separation can be achieved. One of the most popular studies was the practice of the reflectance spectra of several narrow bands and the creation of indices mainly for the

identification of chlorophyll by dividing the values in usually two narrow bands. Testing three bands to develop indices have been mostly used at leaf scale (Gitelson et al., 2003; le Maire et al., 2004; Dash & Curran, 2004; Gitelson et al., 2005). Even four band indices have been developed (Thenkabail et al., 2002). Thenkabail et al., (2002), concluded that broadband data is not sufficient for obtaining indices, while narrowband data has a lot of autocorrelation causing redundancy.

The researchers, who had originally found the role of xanthophylls in the photosynthetic activity of plants, actually examined the role of chlorophyll a, chlorophyll b and carotenoids using their absorption spectra (Gamon et al., 1992; Penuelas, J. et al., 1994). Sims & Gamon, 2002 investigated how they could extract the chlorophyll content at the existence of all the other pigments in a leaf. Chlorophylls were observed as a whole, indicating that the total chlorophyll content overlaps the absorbance of the carotenoids and could therefore not be used to estimate the chlorophyll content.

Sims and Gamon, (2002) also investigated the estimation of the content of leaf carotenoids from reflectance, which is considerably difficult to calculate than the estimation of chlorophyll due to overlapping peaks of chlorophyll with carotenoid absorption, and due to the high concentration of chlorophyll in most leaves than carotenoid (Blackburn, 2006). Reflectance indices have therefore been more successful in estimating the ratio of carotenoid to chlorophyll than in estimating the absolute content of carotenoid (Penuelas et al., 1995; Merzlyak et al., 1999).

The importance of new methods to identify overlapping pigment absorptions was emphasized by (Ustin, et al., 2009), suggesting that it would lead to significant advances in the understanding of plant functions and other plant properties. They underlined that new methods developed to evaluate pigment content and composition from remote sensing data would provide an understanding of photosynthetic processes in a more advanced manner.

On the separation of anthocyanins from the total spectrum of plant pigments, little research has been done. (Gamon and Surfus, 1999) looked at the red area where the anthocyanins absorbed light and created a red / green index. However, the index tests showed no relation to anthocyanins. As a similar study to that of (Gamon and Surfus, 1999), (Neill and Gould, 2000) have highlighted the problem as the existence of chlorophyll obscuring the discrimination of anthocyanins in the reflectance spectrum. (Gitelson et al., 2001) overcame this problem by creating a narrowband index at 550 nm and 700 nm. This new index has been proven to be successful in different plant types.

These studies indicate that, although spectral discrimination of single pigments is possible at some level, the immediate separation of all pigments at the canopy level from satellite data is still a problem without the use of complex non-linear models.

In this thesis, we propose a method for the linear unmixing of the observed spectrum into endmembers, which presumably correspond to the pigments and can even be used directly instead of indices

.

CHAPTER 3

MATERIALS AND METHODS

In this section, the data and the study area will be introduced, to be followed by the step-by-step methodology of the study explained in this thesis.

3.1 The data and the study area

Within the scope of a “Smart Agriculture” project implemented by the Turkish Ministry of Agriculture and Forestry, the yield maps of the combine harvesters with yield mapping technology and barley and wheat products were produced (Sönmez et al., 2015). These maps were transferred to GIS and prepared for use with electro-optic and SAR satellite images. Instant georeferenced yield values were obtained with the combine harvesters equipped with accurate scaling, recording and measuring devices. The combine harvester storage capacity was around 6 tons and its width was 5 m. As the combine harvester moved at a speed of about 7 km/h, it actually displaced 2 m/s and thus yield points in 2x5 meter grids. Coordinated yield distribution maps were prepared in the yield software using the raw data of the harvested result. The efficiency system in the combine harvester was integrated with a DGPS receiver via a display. During the operation, the data were recorded at 1 second intervals with location information.

3.1.1 Yield data

A harvester system was used as part of a project covering fields from various regions of Turkey, for the year 2015. The GPS-equipped harvester is integrated

with high-tech sensors that automatically weighs the crop. These records were disseminated by the Ministry of Agriculture and Forestry, enabling researchers to reach the exact locations and yields of the fields in which the harvester operated. 142 wheat fields were selected in 31 cities in six regions of Turkey, as study areas. These fields were all rain-fed fields in different regions of Turkey having distinct climatic conditions.

A snapshot of the image of a harvested field within the Harvester Project can be seen in Figure 3.1, the stripes seen in the figure shows the harvesting direction of the field. The speed of the harvester during the process was also displayed in the popup.

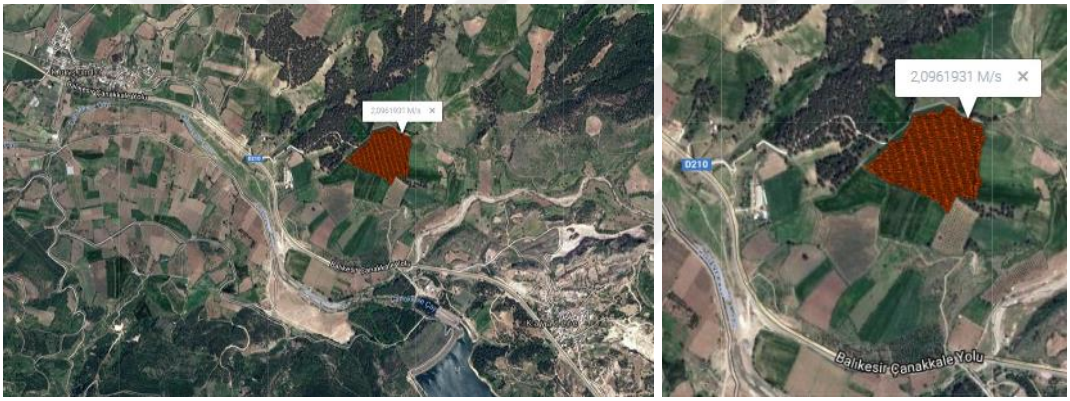

Figure 3.1. GPS data of the harvester forming yield grids overlaid on Google earth base image. The image on the right is the enlarged image of the left one to visualize the speed information of the harvester.

In order to link the fields in the satellite images to harvester yield records, the exact boundaries and coordinates of the agricultural field polygon was extracted from Google Maps.Figure 3.2 shows the locations of 142 selected agricultural fields in 31 Turkish provinces and the station locations that JRC gathered the agrometeorological data from.

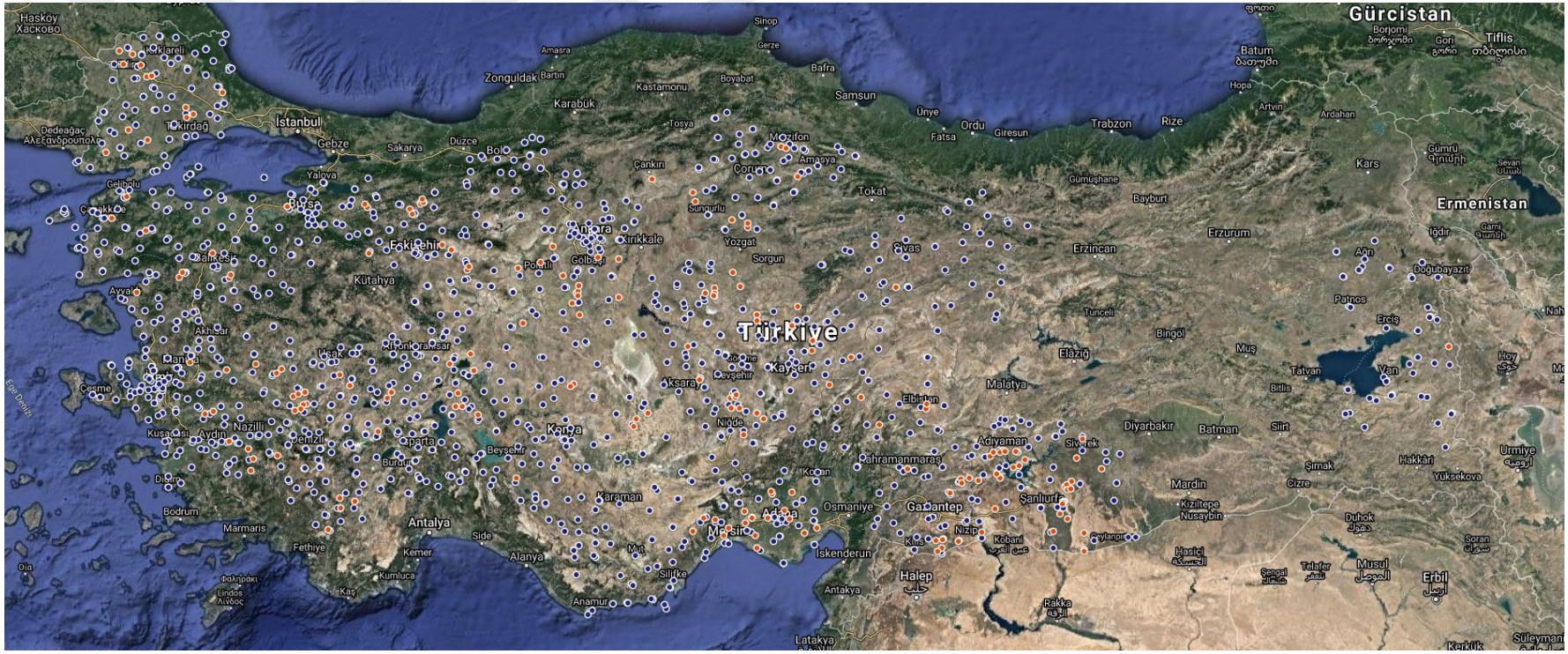

Figure 3.2. Locations of the 142 fields spread around 31 provinces of Turkey. The blue circles represent the meteorology stations and the red circles stand for the fields.

3.1.2 Meteorological data

In most cases, the estimation of yield requires agrometeorological data, which are meteorological data or data that are derived from meteorological data and used in agricultural studies in order to obtain qualitative and quantitative improvement in agricultural production (WMO, 2010). If the area of interest, i.e. the crop field, is in the vicinity of a meteorological station, the data collected from that station can be used directly. However, if the Area of Interest (AOI) happens to be in between several meteorological stations, it is ideal to interpolate the agrometeorological data. The objective of this study is to serve a general approach in the estimation of yield so that people involved in this type of work can find somewhat more general solutions. Therefore, the JRC data (Toreti, 2014) has been used instead of the meteorological station data as in the studies of (Fernandes et al., 2011) and (Salvador et al., 2020).

The JRC collects air temperature, precipitation, radiation, air humidity, and wind speed data from 117 weather stations in Turkey. The data are checked for inconsistencies, errors and duplications and only after these evaluations; the values are converted into daily values that fit into a uniform weather database for the station. The measured data are derived from some variables such as solar radiation or evapotranspiration are also added to the database. These data are, however, obtained from stations that are at point locations and thus have an irregular distribution and density. A conversion is required to disseminate these data to locations between stations. JRC uses interpolation, aggregation and analysis and controls the regularity by using side by side grids of size 25 km by 25 km, which covers the entire area of interest (Weather Monitoring - Agri4castWiki).

The total number of agrometeorological parameters both obtained from JRC and the calculated ones, vegetation indices used in the literature (*MSAVI*, *MTVI*, etc.) and the abundances together with their interactions is 24. Using all the parameters can be unnecessary, because they can be irrelevant or highly correlated to others.

3.1.3 Satellite data

Due to their global coverage and temporal resolution, Landsat 8 satellite images were used in this study. Landsat 8 satellite has a 16-day temporal resolution and contains 11 bands, however only the first five bands were used in this study. The properties of these bands can be found in Table 3.1. The Landsat 8 images of the selected areas from April to June 2015 were downloaded from https://earthexplorer.usgs.gov/ website and processed. At least two different images were found for each field at 0% cloud coverage prerequisite. The image having the highest NDVI value for each field was selected as the input image to be used in the forthcoming processes. At this time of the crop development, it is assumed that the vegetation cover is virtually 100% and there is negligible soil contribution to the spectrum.

Table 3.1 Landsat 8 satellite bands used in this study and their properties

Bands	Wavelength (micrometers)	Spatial Resolution (meters)
Band 1 - Coastal aerosol	0.43-0.45	30
Band 2 - Blue	0.45-0.51	30
Band 3 - Green	0.53-0.59	30
Band 4 - Red	0.64-0.67	30
Band 5 - Near Infrared (NIR)	0.85-0.88	30

3.2 Methodology

The first step is the preparation of the satellite images and extraction of the selected fields from the relevant satellite images. Then, the endmembers are determined and analysed at the time of maximum NDVI before finding their abundance in the fields. The abundances, NDVI and selected agrometeorological parameters are trained in GLM and RF algorithms to predict the yields. The details are presented in the sub-sub-sections below. A flowchart of the algorithm is given in Figure 3.3.

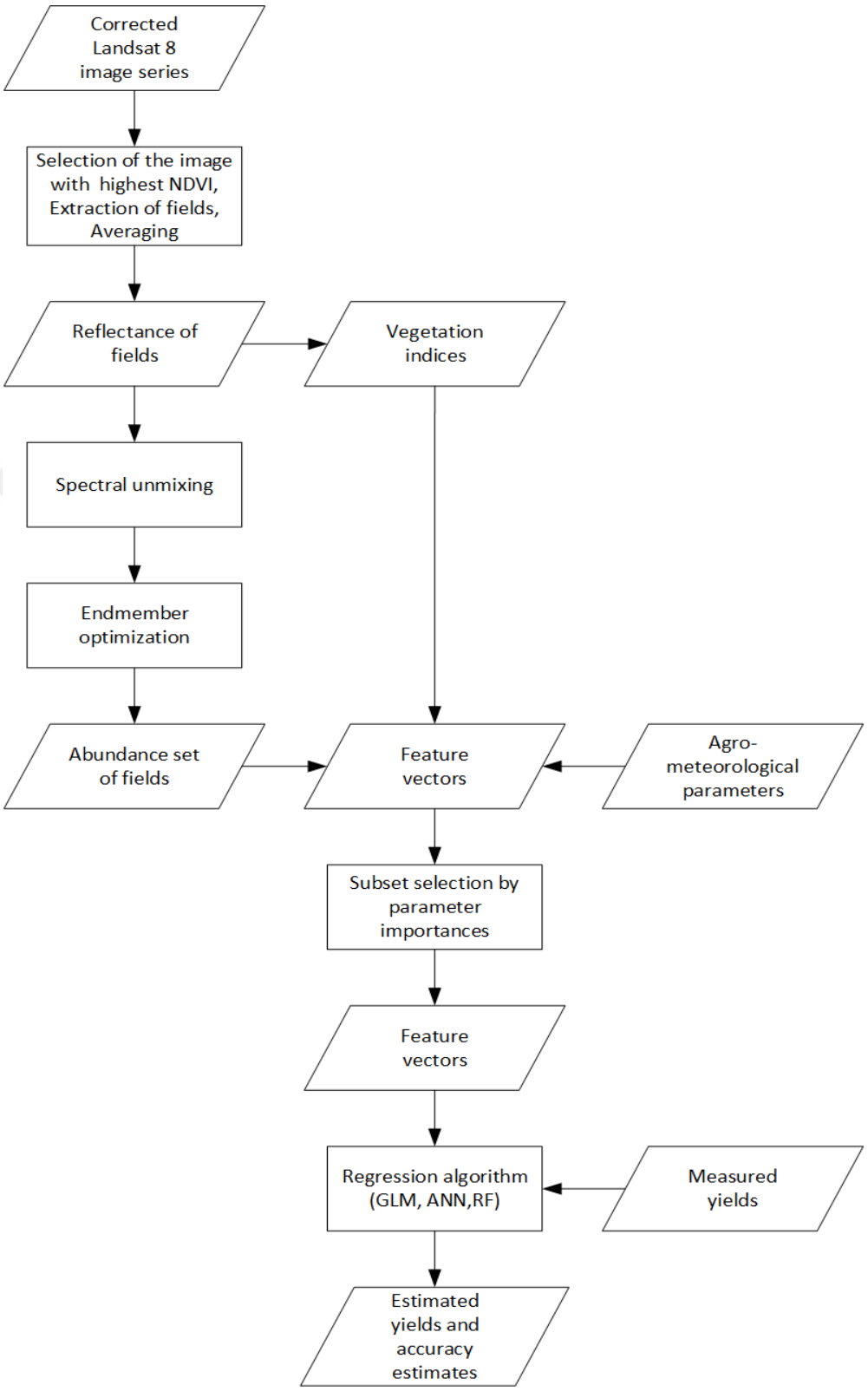

Figure 3.3 The flowchart of the proposed algorithm.

3.2.1 Preparation of the data

All the Landsat 8 images of the area of interests were corrected for radiometric and atmospheric effects using FLAASH in ENVI 5.3. 141 fields extracted from Landsat 8 images after they were geometrically and radiometrically corrected. The image with the highest average NDVI value was selected for each field and used as the dataset for the extraction of the endmembers and the abundance calculation.

Two agrometeorological parameters were used as indices in this study, to observe if they have any effect on the estimation of the wheat yields, namely *NoPRECIPITATIONdays* and *Cons_noPrec*. *NoPRECIPITATIONdays* is the total number of days there was no rain at the area of the field starting from the sowing time until the day of harvest, and *Cons_noPrec* is the consecutive number of days when there was no precipitation from sowing time until the harvest day and helps to estimate the accurate yield in case there is drought. It is also a very useful indicator to find the time when there are many consecutive days of no precipitation and if and/or how it affects the yield at that certain time period.

3.2.2 Spectral Unmixing

Since the pixel sizes of multispectral sensors are large to contain various components, the extraction of desired components has become important for research purposes. Spectral unmixing is applied (Keshava & Mustard, 2002; Somers et al., 2011) in linear spectral unmixing, or Linear Mixing Model (LMM), in which it is assumed that a mixed pixel consists of a set of constituent spectral signatures (endmembers) weighted by a fractional subpixel cover (abundance) as shown in Eq. 3.

$$\boldsymbol{Y} = \boldsymbol{E}\,\boldsymbol{A} + \boldsymbol{\varepsilon} \tag{3}$$

where $\mathbf{Y_{d,n}}$ is composed of observed *n* pixels with *d* bands, $\mathbf{E_{d,m}}$ is the matrix where each column is an endmember, $\boldsymbol{A_{m,n}}$ is the abundance matrix where each column

represents the fractional cover occupied by the endmembers for the corresponding pixel and **ε** is the error, the portion of the spectrum that cannot be modelled using the endmembers (including sensor noise, endmember variability and other model inadequacies).

Normally, the LMM can actually have two constraints:

(1) Non-negativity constraint: All abundances have to be non-negative. $\mathbf{A} \geq 0$ and

(2) Full-additivity constraint: $\mathbf{1}_{\boldsymbol{m}}^{\boldsymbol{T}}\boldsymbol{A} = \mathbf{1}_{\boldsymbol{n}}^{\boldsymbol{T}}$.

Crops consist ideally only of photosynthetic pigments that are intimately mixed in the vegetation. The endmember extraction of these intimate mixtures requires the use of complex non-linear techniques rather than linear ones. The endmembers are the inputs for the next processing step, which generates the abundances that are the inputs of the yield prediction method.

Spectral unmixing is very important and it has recently gained attention in the hyperspectral studies. However, the unmixing algorithms have limited usage in the multispectral studies so far and the researchers usually perform the unmixing in a significantly large image area where a large variety of land cover types exist. However, if the scene is a crop field consisting of full coverage leaves, only plant pigments are expected to be the endmembers. Given that these pigments form intimate mixtures, the standard LMM performance would therefore be limited. As a first approximation of the endmembers, a linear method, Robust Collaborative Nonnegative Matrix Factorization (R-CoNMF) is used (Li et al., 2016). R-CoNMF computes the abundances from the endmembers, which are then used in estimating the yields. When finding the endmembers, R-CoNMF actually performs linear unmixing to the pure crop pixel that presumably consists of intimately mixed photosynthetic pigments. One very important property of R-CoNMF that we took advantage of in this study is that the endmembers do not necessarily correspond to pure pixels, because there are not pixels composed of pure photosynthetic

pigments. Although the initial endmembers are determined by a pure pixel algorithm, they are iteratively updated so that the final endmembers do not necessarily appear as pure pixels in the image.

Still, the endmembers determined by the R-CoNMF algorithm are not optimal. For that purpose, we propose an optimization scheme that increases the predictive power of the endmembers. The optimized endmembers are defined as follows:

$$E_{opt} = \text{argmin}_E(\text{Rsquared\_}glm(SUnSAL(\boldsymbol{E}, \boldsymbol{Y}), y)) \quad (4)$$

where $\boldsymbol{E}$ represents the endmembers, $\boldsymbol{Y}$, the field average pixels as earlier, while *SUnSAL*, an abundance estimation method proposed in (Li et al., 2016), returns the abundance matrix $\boldsymbol{A}$ (both constraints are used). y are yields in our study and Rsquared_*glm* is the coefficient of determination given by MATLAB™ function 'fitglm' with interactions so that non-linearity is partially modelled. The function is minimized using unconstrained multivariable minimization as implemented in 'fminsearch' function of MATLAB™ and the endmembers found by R-CoNMF are used as the initial values. The solution of R-CoNMF is used as the initial value of the optimization algorithm. That is, the endmembers are modified so as to maximize the R^2 value between the yields and the abundances.

3.2.3 Linear and non-linear regression

The abundances were used as regression algorithm inputs to achieve the ultimate intention to find early yield estimates. The selected methods for achieving the final goal were GLM, ANN and RF. All GLM, ANN and RF algorithms were executed in MATLAB™.

It is important to decide which parameters are really useful in estimating yields. Depending on the parameter set, the importance of selected parameters also changes. In this study, three approaches were used to select parameters and estimate yield using different datasets such as;

- Only the abundances,
- Abundances with their interactions with each other (to partially address the non-linear mixing),
- Only the agrometeorological parameters,
- The agrometeorological parameters with NDVI and
- The abundances together with all the other parameters.

The selection and use of appropriate parameters is of vital importance in all approaches in order to achieve the desired results. The parameter importance values that were estimated by using out of bag samples by the random forest algorithm were used. RF, ANN and GLM were used to estimate yield after selecting the appropriate parameters in each approach.

CHAPTER 4

RESULTS AND DISCUSSION

The endmembers and the abundances found by following the steps of the methodology are given in this section. The outcomes of the machine learning algorithms by making use of all available data as well as a selected portion of all the data, that are considered to be the most important, are compared in the discussion section.

4.1 The endmembers

The endmembers can be considered as the spectral signatures of the dominantly existing textures in the field of interest. In our case, since we are looking at wheat fields that are almost totally covered and green (full closure, maximum NDVI), one would expect to find the plant pigments as endmembers. The endmembers were calculated using R-CoNMF (Li et al., 2016) in MATLAB™ for the 142 fields in 31 cities where Landsat 8 images existed just before the harvesters recorded yield data. The first four multispectral bands of Landsat 8 images, where the photosynthetic pigments were mostly absorbent, were used to obtain four endmembers. The parameters used for the implementation of R-CoNMF can be seen in Table 4.1. Figure 4.1 shows the automatically calculated mean square error, projection error and noise power used for the implementation of R-CoNMF as a function of number of endmembers when the R-CoNMF code is run.

Table 4.1 Parameters and their values used to perform R-CoNMF

Parameter	Value	Explanation
Positivity	yes	Enforces the positivity constraints
Alpha	0.1*(1e-8)*sqrt(nd)	Regularization parameter, nd is the number of pixels of all the fields, which id equal to 7576 in this study
Beta	50*10^(-3)*(nd*m_em)	Minimum volume regularization parameter, n_em is the number of endmembers which is equal to 4 in this study
Addone	yes	Enforces the positivity constraints
AO_Iters	100	Number of iterations
Delta	1e-4	(STOP) relative reconstruction error
Csunsal_Iters	100	SUnSAL number of iterations
Mu_A	0.1e-4*(nd*n_em)	Proximity weight for A, optimization variables linked with the mixing matrix
Mu_X	1e-2	Proximity weight for Y, optimization variables linked with the abundance matrix
Spherize	M	{'no','cov', 'M'}, M is the estimated mixing matrix containing the 4 endmembers
Min_Volume	center	{'boundary', 'center', 'totalVar'}

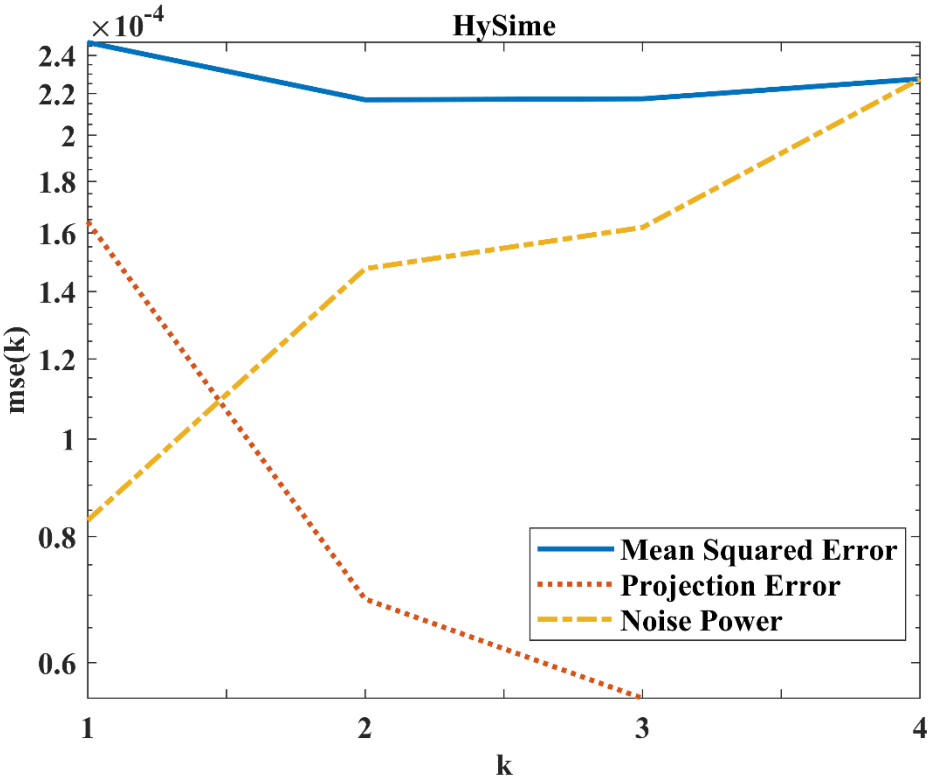

Figure 4.1 Mean square error, projection error and noise power used for the implementation of R-CoNMF as a function of number of endmembers (k)

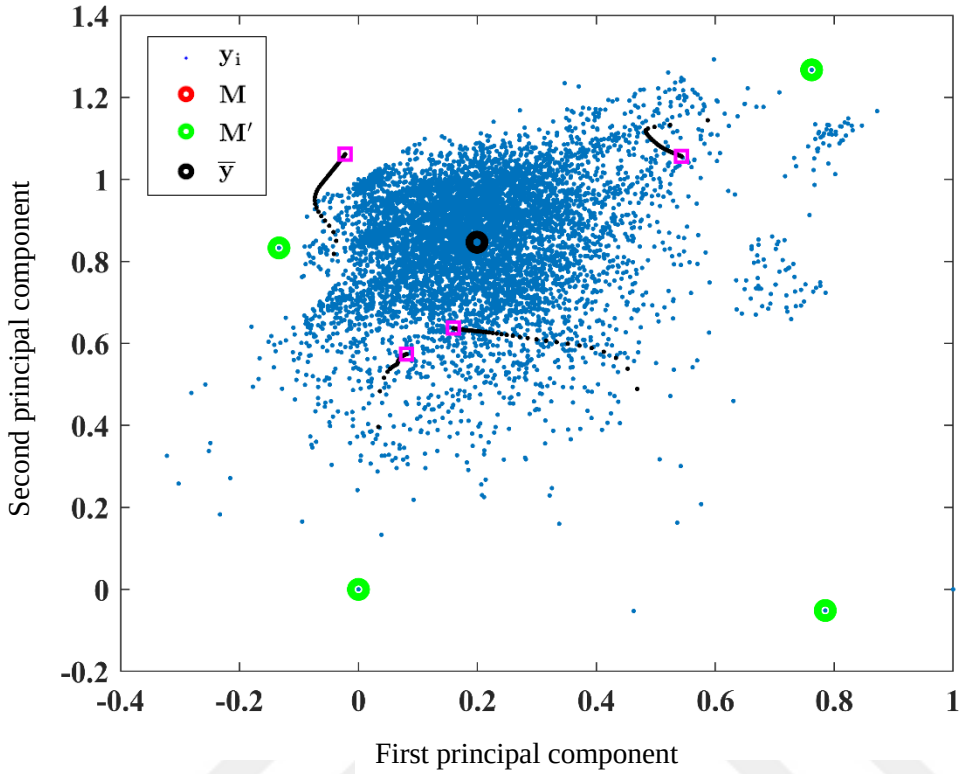

Figure 4.2 Projection of the spectral vectors on the endmembers shown on the first two principal components. Projection of the spectral vectors $\bar{\boldsymbol{y}}_i$, for i = 1, . . . ,n (blue), (where $\bar{\boldsymbol{y}}$ is the sample mean vector); of the endmember signatures $\boldsymbol{m}_i$, for i = 1, . . . ,p (magenta); and of the columns of $\boldsymbol{A}$, which are not endmembers (green). The spectral mean value is shown in black.

Figure 4.2 shows a vector scatter plot $\bar{\boldsymbol{y}}_i$, for i = 1,. .n, projected onto the affine set identified by the $\boldsymbol{M}$, which is a so-called mixing matrix containing $\boldsymbol{p}$ endmembers columns centered at $\bar{\boldsymbol{y}}$, which are plotted in black. It also shows the projection of matrix $\boldsymbol{A}$, where $\boldsymbol{M'}$ contains 5 spectral vectors on the facets of the simplex defined by $\boldsymbol{M}$. The $\boldsymbol{M}$ and $\boldsymbol{M'}$ projections are in red and green, respectively. The black dots ending at the magenta endmembers represent the solution found by R-CoNMF with β from Inf to 0. If β is set well, the final endmembers will be close to the real ones.

The mean reflectance spectra of the photosynthetic pigments (Figure 5.3(a)) are obtained from the absorbance values in Lichtenthaler (1987) by using the conversion factors in Gitelson and Solovchenko (2018).

Lichtenthaler (1987) provides the absorption spectra of the photosynthetic pigments, among others. However, this study and probably some other remote sensing studies require the reflectance spectra of the pigments to analyse the spectra pigments and other materials in satellite images. Still, the transformation of absorbance to reflectance is not a very straight-forward process. Although there are complicated mathematical models to estimate the reflectance from absorbance (Dawson et al. 1998; Jacquemoud and Baret, 1990), we prefer to use empirical data in this study. So the absorbance and reflectance values of Virginia creeper leaf reported in Gitelson and Solovchenko (2018) were used.

The values were used in a MATLAB™ function which basically uses these values at each wavelength where the photosynthetic pigments are effective (400 – 700 nm) in the transformation formula:

$$R^{-1}_{pigment}(\lambda) = \frac{R^{-1}_{leaf}(\lambda)}{c\, A_{leaf}(\lambda)} A_{pigment}(\lambda)$$

(5)

where λ is the wavelength, $R_{pigment}$ is the reflectance of the pigment, $A_{pigment}$ is the absorbance of the pigment, , R_{leaf} is the reflectance of the leaf, A_{leaf} is the absorbance of the leaf at λ (Table 4.2), and c is a constant that brings the range of the two absorbance values in the same range because the absorbance units are different in the two sources.

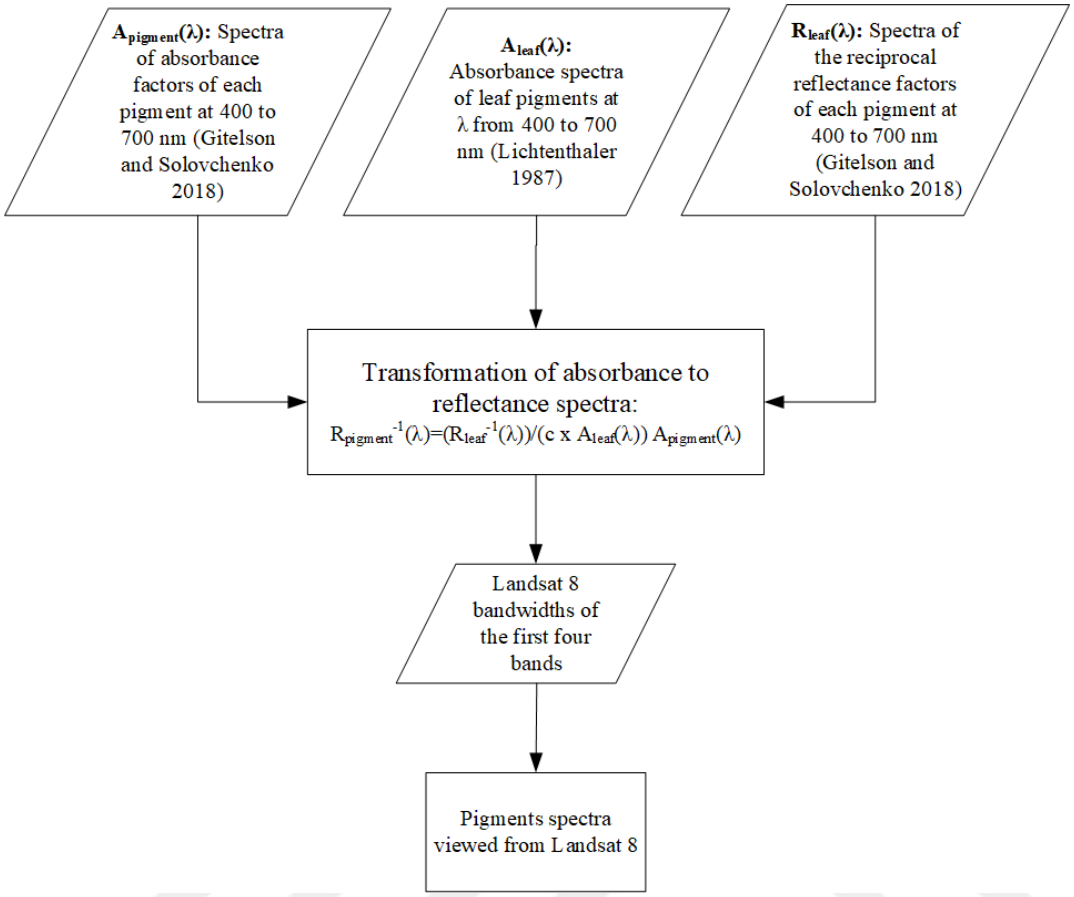

Figure 4.3 The transformation of the absorbance spectra of photosynthetic pigments to reflectance spectra.

The newly found reflectance spectra of each pigment was then processed with the bandwidths of the first four bands of Landsat 8 satellite. The transmissivity of each band is assumed to be unity in the passbands and the reflectance spectra is simply summed in each band to obtain the reflectance values as seen by Landsat 8. The flowchart of this process can be seen in Figure 4.3.

Table 4.2 Estimated absorbance and reflectance values of Virginia creeper leaf reported as graphs in Gitelson and Solovchenko 2018

Wavelength (nm)	**Absorbance values** $\mathbf{A}_{leaf}$ (λ)	**Reflectance values** $\mathbf{R}_{leaf}$ (λ)
400	2.25	11.5
425	2.1	12
450	1.9	12.5
500	1.5	13
550	0.8	6
600	1	9
650	1.3	14
680	1.75	15.5
700	0.8	6
750	0.35	2

Figure 4.4 shows the final reflectance spectra of the pigments obtained using the methodology described above. Figure 4.5(a) shows how the spectra of the pigments would look if observed directly from the first four bands of Landsat 8, for comparison to the endmembers. The endmembers obtained from the R-CoNMF and optimization are shown in Figure 5.4(b).

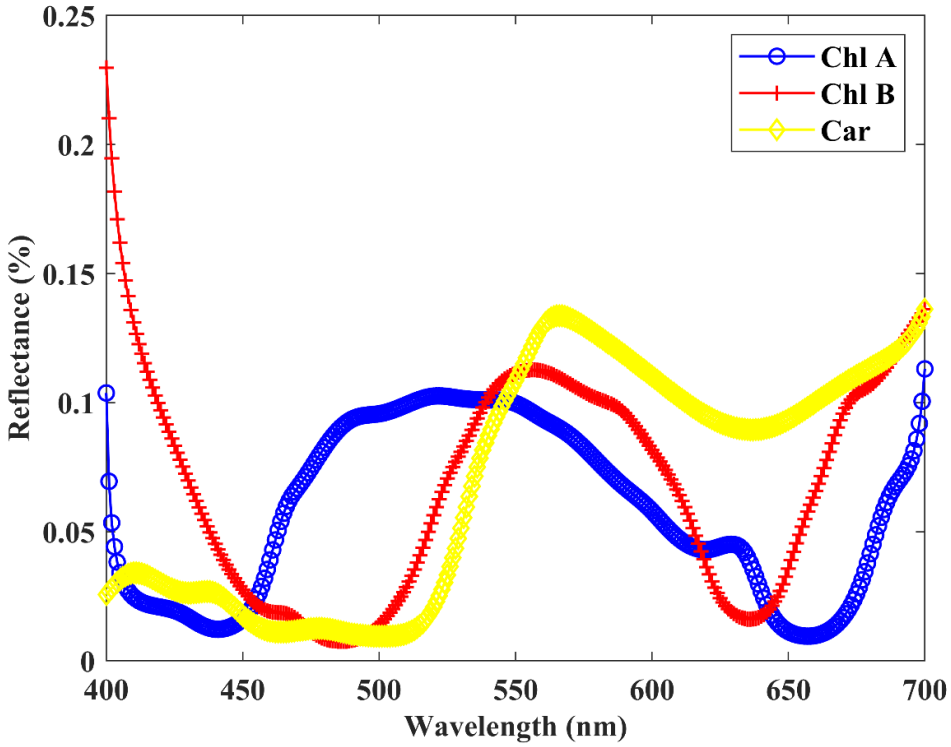

Figure 4.4 The reflection spectra of the major pigments obtained from the absorption spectra in Figure 2.2.

The absorbance and reflectance spectra of photosynthetic pigments are mostly drawn in laboratory conditions with spectrometers in a narrow-band format. The bands where these pigments are most active is between 400nm to 700nm. As Landsat 8 has only four bands in that range, we can only observe a very generalized view of the actual signatures.

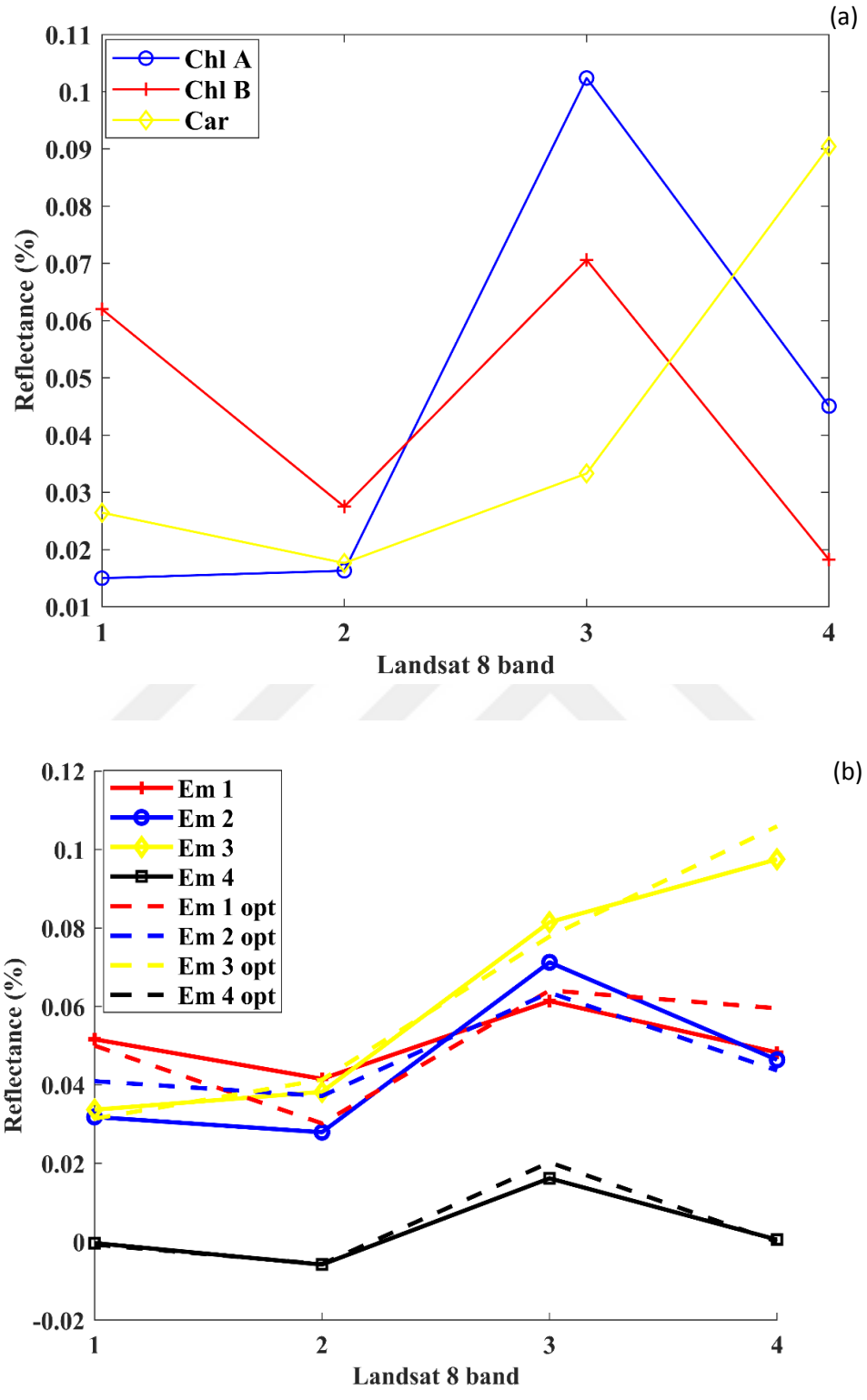

Figure 4.5 (a) The calculated pigment reflectance of the first four bands of Landsat 8; (b) The endmembers found from R-CoNMF algorithm and optimization.

R-CoNMF is used in this study as it successfully gives endmembers similar to spectral signatures of the photosynthetic pigments (compare Figure 4.5(b) to Figure 4.5(a)). The optimization, helped the endmembers to become even more similar to the real pigment endmembers shown in Figure 4.5(a). Therefore, the endmembers in Figure 4.5(b) can be interpreted as follows: Endmember 1 is related to chlorophyll b, Endmember 2 is related to chlorophyll a, Endmember 3 is related to carotenoids. The fourth endmember shows any non-modelled elements and presents its existence as a results of the non-linearities that are inevitably present within the structure of a plant.

4.2 The abundances

The four abundances that were also calculated using R-CoNMF are plotted against the known yields of the fields. The relationships of each abundance with the yields can be seen in Figure 4.6.

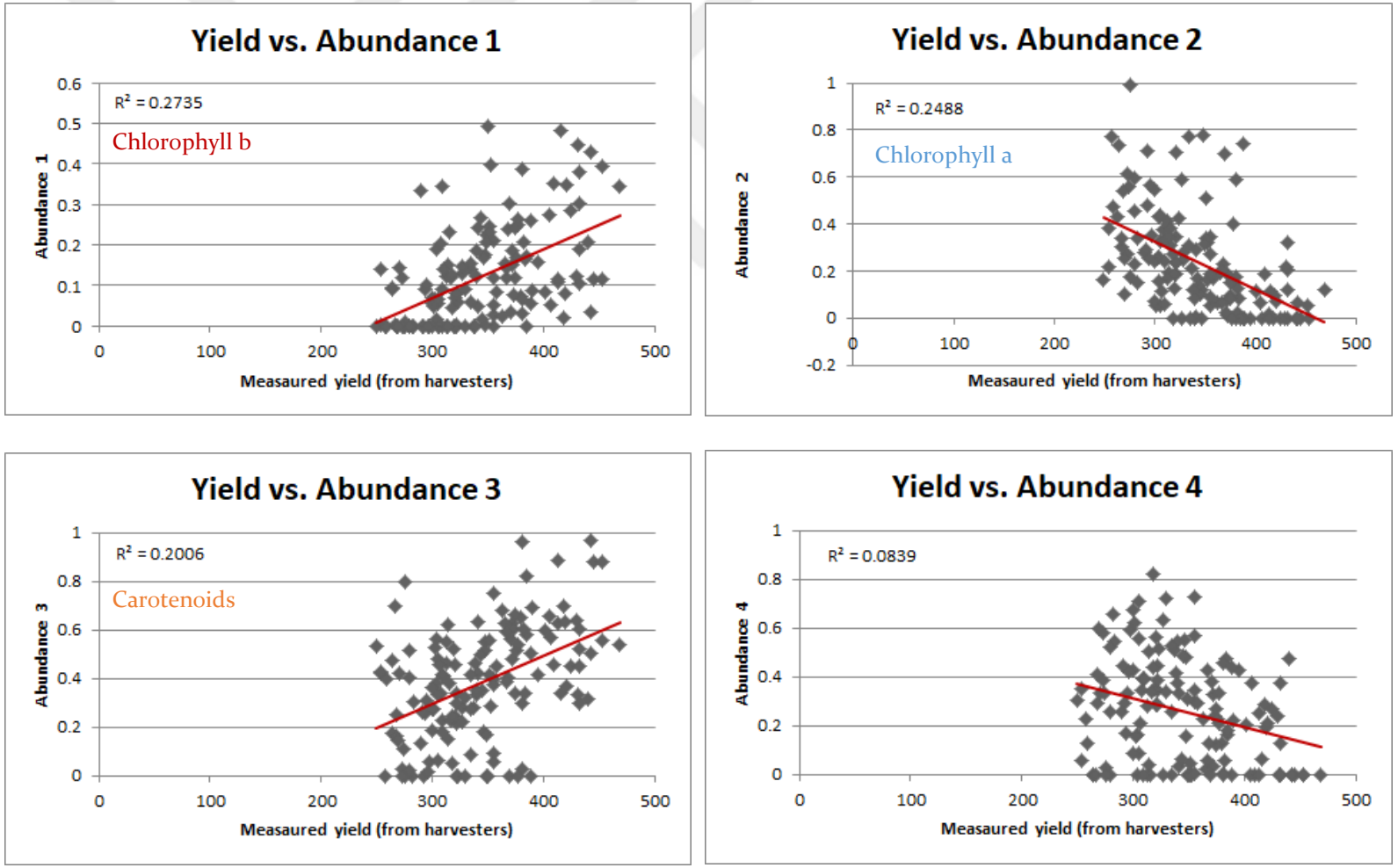

Figure 4.6 Real yields vs. abundances with R^2s. Each dot represents one agricultural field.

The abundances of the endmembers related to chlorophyll b and carotenoids (Abundance 1 and Abundance 3, respectively) have strong correlation with the yield, whereas the abundance of the endmember related to chlorophyll b (Abundance 1) has negative correlation. The fourth abundance carries the information coming from the fourth endmember which represents all the other non-modelled elements. It is shown that chlorophylls and carotenoids are positively correlated with dry mass in wheat (Sabo et al. 2002), however chlorophyll b has the least correlation. Since an increase in one of the quantities will result in a decrease in others, chlorophyll b can have negative correlation with the yield due to the sum-to-one constraint.

4.3 Parameter selection and interactions

The list of all the parameters, namely the abundances, agrometeorological parameters and the vegetation indices that have been either collected or calculated from our dataset can all be found in Table 4.3. All possible interactions of the abundances are also included in the table. The importance of the parameters in Table 4.3 are examined by using the predictor importance property of random forest.

In order to investigate the effect of the soil of the fields, we obtained a categorical soil map of Turkey and added the categorical parameters as binary into the dataset. Their contributions individually and the contribution of their interactions with all the other parameters were investigated. The contribution of any of these parameters were very insignificant, therefore all soil parameters were removed from the dataset.

Table 4.3 Names and abbreviations of all the parameters and interactions

Abbreviation	**Name of the parameter**
A1	Abundance 1
A2	Abundance 2
A3	Abundance 3
A4	Abundance 4
A1A2	Interaction of Abundances 1 and 2
A1A3	Interaction of Abundances 1 and 3
A1A4	Interaction of Abundances 1 and 4
A2A3	Interaction of Abundances 2 and 3
A2A4	Interaction of Abundances 2 and 4
A3A4	Interaction of Abundances 3 and 4
T_MAX	Maximum temperature (^{o}C)
T_MIN	Minimum temperature (^{o}C)
VPD	Average vapour pressure deficit (hPa)
E0	Potential evapotranspiration of open water (mm/day)
SOILFREEZE	Total number of days the soil temperature was below or equal to 0^{o}C (day)
NoPRECIPITATIONdays	Number of days there was no precipitation until harvest (day)
RADIATION	Average radiation (KJ/m^{2}/day)
WINDSPEED	Average speed of the wind at 10m (m/s)
Cons_noPrec	Consecutive no precipitation days until harvest
Elevation	Average elevation from sea level (m)
PTU	Photo Thermal Unit
NDVI	Normalized Difference Vegetation Index
MTVI	Modified Triangular Vegetation Index
MSAVI	Modified Soil-Adjusted Vegetation Index
EVI	Enhanced Vegetation Index

The relative importance of all the 25 parameters used in estimating the wheat yields were calculated using "Out of Bag Predictor Importance Estimates" of the RF algorithm of MATLAB™, as can be seen in Figure 4.7. When combined all together, the most important parameters were calculated to be the *VPD, T_MIN*, *T_MAX*, *RADIATION* and *A2A4*. The predictor association test showed the high correlation between *VPD* and *Elevation*, therefore, given the high importance of *VPD* compared to all the other parameters, Elevation was removed from the predictor list. *EVI* and *MTVI* were also highly correlated and *MTVI* was removed due to *EVI* being more significant. *PTU* and *A1A4*, *A2A4*, *A3A4*, *NDVI* and *MSAVI* showed low influence in the prediction of yield, thus were removed from the list. *Windspeed*, compared to all parameters, showed insignificant importance, which led to its removal from the parameter list. All abundances and their most of their interactions seemed to show significant contribution to the yield, according to the "Predictor Importance Estimates" algorithm, thus they remained with the most important agrometeorological parameters and *EVI*, which proved to be more important than *NDVI* in this study. The insignificancy of the NDVI can be explained with the fact that the study area is very diverse, with different climate and soil conditions. However, since this study aims to find the yield across the country, the parameters for this wider area is used in this thesis. Still, NDVI plays a very important role here, as the selection of the satellite data that the abundances are found from, is selected by the assumption of the fields having the highest NDVI value at the full closure of the green wheat field.

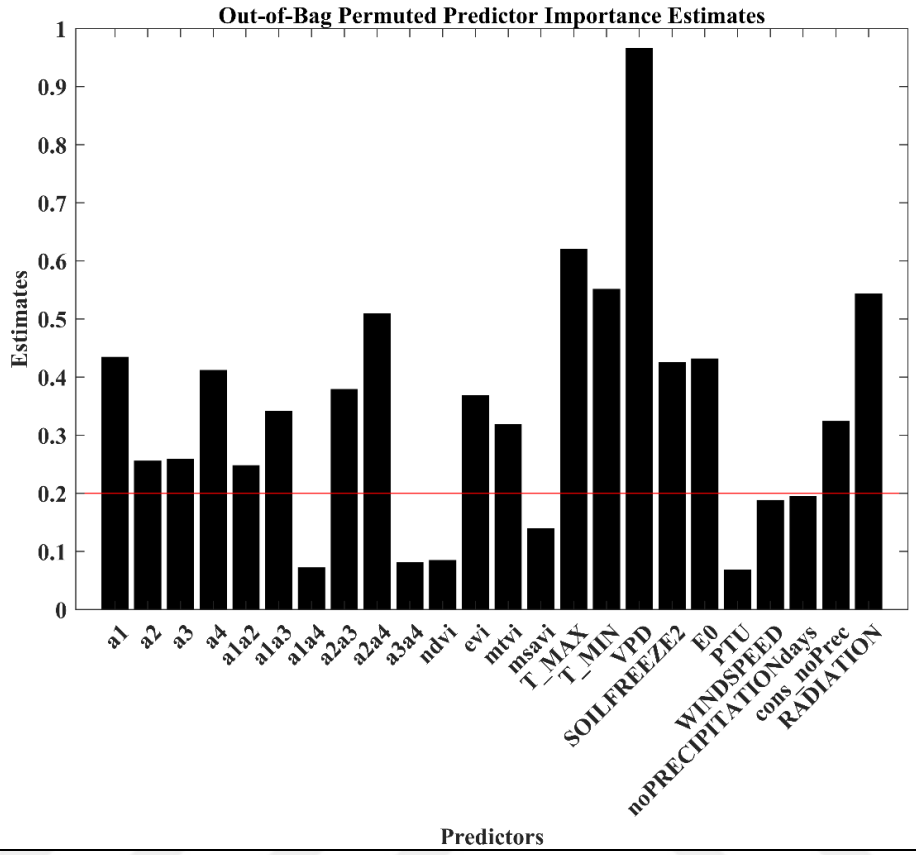

Figure 4.7 Out-of-bag importance of all parameters.

The comparison of predictor importance estimates by permuting out-of-bag observations and those estimates obtained by summing gains in the mean squared error due to splits on each predictor can be seen in Figure 4.8.

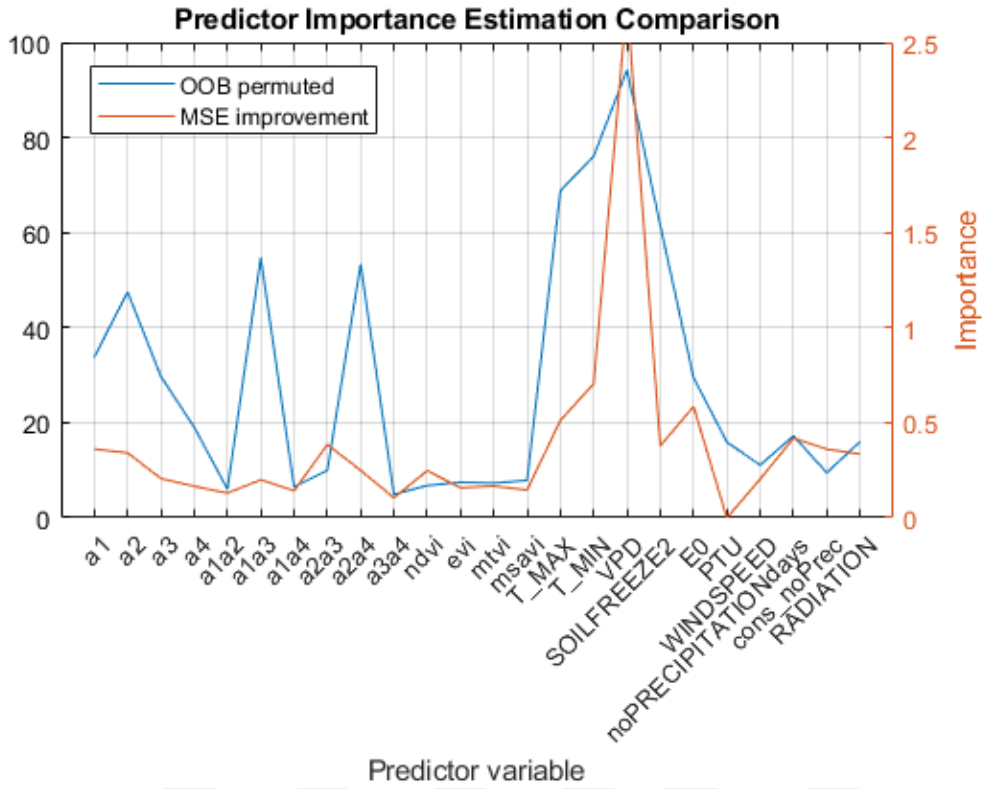

Figure 4.8 Predictor importance estimation comparison.

The predictor association measures estimated by surrogate splits were also observed (Figure 4.9). Predictor association is a 24x24 matrix of predictor association measures of all the parameters that can be potentially used to estimate wheat yields. The strength of the relationship between pairs of predictors can be inferred using the elements of the predictor association. Larger values indicate more highly correlated pairs of predictors. The largest association in our dataset was between *MTVI* and *EVI* with 78.2% relationship, however this value was not high enough to indicate a strong relationship between the two predictors, that one of them should be removed from the dataset.

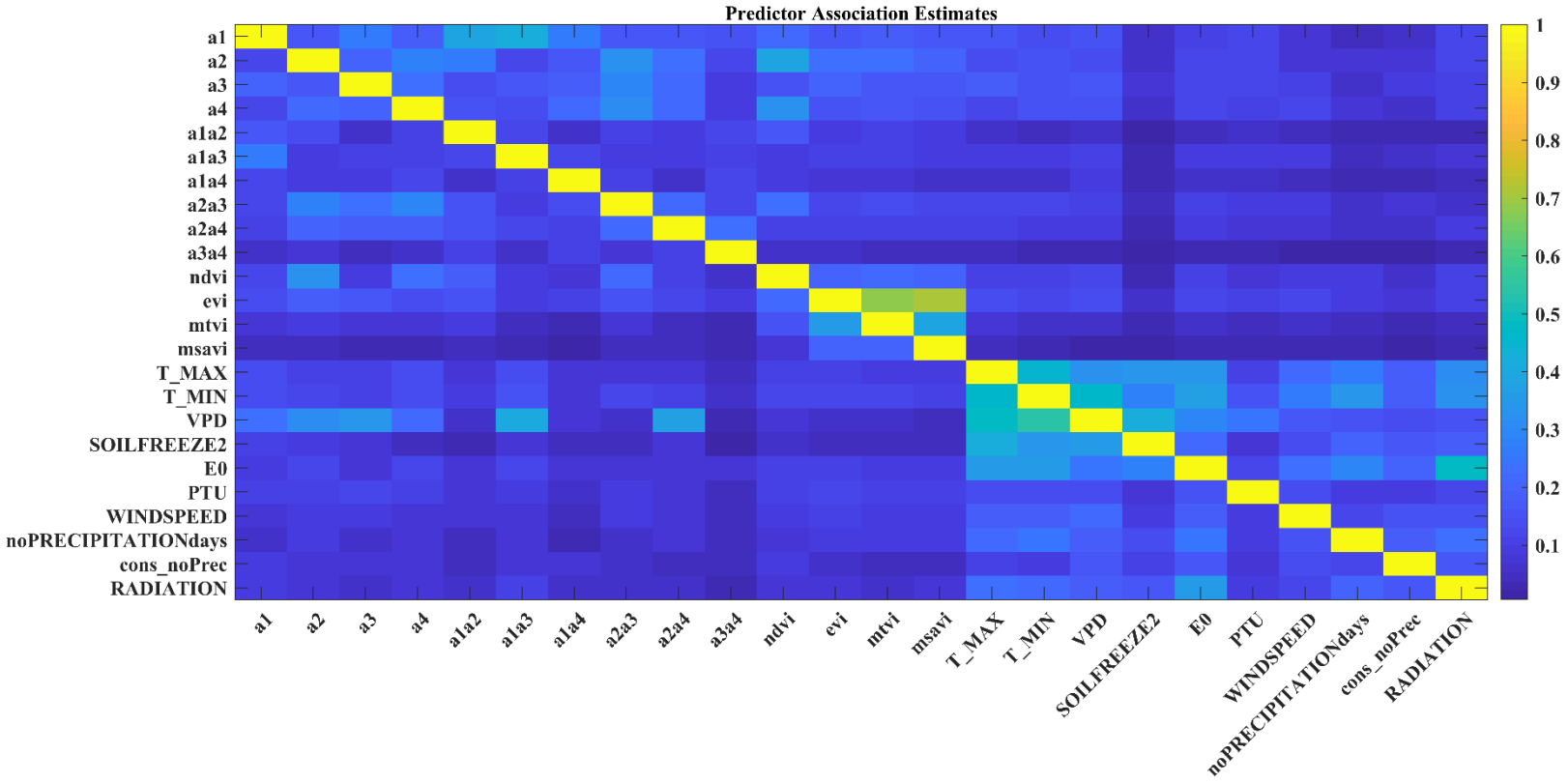

Figure 4.9 Predictor association estimates of all parameters.

After examining the results of the important parameter selection algorithms, the remaining abundances and the agrometeorological parameters under a certain threshold, that made the highest contribution to the estimation of yields were selected. The final parameter list can be seen in Table 4.4.

Table 4.4 Names and abbreviations of all the selected parameters and interactions

Abbreviation	**Name of the parameter**
A1	Abundance 1
A2	Abundance 2
A3	Abundance 3
A4	Abundance 4
A1A3	Interaction of Abundances 1 and 3
A2A3	Interaction of Abundances 2 and 3
A2A4	Interaction of Abundances 2 and 4
T_MAX	Maximum temperature (^{o}C)
T_MIN	Minimum temperature (^{o}C)
VPD	Average vapour pressure deficit (hPa)
E0	Potential evapotranspiration of open water (mm/day)
SOILFREEZE	Total number of days the soil temperature was below or equal to 0^{o}C (day)
NoPRECIPITATIONdays	Number of days there was no precipitation until harvest (day)
RADIATION	Average radiation (KJ/m^{2}/day)
Cons_noPrec	Consecutive no precipitation days until harvest
EVI	Enhanced Vegetation Index

In order to justify the selected parameters, the same Unbiased and Out-of-bag parameter selection algorithms together with predictor association algorithm were applied to the selected parameters and the importance of the selected parameters can be found in Figure 4.10.

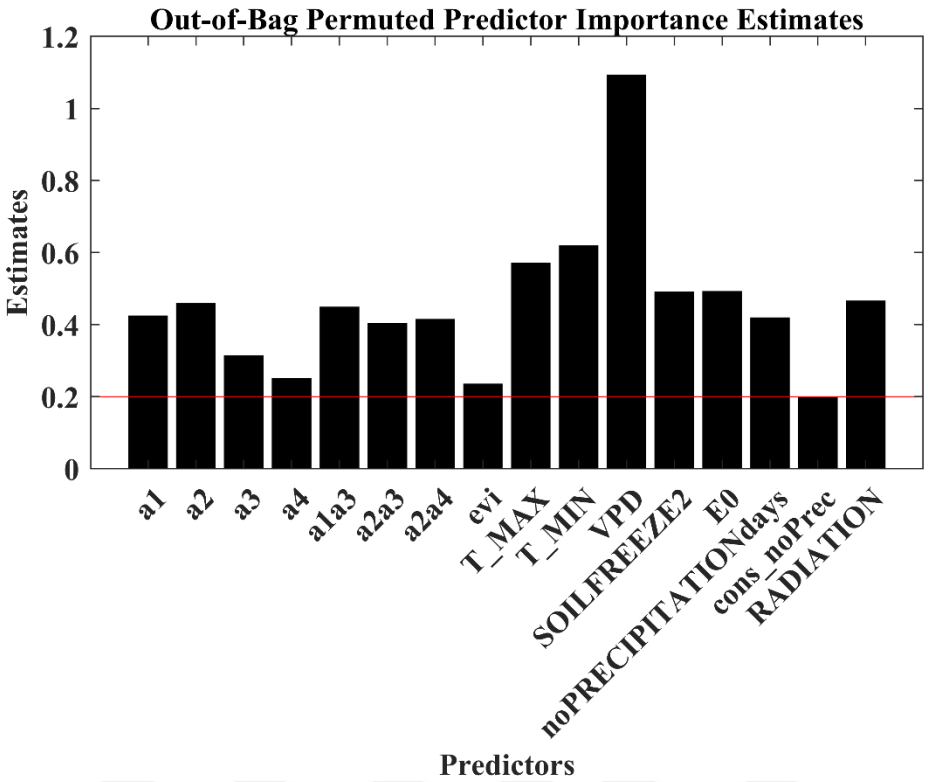

Figure 4.10 Out-of-bag importance of selected parameters.

The comparison of predictor importance estimates by permuting out-of-bag observations and those estimates obtained by summing gains in the mean squared error due to splits on each of the selected predictors can be seen in Figure 4.11.

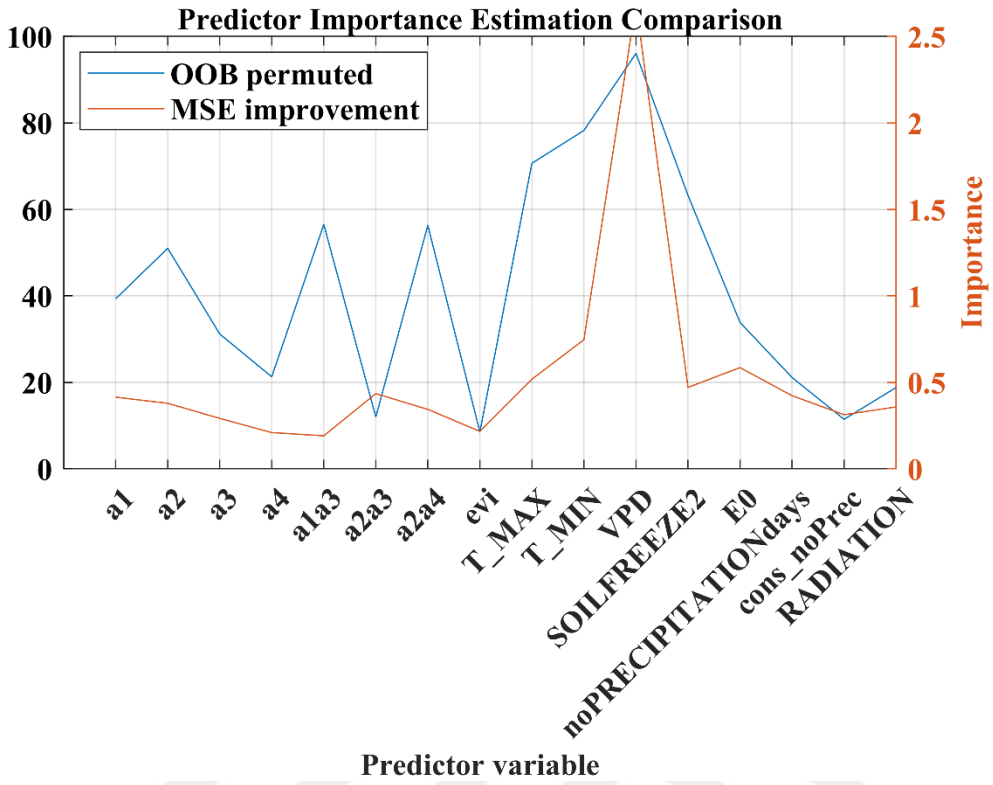

Figure 4.11 Predictor importance estimation comparison.

The predictor association measures estimated by surrogate splits were also observed for the selected parameters (Figure 4.12). Predictor association is now a 16x16 matrix of selected predictor association measures of all the parameters that can be used to estimate wheat yields. The largest association in the selected parameters dataset was between *T_MAX* and *T_MIN* with 53.4% relationship, which indicated that all of the parameters were independently contributing to the estimation of the wheat yields.

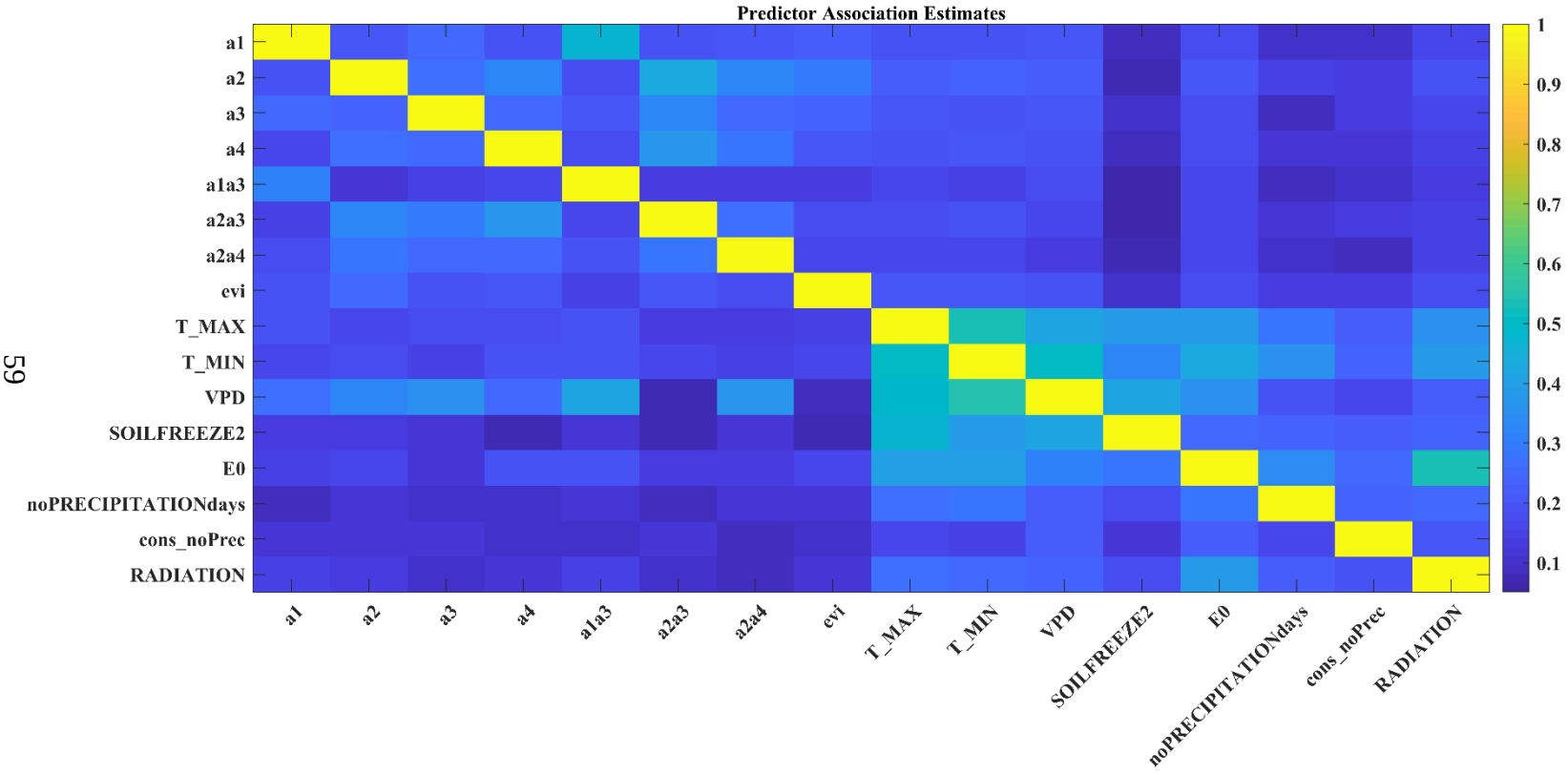

Figure4.12 Predictor association estimates of all parameters.

4.4 Yield estimation using three different machine learning approaches

This study was done with Landsat 8 – a medium-resolution satellite data – using the procedures that give best results when used with hyperspectral imagery. However, the goal of this research is to obtain the best possible outcomes with free and accessible data for repeatable research. Three different machine learning approaches were applied to the dataset to compare their performance with each other. These methods are namely Generalized Linear Model, Artificial Neural Network and Random Forests all ran in MATLAB™. Due to the limited number of data that could be used in this study, the validation procedure was conducted by using cross-validation techniques. 10-fold-cross-validation is often used in the literature and therefore was also used in all our models (GLM, ANN and RF) to establish a certain consistency.

4.4.1 The Generalized Linear Model (GLM) approach

Linear regression models describe the linear relationship between the response and the predictive parameters. There may, however, be a nonlinear relationship between the parameters most of the times. Nonlinear regression describes these general non-linear models. The GLM is a special class of nonlinear models that use linear methods (Generalized Linear Models 2019) for regression fitting.

The GLM model is selected as linear and the distribution as 'Poisson'. The result of the GLM gives a coefficient of determination (R^2) of the real yield and the estimated yield of 0.64 and an RMSE of 31.53 when only the abundances and their selected interactions are used as input parameters. When the selected agrometeorological parameters and NDVI are used as the input set for the model, the R^2 is 0.60 and RMSE is 33.53. The R^2 reached its highest value of 0.67 with an RMSE of 31.98 when all the selected parameters in Table 4.4 were used in the GLM. Accuracies of all relevant correlations can be found in Table 4.5.

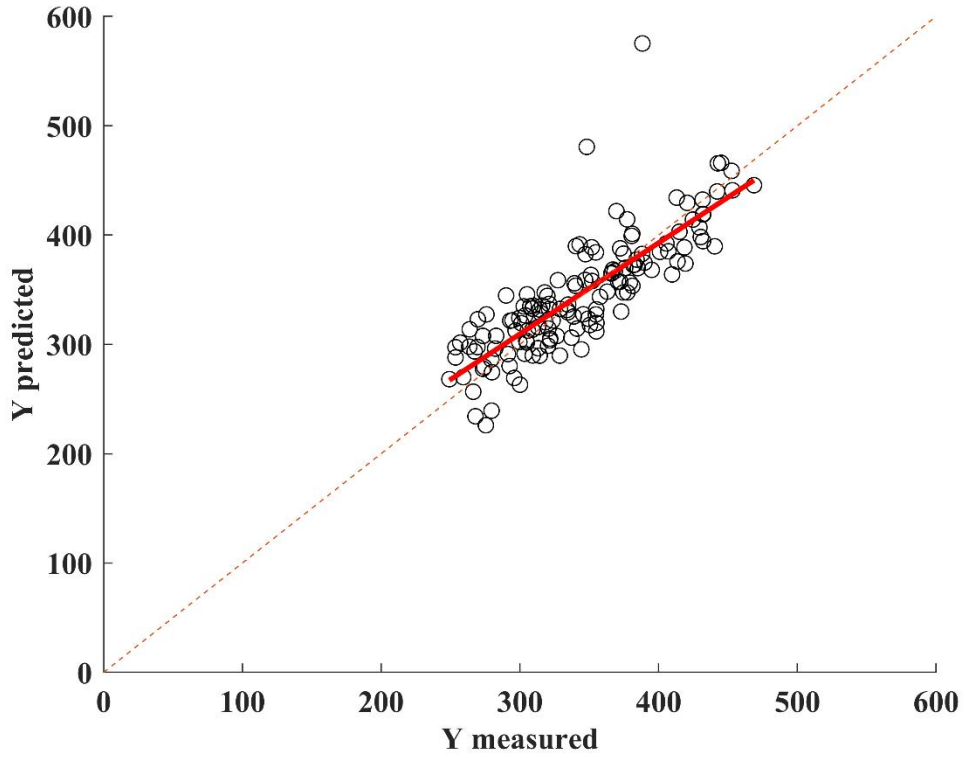

Figure 4.13 Relationship between real and predicted yields found by using all the parameters of Table 5.3 in GLM algorithm (R^2 = 0.67).

4.4.2 The neural network approach

The matrix of all abundances and their interactions are selected as the input set of the Levenberg-Marquardt backpropagation algorithm, which is the ANN algorithm used in this study. The target is the yield matrix, while the hidden layer size changes between 5 to 16 depending on the number of parameters according to the 2/3 rule of thumb.

The R^2 is 0.63 when the abundances and their interactions are used as the input set. The hidden layer size is five and the RMSE was calculated as 31.86. When the input set of the selected agrometeorological parameters and NDVI are used as the input set, the R^2 is found as 0.75 with an RMSE of 26.52. When all the selected parameters of Table 4.4 are used, making the hidden layer size 11, the R^2 reaches

its top value of 0.78 with an RMSE of 25.00. The relationship of the real and predicted yields of the neural networks approach can be seen in Figure 4.13.

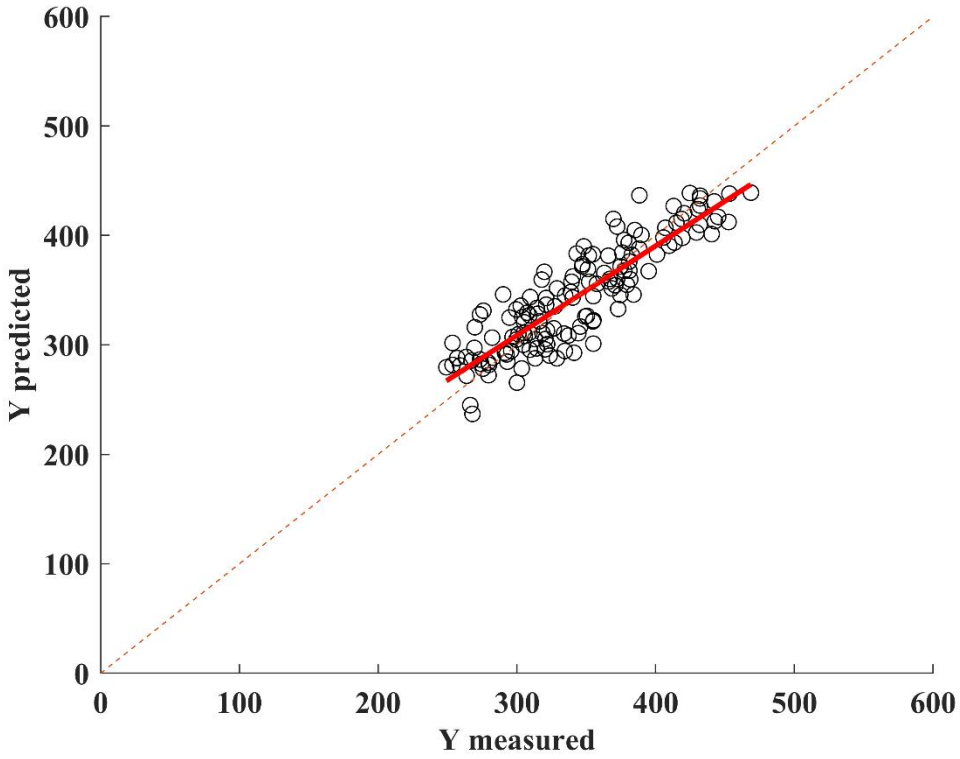

Figure 4.14 Relationship between real and predicted yields found by using all the parameters of Table 4.5 in ANN algorithm (R^2 = 0.78).

4.4.3 The random forests approach

The selected parameters are used in the Treebagger Algorithm, growing 500 trees in the forest. The R^2 is found to be 0.63 with an RMSE of 32.45 when the abundances and their selected interactions are used as the input set of six parameters. The R^2 improves significantly when agrometeorological parameters and NDVI are used, to 0.78 with RMSE equal to 24.72 and running the selected parameters of Table 4.4 increases the R^2 to 0.82 (RMSE=22.51) reaching the best value of all the tests. The relations between real and predicted yields when all

parameters of Table 4.4 were used in GLM (a), ANN (b) and RF (c) can be found in Figure 4.14.

(Heremans et al., 2015) used 262 input variables consisting of overall fertilizer use, 27 meteorological parameters and 234 cumulative NDVI values for 12 years. Their results showed over 0.80 R^2 values for the RF. The dataset used in this study was relatively small compared to their set and it can be seen that the RF showed similar results when only agrometeorological parameters and NDVI were used as inputs (R^2 = 0.77).

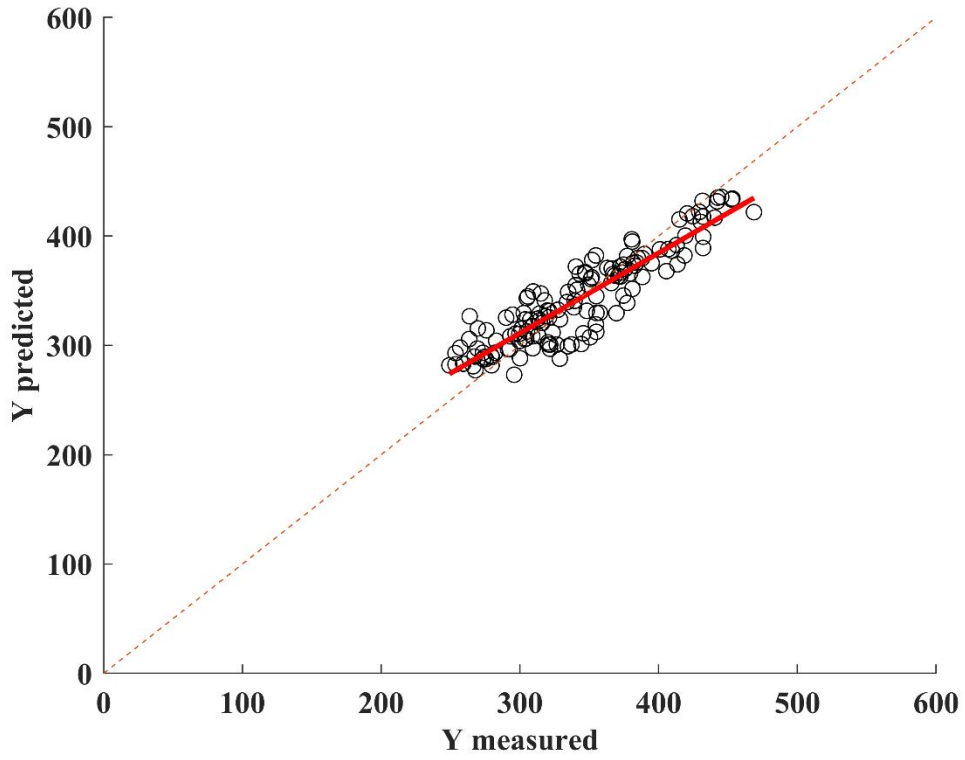

Figure 4.15 The relations between real and predicted yields using abundances and their interactions and selected agrometeorological parameters in RF (R^2=0.82).

The correlations of real and predicted yields when different machine learning algorithms are applied to all possible combination of parameters of Table 4.4, namely all parameters used in this study, can be found in Table 4.5.

Table 4.5 The real vs. predicted yield accuracies and RMSE, of the applied methods: GLM, ANN and RF according to different parameter combinations for the training sets.

Parameters	GLM		ANN			RF	
	R^2	RMSE	Hidden layer size	R^2	RMSE	R^2	RMSE
Four abundances	0.59	33.71	3	0.73	27.73	**0.84**	22.32
Four abundances and their selected interactions	0.68	29.85	5	0.73	27.4	**0.84**	22.1
Agrometeorological parameters	0.68	29.73	5	0.83	22.05	**0.89**	18.11
Agrometeorological parameters and NDVI	0.68	29.61	6	0.88	18.64	**0.89**	17.78
All selected parameters	0.82	22.09	11	0.85	20.58	**0.93**	15.21
All parameters	0.83	21.40	16	0.78	24.65	**0.93**	15.11

Table 4.6 The real vs. predicted yield accuracies and RMSE, of the applied methods: GLM, ANN and RF according to different parameter combinations for the test sets.

Parameters	GLM		ANN			RF	
	R^2	RMSE	Hidden layer size	R^2	RMSE	R^2	RMSE
Four abundances	0.56	34.89	3	**0.62**	32.55	**0.62**	32.45
Four abundances and their selected interactions	**0.64**	31.53	5	0.63	31.86	0.63	32.32
Agrometeorological parameters	0.63	32.05	5	0.76	25.92	**0.78**	24.96
Agrometeorological parameters and NDVI	0.6	33.53	6	0.75	26.52	**0.78**	24.72
All selected parameters	0.67	31.98	11	0.78	25	**0.82**	23.51
All parameters	0.61	36.21	16	0.61	55.59	**0.80**	24.15

Table 4.5 shows the results of the machine learning algorithms applied to different datasets consisting of different parameter combinations by using only the training sets. Table 4.6 shows the results of the same procedures applied to test sets, which were obtained by using 10-fold cross validation technique. The training set results showed that RF was highly capable of estimating the yields even when only the abundances were used (R^2=0.84) and when all the selected parameters were used, the R^2 reached 0.93. Results of the three different estimation methods using the test sets can be compared by examining Table 4.6. The best outcomes are obtained when all the parameters of Table 4.4 are used, which are selected from Table 4.3 and they are the parameters that make the greatest contribution to the artificial intelligence models when predicting the yields. RF can predict 82% of the yields when all parameters are used. And the importance of selecting the most important parameters is openly demonstrated in the results too. The accuracies increased for all estimation methods quite significantly when the selected parameters were used, with the smallest increase observed in RF as it is resistant to overfitting by nature. This result also showed that the parameters selected by using an RF algorithm also serves the other methods, even more than it serves RF itself.

ANN and GLM normally need more predictors for better accuracy, whereas RF improves model accuracy by randomly changing the predictors and training data for each decision tree. RF is resistant to noise within the data and also to over-fitting problem. Despite these facts, NDVI seemed to have no major contribution to the accuracy when used with the agrometeorological parameters to find the yields. Agrometeorological parameters are highly capable of estimating the yield on their own with RF (78% accuracy). ANN can also predict with good enough accuracy, over 76% but GLM cannot predict the yield with agrometeorological parameters as good as the other models (63%).

RMSE is better for RF in most of the cases, although RF and ANN performed close enough except when the set of selected parameters is used to predict the yields. RMSE is calculated as low as 22.5 when only selected parameters are used. These

results show that when using all the important parameters and their interactions, RF is the best method for estimating wheat yield.

This research proves the importance and power of extraction of intimately mixed endmembers, presumably the photosynthetic pigments in yield estimation. This is succeeded despite the fact that the intimate mixture of photosynthetic pigments in the wheat crop is treated linearly when unmixing with R-CoNMF. In addition to finding the endmembers with R-CoNMF, the yield estimation performance increases significantly after the optimization of the endmembers with the GLM algorithm. Before the optimization was included in the calculation steps, the R^2 of the abundances and their interactions could only go as high as 0.55 and RMSE 35.10 with the RF model, whereas after the optimization it reached 0.63 with an RMSE of 32.32.

The endmembers are related to the pigments as they are similar to the spectral signatures of the pigments when reflected from Landsat 8 bands. When the interactions are also included, it can be said that a bilinear method for a non-linear mixture is used (Heylen et al., 2014) and the predictions get better. Interactions may make a greater contribution to increasing R^2 than some of the abundances themselves (Figure 4.7, Figure 4.10). These abundances and their interaction can estimate almost 65% of the yield all by themselves in the test set created by using 10-fold cross-validation when using medium resolution Landsat 8 data, and over 82% when all the important parameters are used. When a fine decomposition algorithm of the inner structure of the crops with the evolving technology is available, the yield estimate is bound to be better.

Heremans et al., (2015) had used 262 input variables consisting of overall fertilizer use, 27 meteorological parameters and 234 cumulative NDVI values for 12 years. Their results showed over 0.80 R^2 values for RF. The dataset used in this study is relatively quite small compared to their set, yet still, RF shows similar results when only agrometeorological parameters and NDVI were used as inputs (R^2 = 0.77). Our study is done with agrometeorological parameters and endmembers calculated

with the data of only one year (2015) using only 17 selected parameters which proves its efficiency and easy data collection process and calculations when compared to not only Heremans et al., (2015) study but many other studies that were conducted using agrometeorological parameters and NDVI.

The most important thing to note is that the results of this study, especially the importance of the abundances in yield estimation, would probably have increased rapidly if hyperspectral satellite data were to be used. This is mainly because there would be more bands to use and secondly, the used algorithms are actually developed for hyperspectral data.

CHAPTER 5

CONCLUSION

This thesis demonstrates a field-level wheat yield estimation method using the first four bands of pure Landsat 8 pixels of wheat crop whose performance is tested on data from 142 fields in 31 provinces, belonging to different regions with distinct climatic conditions. Harvester data obtained from the Ministry of Agriculture is used as ground truth of these fields. With the linear unmixing algorithm called R-CoNMF that do not need a pure pixel for the unmixing process, we are able to unmix intimately mixed pure wheat crop pixel containing almost only photosynthetic pigments to find the endmembers representing these pigments. The endmembers are further optimized for predicting the yields more accurately by them and their interactions. The endmembers calculated by the algorithm show a similar pattern to the spectral signatures of chlorophylls and carotenoids, whose spectral signatures are processed with Landsat 8 bands to obtain a view of how they would look from a medium resolution satellite point of view. Abundances found from the endmembers by using the same algorithm acts as new indices and the nonlinearity was handled by including the interactions of the abundances in the parameter list of the three machine learning algorithms (GLM, ANN and RF) that are used to predict the yields.

GLM predicts over 64% of the yields by only using the abundances and the most important interactions. Adding the agrometeorological parameters and the VIs in the picture helps RF to attain an R^2 of 0.82, which can be considered a big success considering a multispectral satellite is used in the process.

This thesis contributed to the literature by

- demonstrating a novel point of view in estimating the yields, by using soft computing methods for unmixing the pigments within a crop,

- introducing a novel point of view in unmixing intimate mixtures, without using non-linear methods,
- improving the timeline of estimation of the yield, as with the method given in this study, the yield can be estimated at least a month before the harvest and
- reducing the number of sources and parameters that need to be used for yield estimation. By doing this, the focus could be directed on the parameters that are easily accessible or can be calculated with no additional cost.

For future studies Sentinel-2 data can be used, as it is available freely just like Landsat 8, it has more bands and a 5-day temporal resolution, which is also better than that of Landsat 8, which is 16 days. It would be better to use hyperspectral satellite data obtained from future hyperspectral space missions like HyspIRI to perform the analysis and find the yields, as every increase in the number of bands would make a finer endmember calculation, resulting in finding more accurate yields.

The most important future work would be to carry out measurements of ground truth with spectroradiometers and laboratory work to determine to what extend the endmembers actually represent the photosynthetic pigments within the crops. It would also be a good future work to take the absorption and reflection measurements of the wheat leaves in the field, in order to calculate the actual values for the transformation from the absorbance spectra of the pigments to reflectance spectra for wheat. Implementation of the process for crops other than wheat would also be a fruitful study. The most useful future study; however, would be to embed the process in Google Earth Engine and record the yield estimation of the wheat fields on a weekly basis, if Sentinel-2 data is to be used. This can be especially easier to do and very useful if the endmember signatures are proven not to change throughout the years. However, the algorithm needs to be integrated with agricultural parcel segmentation and crop type detection algorithms to have an operational service.

REFERENCES

Andrea Toreti. (2014). *Gridded Agro-Meteorological Data in Europe. European Commission, Joint Research Centre (JRC) [Dataset] PID: http://data.europa.eu/89h/jrc-marsop4-7-weather_obs_grid_2019.* Retrieved from http://data.europa.eu/89h/jrc-marsop4-7-weather_obs_grid_2019

Apostol, S., Viau, A., Tremblay, N., Briantais, J.-M., Prasher, S., Parent, L.-E., & Moya, I. (2003). Laser-induced fluorescence signatures as a tool for remote monitoring of water and nitrogen stresses in plants. *Canadian Journal of Remote Sensing, 29*(1), 57-65.

Azzari, G., Jain, M., & Lobell, D. (2017). Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries. *Remote Sensing of Environment*.

Barker, D., Seaton, G., & Robinson, S. (1997). Internal and external photoprotection in developing leaves of the CAM plant Cotyledon orbiculata. *Plant, Cell and Environment, 20*(5), 617-624.

Bartley, G., & Scolnik, P. (1995). Plant carotenoids: pigments for photoprotection, visual attraction, and human health. *The Plant cell, 7*(7), 1027-38.

Basso, B., Cammarano, D., & Carfagna, E. *Review of Crop Yield Forecasting Methods and Early Warning Systems.*

Basso, B., Ritchie, J., Pierce, F., Braga, R., & Jones, J. (2001). Spatial validation of crop models for precision agriculture. *Agricultural Systems, 68*(2), 97-112.

Basso, B., T.R. McVicar, B. (2007). Remote sensing and GIS applications in agrometeorology. In B. Basso, B., T.R. McVicar, *Remote sensing and GIS applications in agrometeorology*.

Benedetti, R., & Rossini, P. (1993). On the use of NDVI profiles as a tool for agricultural statistics: The case study of wheat yield estimate and forecast in Emilia Romagna. *Remote Sensing of Environment, 45*(3), 311-326.

Bilger, W., Björkman, O., & Thayer, S. (1989). Light-induced spectral absorbance changes in relation to photosynthesis and the epoxidation state of xanthophyll cycle components in cotton leaves. *Plant physiology, 91*(2), 542-51.

Blackburn, G. (2006). Hyperspectral remote sensing of plant pigments. *Journal of Experimental Botany, 58*(4), 855-867.

Buschmann, C., & Nagel, E. (1993). In vivo spectroscopy and internal optics of leaves as basis for remote sensing of vegetation. *International Journal of Remote Sensing, 14*(4), 711-722.

Cai, Y., Guan, K., Lobell, D., Potgieter, A. B., Wang, S., Peng, J., Asseng, S. (2019). Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agricultural and Forest Meteorology, 274*, 144-159. doi:https://doi.org/10.1016/j.agrformet.2019.03.010

Cai, Y., Guan, K., Lobell, D., Potgieter, A., Wang, S., Peng, J., Peng, B. (2019). Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agricultural and Forest Meteorology*.

Chappelle, E., Kim, M., & McMurtrey, J. (1992). Ratio analysis of reflectance spectra (RARS): An algorithm for the remote estimation of the concentrations of chlorophyll A, chlorophyll B, and carotenoids in soybean leaves. *Remote Sensing of Environment, 39*(3), 239-247.

Coefficient of Determination, M. (n.d.). *Coefficient of Determination (R-Squared) - MATLAB & Simulink*. Retrieved from https://www.mathworks.com/help/stats/coefficient-of-determination-r-squared.html

Cohen, Y., Alchanatis, V., Meron, M., Saranga, Y., & Tsipris, J. (2005). Estimation of leaf water potential by thermal imagery and spatial analysis. *Journal of Experimental Botany, 56*(417), 1843-1852.

Crop simulation model - Wikipedia. (2020). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Crop_simulation_model

Dash, J., & Curran, P. (2004). The MERIS terrestrial chlorophyll index. *International Journal of Remote Sensing, 25*(23), 5403-5413.

Datt, B. (1998). Remote sensing of chlorophyll a, chlorophyll b, chlorophyll a+b, and total carotenoid content in eucalyptus leaves. *Remote Sensing of Environment, 66*(2), 111-121.

Dawson, T., Curran, P., & Plummer, S. (1998). LIBERTY - Modeling the effects of Leaf Biochemical Concentration on Reflectance Spectra. *Remote Sensing of Environment, 65*(1), 50-60.

Defourny, P., Bontemps, S., Bellemans, N., Cara, C., Dedieu, G., Guzzonato, E., Koetz, B. (2019). Near real-time agriculture monitoring at national scale at parcel resolution: Performance assessment of the Sen2-Agri automated system in various cropping systems around the world. *Remote Sensing of Environment*.

Demarez, V. (1999). Seasonal variation of leaf chlorophyll content of a temperate forest. Inversion of the PROSPECT model. *International Journal of Remote Sensing, 20*(5), 879-894.

Dimitruk, P., Schermelleh-Engel, K., Kelava, A., & Moosbrugger, H. (2007). Challenges in nonlinear structural equation modeling. *Methodology, 3*(3), 100-114.

Dorigo, W., Zurita-Milla, R., de Wit, A., Brazile, J., Singh, R., & Schaepman, M. (2007). A review on reflective remote sensing and data assimilation

techniques for enhanced agroecosystem modeling. *International Journal of Applied Earth Observation and Geoinformation, 9*(2), 165-193.

Dourado-Neto, D., Teruel, D., Reichardt, K., Nielsen, D., Frizzone, J., & Bacchi, O. (1998). Principles of crop modeling and simulation: I. uses of mathematical models in agricultural science. *Scientia Agricola, 55*(Special Issue), 46-50.

Feild, T., Lee, D., & Holbrook, N. (2001). Why leaves turn red in autumn. The role of anthocyanins in senescing leaves of red-osier dogwood. *Plant physiology, 127*(2), 566-74.

Fortin, J., Anctil, F., Parent, L.-É., & Bolinder, M. (2010). A neural network experiment on the site-specific simulation of potato tuber growth in Eastern Canada. *Computers and Electronics in Agriculture, 73*(2), 126-132.

Franch, B., Vermote, E., Skakun, S., Roger, J., Becker-Reshef, I., & Murphy, E. J. (2019). Remote sensing based yield monitoring: Application to winter wheat in United States and Ukraine. *International Journal of Applied Earth Observation and Geoinformation, 76*, 112-127. doi:https://doi.org/10.1016/j.jag.2018.11.012.

Franch, B., Vermote, E., Skakun, S., Roger, J., Becker-Reshef, I., Murphy, E., & Justice, C. (2019). Remote sensing based yield monitoring: Application to winter wheat in United States and Ukraine. *International Journal of Applied Earth Observation and Geoinformation*.

Frank B. Salisbury, C. (1992). *Plant physiology* (4th ed.). Wadsworth Pub. Co.

Gamon, J. (2010). The Photochemical Reflectance Index (PRI) – a measure of photosynthetic light-use efficiency. *International Journal of Remote Sensing*.

Gamon, J., & Surfus, J. (1999). *Assessing Leaf Pigment Content and Activity with a Reflectometer*.

Gamon, J., Peñuelas, J., & Field, C. (1992). A narrow-waveband spectral index that tracks diurnal changes in photosynthetic efficiency. *Remote Sensing of Environment, 41*(1), 35-44. doi:https://doi.org/10.1016/0034-4257(92)90059-S

Ganapol, B., Johnson, L., Hammer, P., Hlavka, C., & Peterson, D. (1998). LEAFMOD: A New Within-Leaf Radiative Transfer Model. *Remote Sensing of Environment, 63*(2), 182-193.

Generalized Linear Models - MATLAB & Simulink. (n.d.). Retrieved from https://www.mathworks.com/help/stats/generalized-linear-regression.html

Gitelson, A., & Merzlyak, M. (1994). Quantitative estimation of chlorophyll-a using reflectance spectra: Experiments with autumn chestnut and maple leaves. *Journal of Photochemistry and Photobiology B: Biology, 22*(3), 247-252.

Gitelson, A., & Merzlyak, M. (1994). Spectral Reflectance Changes Associated with Autumn Senescence of Aesculus hippocastanum L. and Acer platanoides L. Leaves. Spectral Features and Relation to Chlorophyll Estimation. *Journal of Plant Physiology, 143*(3), 286-292.

Gitelson, A., & Merzlyak, M. (1996). Signature analysis of leaf reflectance spectra: Algorithm development for remote sensing of chlorophyll. *Journal of Plant Physiology, 148*(3-4), 494-500.

Gitelson, A., & Solovchenko, A. (2018). Non-invasive quantification of foliar pigments: Possibilities and limitations of reflectance- and absorbance-based approaches. *Journal of Photochemistry and Photobiology B: Biology, 178*(September 2017), 537-544.

Gitelson, A., Gritz, Y., & Merzlyak, M. (2003). Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. *J. Plant Physiol, 160*, 271-282.

Gitelson, A., Merzlyak, M., & Chivkunova, O. (2001). *Optical Properties and Nondestructive Estimation of Anthocyanin Content in Plant Leaves* ¶.

Gitelson, A., Viña, A., Ciganda, V., Rundquist, D., & Arkebauer, T. (2005). Remote estimation of canopy chlorophyll content in crops. *Geophysical Research Letters, 32*(8), 1-4.

Gouveia-Neto, A. (2011). Abiotic Stress Diagnosis via Laser Induced Chlorophyll Fluorescence Analysis in Plants for Biofuel. In A. Gouveia-Neto, *IntechOpen* (pp. 3-22).

Groten, S. (1993). NDVI—crop monitoring and early yield assessment of Burkina Faso. *International Journal of Remote Sensing, 14*(8), 1495-1515.

Guo, P.-T., Li, M.-F., Luo, W., Tang, Q.-F., Liu, Z.-W., & Lin, Z.-M. (2015). Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach. *Geoderma, 237-238*, 49-59.

Gutiérrez, P., López-Granados, F., Peña-Barragán, J., Jurado-Expósito, M., & Hervás-Martínez, C. (2008). Logistic regression product-unit neural networks for mapping Ridolfia segetum infestations in sunflower crop using multitemporal remote sensed data. *Computers and Electronics in Agriculture, 64*(2), 293-306.

Heremans, S., Dong, Q., Zhang, B., Bydekerke, L., & Van Orshoven, J. (2015). Potential of ensemble tree methods for early-season prediction of winter wheat yield from short time series of remotely sensed normalized difference vegetation index and in situ meteorological data. *Journal of Applied Remote Sensing, 9*(1), 097095.

Heylen, R., Parente, M., & Gader, P. (2014). A review of nonlinear hyperspectral unmixing methods. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 7*(6), 1844-1868.

Huang, J., Wei, C., Zhang, Y., Blackburn, G., Wang, X., Wei, C., & Wang, J. (2015). Meta-Analysis of the Detection of Plant Pigment Concentrations Using Hyperspectral Remotely Sensed Data.

Huang, W., Zhou, X., Kong, W., & Ye, H. (2018). Monitoring Crop Carotenoids Concentration by Remote Sensing. In L. Q. Zepka, E. Jacob-Lopes, & V. V. Rosso, *Progress in Carotenoid Research.* InTech. doi:10.5772/intechopen.78239

Huang, W.-D., Lin, K.-H., Hsu, M.-H., Huang, M.-Y., Yang, Z.-W., Chao, P.-Y., & Yang, C.-M. (2014). *Eliminating interference by anthocyanin in chlorophyll estimation of sweet potato (Ipomoea batatas L.) leaves.*

Huang, Y., Lan, Y., Thomson, S., Fang, A., Hoffmann, W., & Lacey, R. (2010). Development of soft computing and applications in agricultural and biological engineering. *Computers and Electronics in Agriculture, 71*, 107-127.

Hughes, N., Morley, C., & Smith, W. (2007). Coordination of anthocyanin decline and photosynthetic maturation in juvenile leaves of three deciduous tree species. *New Phytologist, 175*(4), 675-685.

Hunt, M., Blackburn, G., Carrasco, L., Redhead, J., & Rowland, C. (2019). High resolution wheat yield mapping using Sentinel-2. *Remote Sensing of Environment, 233*(December 2018), 111410.

Improve linear regression model by adding or removing terms - MATLAB step. (n.d.). Retrieved from https://www.mathworks.com/help/stats/linearmodel.step.html

Jacquemoud, S., & Baret, F. (1990). PROSPECT: A model of leaf optical properties spectra. *Remote Sensing of Environment, 34*(2), 75-91.

Jiang, P., Thelen, K. (2004). Effect of Soil and Topographic Properties on Crop Yield in a North-Central Corn-Soybean Cropping System. *AGRONOMY JOURNAL, 96*, pp. 252-258.

Jones, H., & Vaughan, R. (2010). *Remote sensing of vegetation : principles, techniques, and applications.* Oxford University Press.

JRC. (n.d.). *Weather Monitoring - Agri4castWiki*. Retrieved from https://marswiki.jrc.ec.europa.eu/agri4castwiki/index.php/Weather_Monitoring#Interpolation

Kang, Y., & Özdoğan, M. (2019). Field-level crop yield mapping with Landsat using a hierarchical data assimilation approach. *Remote Sensing of Environment, 228*, 144-163. doi:https://doi.org/10.1016/j.rse.2019.04.005

Kang, Y., & Özdoğan, M. (2019). Field-level crop yield mapping with Landsat using a hierarchical data assimilation approach. *Remote Sensing of Environment, 228*, 144-163.

Kaul, M., Hill, R., & Walthall, C. (2005). Artificial neural networks for corn and soybean yield prediction. *Agricultural Systems, 85*(1), 1-18.

Kenny, D., & Judd, C. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin, 96*(1), 201-210.

Keshava, N., & Mustard, J. (2002). Spectral unmixing. *IEEE Signal Processing Magazine, 19*(1), 44-57.

Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika, 65*(4), 457-474.

Kravchenko, A., & Bullock, D. (2000). Correlation of corn and soybean grain yield with topography and soil properties. *AGRONOMY JOURNAL, 92*, 75-83.

Kumar, P., Prasad, R., Gupta, D., Mishra, V., Vishwakarma, A., Yadav, V., Avtar, R. (2018). Geocarto International Estimation of winter wheat crop growth

parameters using time series Sentinel-1A SAR data Estimation of winter wheat crop growth parameters using time series Sentinel-1A SAR data. *Geocarto InternatIonal, 33*(9), 942-956.

Lachman, J., Martinek, P., Kotíková, Z., Orsák, M., & Šulc, M. (2017). Genetics and chemistry of pigments in wheat grain – A review. *Journal of Cereal Science, 74*, 145-154.

le Maire, G., François, C., & Dufrêne, E. (2004). Towards universal broad leaf chlorophyll indices using PROSPECT simulated database and hyperspectral reflectance measurements. *Remote Sensing of Environment, 89*(1), 1-28.

Leroux, L., Castets, M., Baron, C., Escorihuela, M.-J., Bégué, A., & Lo Seen, D. (2019). Maize yield estimation in West Africa from crop process-induced combinations of multi-domain remote sensing indices. *European Journal of Agronomy, 108*, 11 - 26. doi:https://doi.org/10.1016/j.eja.2019.04.007

Leroux, Louise; Castets, Mathieu; Baron, Christian; Escorihuela, Maria-Jose; Bégué, Agnès; Lo Seen, D. (2019). Maize yield estimation in West Africa from crop process-induced combinations of multi-domain remote sensing indices. *European Journal of Agronomy, 108*, 11 - 26.

Li, J., Bioucas-Dias, J., Plaza, A., & Liu, L. (2016). Robust Collaborative Nonnegative Matrix Factorization for Hyperspectral Unmixing. *IEEE Transactions on Geoscience and Remote Sensing, 54*(10), 6076-6090.

Li, Z., Wang, J., Tang, H., Huang, C., Yang, F., Chen, B., Ge, Y. (2016). Predicting Grassland Leaf Area Index in the Meadow Steppes of Northern China: A Comparative Study of Regression Approaches and Hybrid Geostatistical Methods. *Remote Sensing, 8*(8), 632.

Liang, L., Di, L., Zhang, L., Deng, M., Qin, Z., Zhao, S., & Lin, H. (2015). Estimation of crop LAI using hyperspectral vegetation indices and a hybrid inversion method. *Remote Sensing of Environment, 165*, 123-134.

Lichtenthaler, H. (1987). Chlorophylls and Carotenoids: Pigments of Photosynthetic Biomembranes. *Methods in Enzymology, 148*(C), 350-382.

Lichtenthaler, H., Gitelson, A., & Lang, M. (1996). Non-Destructive Determination of Chlorophyll Content of Leaves of a Green and an Aurea Mutant of Tobacco by Reflectance Measurements. *Journal of Plant Physiology, 148*(3-4), 483-493.

Liu, B., Yue, Y., Li, R., Shen, W., & Wang, K. (2014). Plant leaf chlorophyll content retrieval based on a field imaging spectroscopy system. *Sensors (Switzerland), 14*(10), 19910-19925.

Lobell, D., Asner, G., Ortiz-Monasterio, J., & Benning, T. (2003). Remote sensing of regional crop production in the Yaqui Valley, Mexico: Estimates and uncertainties. *Agriculture, Ecosystems and Environment, 94*(2), 205-220.

Lobell, D., Cassman, K., & Field, C. (2010). Crop Yield Gaps: Their Importance, Magnitudes, and Causes. *Ssrn*.

Lobell, D., Ortiz-Monasterio, J., Asner, G., Naylor, R., & Falcon, W. (2005). Combining field surveys, remote sensing, and regression trees to understand yield variations in an irrigated wheat landscape. *Agronomy Journal, 97*(1), 241-249.

Magney, T., Eitel, J., & Vierling, L. (2017). Mapping wheat nitrogen uptake from RapidEye vegetation indices. *Precision Agriculture, 18*(4), 429-451.

Maier, S., Lüdeker, W., & Günther, K. (1999). SLOP: A Revised Version of the Stochastic Model for Leaf Optical Properties. *Remote Sensing of Environment, 68*(3), 273-280.

MathWorks. (2019). *Improve linear regression model by adding or removing terms - MATLAB step*. Retrieved 07 18, 2019, from www.mathworks.com/help/stats/linearmodel.step.html

MathWorks, B.. *Backpropagation (Neural Network Toolbox)*. Retrieved from https://edoras.sdsu.edu/doc/matlab/toolbox/nnet/backpr11.html

MathWorks, B.. *Bootstrap Aggregation (Bagging) of Regression Trees - MATLAB & Simulink*. Retrieved from https://www.mathworks.com/help/stats/regression-treeBagger-examples.html

MathWorks, D.. *Create bag of decision trees - MATLAB - MathWorks United Kingdom*. Retrieved from https://uk.mathworks.com/help/stats/treebagger.html

MathWorks, L.. *Linear Regression - MATLAB & Simulink - MathWorks United Kingdom*. Retrieved from http://uk.mathworks.com/help/matlab/data_analysis/linear-regression.html

Merzlyak, M., Gitelson, A., Chivkunova, O., & Rakitin, V. (1999). Non-destructive optical detection of pigment changes during leaf senescence and fruit ripening. *Physiologia Plantarum, 106*(1), 135-141.

Models, G. L. (2019). *Generalized Linear Models - MATLAB*. Retrieved 07 18, 2019, from www.mathworks.com/help/stats/generalized-linear-regression.html

Monitoring Agricultural ResourceS (MARS) | EU Science Hub. (n.d.). Retrieved from https://ec.europa.eu/jrc/en/mars

Morellos, A., Pantazi, X.-E., Moshou, D., Alexandridis, T., Whetton, R., Tziotzios, G., Mouazen, A. (2016). Machine Learning based Prediction of Soil Total Nitrogen, Organic Carbon and Moisture Content by Using VIS-NIR Spectroscopy. *Biosystems Engineering, 152*, 104-116.

Mulla, D. (2013). Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. *Biosystems Engineering, 114*(4), 358-371.

Neill, S., & Gould, K. (2000). Optical properties of leaves in relation to anthocyanin concentration and distribution. *Canadian Journal of Botany, 77*(12), 1777-1782.

Pantazi, X., Moshou, D., Mouazen, A., Kuang, B., & Alexandridis, T. (2014). Application of Supervised Self Organising Models for Wheat Yield Prediction. *International Federation for Information Processing*, 556-565.

Park, S., Hwang, C., & Vlek, P. (2005). Comparison of adaptive techniques to predict crop yield response under varying soil and land management conditions. *Agricultural Systems, 85*(1), 59-81.

Patel, S., Vediya, S., Science, S., & Modasa, C. (2013). Separation of photosynthetic pigments in Spirogyra specis by means of thin layer Chromatography from Sola lake , Ahmedabad , Gujarat. *INTERNATIONAL JOURNAL OF PHARMACY & LIFE SCIENCES, 4*(7), 2819-2822.

Penuelas, J., Baret, F., & Filella, I. (1995). Semi-empirical indices to assess carotenoids/chlorophyll a ratio from leaf spectral reflectance. *Photosynthetica, 31(2)*, 221-230.

Penuelas, J., Gamon, J.A., Freeden, A.L., Merino, J., F. (1994). Reflectance Indices Associated with Physiological Changes in Nitrogen and Water Limited Sunflower Leaves. *Remote Sensing of Environment, 48*, 135 - 146.

Pinter, P., Hatfield, J., Schepers, J., Barnes, E., Moran, M., Pinter, P., Daughtry, C. (2003). Remote Sensing for Crop Management. *PHOTOGRAMMETRIC ENGINEERING & REMOTE SENSING*.

Plant, R. (2001). Site-specific management: the application of information technology to crop production. *Computers and Electronics in Agriculture, 30*(1-3), 9-29.

Renzullo, L., Blanchfield, A., Guillermin, R., Powell, K., & Held, A. (2006). Comparison of PROSPECT and HPLC estimates of leaf chlorophyll

contents in a grapevine stress study. *International Journal of Remote Sensing, 27*(4), 817-823.

Sabo, M., Teklic, T., & Vidovic, I. (2002). Photosynthetic productivity of two winter wheat varieties (Triticum aestivum L.). *ROSTLINNÁ VÝROBA, 48*(2), 80 - 86.

Salvador, P., Gómez, D., Sanz, J., & Casanova, J. (2020). Estimation of Potato Yield Using Satellite Data at a Municipal Level: A Machine Learning Approach. *ISPRS International Journal of Geo-Information, 9*(6), 343.

Sayago, S., & And Bocco, M. (2018). Crop yield estimation using satellite images: comparison of linear and non-linear models. *AgriScientia, 35*, 1 - 9.

Schepers, J., Blackmer, T., Wilhelm, W., & Resende, M. (19961). Transmittance and Reflectance Measurements of CornLeaves from Plants with Different Nitrogen and Water Supply. *Journal of Plant Physiology, 148*(5), 523-529.

Sharma, B., & Venugopalan, P. (2014). Comparison of Neural Network Training Functions for Hematoma Classification in Brain CT Images. *IOSR Journal of Computer Engineering (IOSR-JCE), 16*(1), 31-35.

Sheela, K., & Deepa, S. (2013). Review on Methods to Fix Number of Hidden Neurons in Neural Networks. *Mathematical Problems in Engineering, 2013*, 1-11.

Shuxiang Xu and Ling Chen. (2008). A Novel Approach for Determining the Optimal Number of Hidden Layer Neurons for FNN's and Its Application in Data Mining. *5th International Conference on Information Technology and Applications (ICITA).*

Sims, D., & Gamon, J. (2002). Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages. *Remote Sensing of Environment, 81*(2-3), 337-354.

Somers, B., Asner, G., Tits, L., & Coppin, P. (2011). Endmember variability in Spectral Mixture Analysis: A review. *Remote Sensing of Environment, 115*(7), 1603-1616.

Song, X., Yang, G., Yang, C., Wang, J., Cui, B., Song, X., . . . Cui, B. (2017). Spatial Variability Analysis of Within-Field Winter Wheat Nitrogen and Grain Quality Using Canopy Fluorescence Sensor Measurements. *Remote Sensing, 9*(3), 237.

Stas, M., Orshoven, J., Dong, Q., Heremans, S., & Zhang, B. (2016). A comparison of machine learning algorithms for regional wheat yield prediction using NDVI time series of SPOT-VGT. *2016 Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics).*

Steele, M., Gitelson, A., Rundquist, D., Merzlyak, M., Steele, M., Gitelson, A., & Rundquist, D. (2009). Nondestructive Estimation of Anthocyanin Content in Grapevine Leaves. *American Journal of Enology and Viticulture*.

Steyn, W., Wand, S., Holcroft, D., & Jacobs, G. (2002). Anthocyanins in vegetative tissues: a proposed unified function in photoprotection. *New Phytologist, 155*(3), 349-361.

Suzuki, T., Ueoka, Y., & Sato, H. (2009). Estimating structure of multivariate systems with genetic algorithms for nonlinear prediction. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *80*(6).

Thenkabail, P., Smith, R., & De Pauw, E. (2000). Hyperspectral Vegetation Indices and Their Relationships with Agricultural Crop Characteristics. *Remote Sensing of Environment, 71*(2), 158-182.

Thenkabail, P., Smith, R., & De Pauw, E. (2002). Evaluation of Narrowband and Broadband Vegetation Indices for Determining Optimal Hyperspectral Wavebands for Agricultural Crop Characterization. *PHOTOGRAMMETRIC ENGINEERING & REMOTE SENSING, 68*(6), 607-621.

Thomas, J., & Gausman, H. (1977). Leaf Reflectance vs. Leaf Chlorophyll and Carotenoid Concentrations for Eight Crops1. *Agronomy Journal, 69*(5), 799.

USDA.. *What Are Crop Simulation Models?* Retrieved from http://www.ars.usda.gov/main/docs.htm?docid=2890

Ustin, S., Gitelson, A., Jacquemoud, S., Schaepman, M., Asner, G., Gamon, J., & Zarco-Tejada, P. (2009). Retrieval of foliar information about plant pigment systems from high resolution spectroscopy. *Remote Sensing of Environment, 113*(SUPPL. 1), 67-77.

Wang, L., Tian, Y., Yao, X., Zhu, Y., & Cao, W. (2014). Predicting grain yield and protein content in wheat by fusing multi-sensor and multi-temporal remote-sensing images. *Field Crops Research, 164*(1), 178-188.

Wang, Q., Shi, W., & Atkinson, P. (2014). Sub-pixel mapping of remote sensing images based on radial basis function interpolation. *ISPRS Journal of Photogrammetry and Remote Sensing, 92*, 1-15.

Wang, Y., Xu, X., Huang, L., Yang, G., Fan, L., Wei, P., Chen, G. (2019). An Improved CASA Model for Estimating Winter Wheat Yield from Remote Sensing Images. *Remote Sensing, 11*(9), 1088.

Weather Monitoring - Agri4castWiki.. Retrieved from https://marswiki.jrc.ec.europa.eu/agri4castwiki/index.php/Weather_Monitoring

Wu, Q., Jin, Y., Bao, Y., Hai, Q., Yan, R., Chen, B., Xin, X. (2015). Comparison of two inversion methods for leaf area index using HJ-1 satellite data in a temperate meadow steppe. *International Journal of Remote Sensing, 36*(19-20), 5192-5207.

Yamasaki, H., Sakihama, Y., & Lkehara, N. (1997). Flavonoid-Peroxidase Reaction as a Detoxification Mechanism of Plant Cells against H 2 0 2. *Plant Physiol, 11*(2), 1405-1406.

Yoder, B., & Waring, R. (1994). The normalized difference vegetation index of small Douglas-fir canopies with varying chlorophyll concentrations. *Remote Sensing of Environment, 49*(1), 81-91.

Yue, J., Feng, H., Yang, G., Li, Z., Yue, J., Feng, H., Li, Z. (2018). A Comparison of Regression Techniques for Estimation of Above-Ground Winter Wheat Biomass Using Near-Surface Spectroscopy. *Remote Sensing, 10*(2), 66.

Zhang, Y. (2011). Chapter 7 Forest Leaf Chlorophyll Study Using Hyperspectral Remote Sensing. In Y. Zhang, *Hyperspectral Remote Sensing of Vegetation* (pp. 263-264). CRC Press.

Zhou, X., Huang, W., Zhang, J., Kong, W., Casa, R., & Huang, Y. (2019). A novel combined spectral index for estimating the ratio of carotenoid to chlorophyll content to monitor crop physiological and phenological status. *International Journal of Applied Earth Observation and Geoinformation, 76*, 128-142. doi:https://doi.org/10.1016/j.jag.2018.10.012

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Özcan, Ayşenur
Nationality: Turkish (TC)
Date and Place of Birth: 24 March 1980, Kahramanmaraş
Marital Status: Single
Phone: +90 505 329 05 35
email: aysenur.ozcan@gmail.com

EDUCATION

Degree	Institution	Year of Graduation
MS	METU Chemical Engineering	2005
BS	Ankara University Chemical Engineering	2003
High School	Atatürk Anadolu High School, Ankara	1997

WORK EXPERIENCE

Year	Place	Enrollment
2018 June - present	UN World Food Programme	Information Management Officer
2010 January	Turkish Statistical Institute	Statistical Expert
2006 January	Turkish Statistical Institute	Deputy Expert

FOREIGN LANGUAGES

Advanced English

PUBLICATIONS

1. Ayşenur Türkmenoğlu, Development of Meat and Milk Production Estimation Methods and Comparison of the Results with Current Statistics, Expert Thesis (153 pages), TurkStat, 2010.

2. Ayşenur Özcan, Halil Kalıpçılar, Preparation of Zeolite A Tubes from Amorphous Aluminosilicate Extrudates, Industrial & Engineering Chemistry Research 45 (14):4977-4984, 2006.

3. Ayşenur Özcan, Investigating the Extrusion of Alumina Silicate Pastes for Synthesis of Monolith Zeolite A, Master's Thesis (222 pages), Middle East Technical University, Department of Chemical Engineering, 2005

Tennis, Scuba diving, Computer technologies, Reading