Laboratory of Geo-Information Science and Remote Sensing

Thesis Report GIRS-2015-20

Forecasting and classifying potato yields for precision agriculture based on time series analysis of multispectral satellite imagery

Tziolas Nikolaos

May 2015

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Forecasting and classifying potato yields for precision agriculture based on time series analysis of multispectral satellite imagery

Tziolas Nikolaos

Registration number 89 07 08 846 040

Supervisors:

Dr. ir. Lammert Kooistra
ir. Gerbert Roerink

A thesis submitted in partial fulfilment of the degree of Master of Science
at Wageningen University and Research Centre,
The Netherlands.

May, 2015
Wageningen, The Netherlands

Thesis code number: GRS-80436
Thesis Report: GIRS-2015-20
Wageningen University and Research Centre
Laboratory of Geo-Information Science and Remote Sensing

Acknowledgement

First of all, I would like to thank my supervisor Lammert Kooistra for guiding this research and leaving me the free to explore several ideas and provided me with feedback in such detail way. Further thanks go to Gerbert Roerink for valuable discussions about approaches to extract representative NDVI temporal profiles. This helped me to see the fitting process with more critical eyes. He also offered the pre-processed RS data.

This research would not have been possible without the detailed ground based data kindly provided by Mr. Jacob van den Borne.

Another gratitude goes to Andreas Matalis and Ioannis Moutsinas were kind enough to proofread the manuscript, before I submitted the latest version of this research.

I would also like to mention the open source communities for providing most all of my research tools: RStudio and GDAL, to name only some. Nowadays, freely accessible software tools enable anyone to solve complex problems more feasible and faster.

Furthermore I want to thank my parents for their unconstrained support during my studies. Finally, my parents and all my friends for supporting and encouraging me during the times when I lose my motivation and get distracted.

Abstract

Since, March 2012 the National Satellite Dataportal of the Netherlands provides Disaster Monitoring Constellation (DMC) images with a time resolution of 2 days and sufficient spatial resolution of 22m, making it an ideal data source for application in precision agriculture sector. For a farm spanning between Belgium and the Netherlands as a research area data from two growing seasons (2013 and 2014) were analyzed, in order to determine if potato yield potential could be estimated utilizing an in-season estimation of normalized difference vegetative index (NDVI) and meteorological data. This research provides important insight into availability of cloud free satellite images during critical periods of agricultural growing season emphasizing that is a key for agricultural monitoring and yield prediction. In this research, inclusion of information related to crop phenology showing significantly improved model performance. Several methods used to provide recommendations for the estimation of yield at the field scale. Linear regression models developed using parameters of NDVI time series profiles were evaluated as a stand-alone yield predictor. Additionally, multivariate regression models were developed introducing bio-climatic variables (solar radiation, temperature and precipitation) in conjunction with NDVI. Beyond the regression models, decision trees were used to analyze a qualitative relationship between yield NDVI and meteorological variables. In general the results were significant and promising. The resulting yield maps provide a unique opportunity to inform agricultural management decisions. Future satellite missions should permit estimation of potato yields using image resolutions that facilitate extraction of information in more frequent times. This analysis has also described cloud cover frequency throughout the agricultural growing season, providing insight into how yield forecasting approach could be impacted by cloud cover.

Table of Contents

1 Introduction

1.1 Context and Background

The world population exceeds the 7 billion and according to the current birth rates it appears to rise by approximately 1 billion over the next several decades and finally surpass 9 billion in 2050 (United Nations, 2013). At the same time the demand for agricultural products is expected to double, according to recent estimations (FAO, 2009).

The agricultural sector is facing one of the largest challenges, being necessary to meet the growing demand for food, and at the same time halting agriculture expansion and eliminating the environmental impacts (Godfray *et al.*, 2010). Since the end of the last century there studies have indicated that both intensification of the agricultural production systems and expansion of arable land contribute to anthropogenic effects on the Earth's biogeochemical cycles (Vitousek *et al.*, 1997). In particular, agriculture is estimated as the direct driver for around 80% of deforestation worldwide (Kissinger *et al.*, 2012) resulting in biodiversity loss and climate change. Likewise, the excessive application of fertilizers, pesticides and water for irrigation use affect negatively the environmental quality and ecosystem services (Matson *et al.*, 1997; Gregory *et al.*, 2002).

All these developments reinforce the importance to pursue an ecological intensification of agriculture to increase current yield levels of existing crops on the same number of hectares and simultaneously protect the environment and the natural resources for the next generations (Cassman, 1999). Hence, the question arises, can we ensure food security using more sustainable ways in agricultural production systems?

Radical changes in agriculture have been introduced over the past century. Mechanized agriculture, advancements and innovations in breeding resulted in increased crop production and productivity from the late 1960s and beyond. Nowadays, advances in global positioning systems (GPS), geographic information systems (GIS), remote sensing (RS) and a series of several sensors can be a new inflection point in agricultural sector (Blackmore, 1994).

All these technologies found in precision agriculture (PA) provide large amount of data that is currently recognized to be the next step of providing a timely and reliable picture of actual field conditions(Robert, 2002). The collection, processing, analysis and understanding of that time series data can enable monitoring agricultural production, estimating variations in crop productivity among fields by making a yield prediction, during and throughout the growing season. Hence, farmers can be assisted to establish proper management action plans about their operations that will drive in significant increases in crop production alone on existing areas by using resources more efficiently.

So, there is need to attempt good techniques for early crop prediction in order to minimize the yield gap by identifying the potential scope for raising average yields via optimization of spatially explicit irrigation, fertilization and application of pesticides(Pinter Jr *et al.*, 2003). However, the necessity of estimating agricultural production is broadened in a wider variety

of applications as it can have a direct influence in marketing and logistical issues and determine of pricing policies of food (Lobell *et al.*, 2003).

1.2 Problem definition

Agricultural vegetation develops from sowing to harvest as a function of meteorological driving variables (e.g., temperature, sunlight, and precipitation). Since the early days of earth observation satellites , RS data has a emerged as a feasible tool to the monitoring task by providing timely, synoptic and repetitive information about agricultural vegetation (Atzberger, 2013). Several previous studies illustrated examples of RS in agricultural sector and a general review is given by Atzberger (2013). For the purpose of this research a brief summary of yield estimation and assessment of crop phenological development is provided below.

Numerous studies (Yang and Anderson, 2000; Lyle *et al.*, 2013) in literature dealing with the estimation of yield heterogeneity via RS have been presented as an alternative to laborious and time consuming field measurements or to complex crop growth models. In summary, past predictions of yield, which incorporated a variety of approaches have ranged from simple statistical relationships (Bolton and Friedl, 2013; Lyle *et al.*, 2013); more advanced relationships, using multivariate linear regression models (Prasad *et al.*, 2006; Balaghi *et al.*, 2008), based on remotely sensed and meteorological data; as well as recent intelligent data mining techniques like artificial neural network, decision trees, and feature selection algorithms, which have the capability to involve hundreds or even thousands of variables (Panda *et al.*, 2010; Fernandes *et al.*, 2011; Gonzalez-Sanchez *et al.*, 2014; Johnson, 2014). The availability of RS time-series data enable for delineating spatial and temporal patterns of crop phenology on a per pixel basis. Taking advantage of this application, recent studies (Funk *et al.*, 2007; Bolton and Friedl, 2013; Wang *et al.*, 2014b) highlighted that the correlation of yield performance both differ between vegetation indices and varies through the crop cycle. Results from these studies have indicated that the perspective of forecasting yield is enhanced by using particular information related to crop phenology.

Although, the potential value of field level evaluation is commonly recognized by the researchers, the majority of the above mentioned studies present approaches for yield estimation at scales broader than individual fields. Two are the main reasons for this. On the one hand, the evaluations of approaches at scales broader than individual fields performed by comparing reported yields for counties or crop reporting districts with the average of RS yields over these domains (Doraiswamy *et al.*, 2005; Becker-Reshef *et al.*, 2010; Lobell *et al.*, 2010). On the other hand, RS estimation of crop yields at field level has been hindered due to the difficulty for access to sufficient data that reflect the production and productivity of actual farmer fields.

A second hindrance, up to now, was located in the limitation to obtain images at sufficient spatial resolution to delineate individual fields (Figure 1). Hence, those studies have employed low resolution sensors because of their availability and frequent revisit time. Landsat imagery allows yield predictions at a higher spatial resolution by distinguishing fields that are roughly 1 ha in size or greater and it has also extensively used in RS applications in several agricultural areas (Lyle *et al.*, 2013). However, with a 16 days temporal resolution the

problem of obtaining satellite imagery in cloudy locations, like the Netherlands, acts as an obstacle for precision farming applications (Zhang and Kovacs, 2012). Moving forward, the innovations in RS technology have positive effect on lower cost and increased number of fine spatial resolution sensors helping to partly mitigate the inherent tradeoff between spatial and temporal resolution. In the light of the above, high spatiotemporal data are becoming available and can be used to monitor crop growth in real-time detection. Thanks to these developments RS has shown great expectations for quantifying yield variations both within and between fields. Although many studies have employed RS in precision agriculture to analyze variations within individual fields (Kooistra *et al.*, 2012; Gevaert *et al.*, 2015) few have been addressed between yield variations across the landscape facing each field as a single unit (Lyle *et al.*, 2013).

However, RS data are still not available on a daily basis, hence some interpolation is recommended to estimate daily VI values. Some researchers proposed to resolve the compromising between spatial and temporal resolution by performing interpolation based on some statistical function such as a local linear interpolation (Mingwei *et al.*, 2008; Pan *et al.*, 2015) or fitting double logistic curves (Zhang *et al.*, 2003; Beck *et al.*, 2006). Extensive description of these methods is provided in the second chapter.

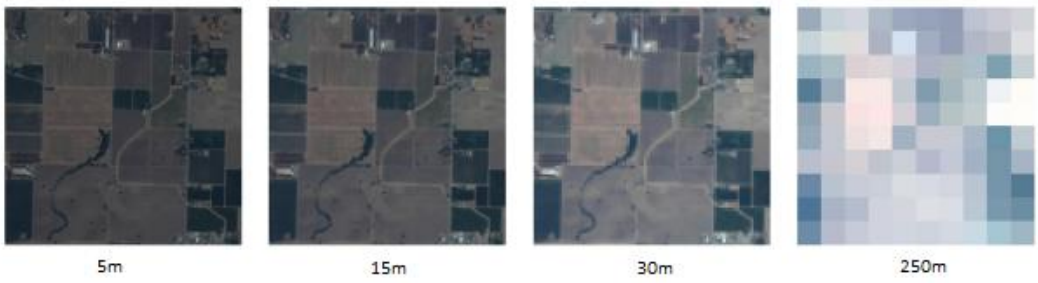


Figure 1: Examples of the effect of image resolution on the ability of remote sensing to monitor individual fields. The left image displays a 2.5 km×2.5 km section of a Rapid Eye 5m x 5m resolution image on agriculture area in Kings County, United States. Moving right, images resampled to represent the coarser resolutions of some common sensors: ASTER (15m, second), Landsat (30m, third) and MODIS (250m, last) are presented (Lobell, 2013).

Finally, most of the above mentioned studies for crop prediction give more emphasis on cereals and rice, as over half of the global population relies on them for their nutrition. However, studies over the past decade demonstrating a slowing of crop yield rates of grain both in farmers field's (Lin and Huybers, 2012) and on broader scales (Cassman *et al.*, 2003; Brisson *et al.*, 2010) raise the concerns for ensuring equitable food production. Possible shortages could be covered by other type of crops like tuber crops (potatoes) which are very easy to grow everywhere and don't have a lot of requirements. In addition, the Netherlands is a world leading country as far as it concerns potato production. The subsequent knowledge and experience in the field of potato growing, harvesting techniques, storage, transport and processing can also offer solutions to overcome food security issue. In general, potato has been also mentioned in few studies (Bala and Islam, 2009; Ramírez *et al.*, 2014) but the number of crop prediction studies focusing on potato is significant limited.

In general, there is clear evidence that the availability of time-series data, with a high spatial and temporal resolution to detect variations in crop productivity and production, is a critical

user requirement for the application of remote sensing in precision agriculture (Hatfield and Prueger, 2010). Crop yield estimation with good accuracy provides an insight on how to face field as a single unit across the landscape in order to improve management and maximize field productivity. For instance, this information on emerging problems like water stress or possible diseases, could be indicated during the growing season. All the above can assist the farmers to make appropriate crop management such as selecting the optimal rate and time of fertilization and irrigation, for optimizing the yield quantity and quality (Wu *et al.*, 2007).

1.3 Research objectives & questions

This research focuses on the development and evaluation of a method to estimate potato yields from remote sensing time-series at the field level, for 2013 and 2014 growing seasons. The aim is twofold, as research is done on both the entire growing season and in several time steps. On the basis of DMC's multispectral remote sensing data, the quantitative relationship between the NDVI and potato yield was analyzed and additionally the optimal period for predicting potato yield within the growing season was determined. Due to the changes of crop phenology among the years, an approach using information related to this phenomenon was investigated, aiming to enhance the performance of yield forecasting models. Having explored how far NDVI can be used as a stand-alone yield predictor, regression models were developed by incorporating meteorological factors. Additionally, besides the regression models, decision trees were used to analyze a qualitative relationship between yield and NDVI metrics, providing means to assess the relative performance of each technique in the context of explaining the yield variability of fields across the landscape.

The following research questions were leading this research in order to meet the objectives of this study :

RQ1 Can remotely sensed information related to crop phenology be used to improve the predictions of yield ?

RQ2 Does the number of satellite images influence the estimated yield accuracy ?

RQ3 How can the meteorological factors be incorporated in a VI-yield prediction model ?

RQ4 Which classification accuracy can decision trees provide a qualitative yield estimation ?

1.4 Thesis outline

This report is structured as follows: Chapter 1, context and background of the research were discussed and the problem definition was introduced. Chapter 2 contains a description of yield prediction methods, crop phenology of the investigated plant and several phenological metrics that can be derived by analyzing remotely sensed times series. Consequently, a brief description of method in order to obtain daily basis VIs time series and an introduction in data mining techniques are presented. Chapter 3 describes the study area and takes a closer look in all data sets used. Also it deals with the methods used to analyze the data to offer answers for the research questions. Chapter 4 summarizes the results from each proposed method, while the results are discussed in a broader context taking into account other limitations and results from other studies are presented in Chapter 5. Finally, Chapter 6 contains an overall conclusion and recommendations for further studies.

2 Literature review

In this chapter a closer look at the theoretical framework is presented, providing a synoptic literature review on the main topics of this study.

2.1 Yield prediction

In general, several methods have been developed and implemented for estimating agricultural production on regional, national and global scale. Each method displays advantages and disadvantages in forecasting the crop yield precisely. Conventional methods based on crop data collection from field visits (Cassman *et al.*, 2003) or surveys using questionnaires among farmers to assess future yields are costly, time consuming, prone to errors and leading to inaccurate crop estimations (Reynolds *et al.*, 2000). Furthermore, many yield forecasting approaches involve the use of mechanistic crop growth models (Grassini *et al.*, 2011; Laborte *et al.*, 2012). However, such models require a large volume of specific inputs which are available for only a limited number of farmers. Subsequently, they need to be calibrated and validated always for robust simulations from different crop types and environmental parameters. Their complexity and high input data requirements have been considered as the main limitations for application of these models at a field level. Remotely sensed data also applied into mechanistic crop growth models offering a chance to enhance yield prediction (Doraiswamy *et al.*, 2005; Chahbi *et al.*, 2014). Due to the fact that RS can provide with timely information at a range of scales, accurate yield prediction is feasible using only RS data, as a consequence crop yield reference data are necessary only in the validation phase. The overview of Gallego *et al.* (2010) presented several approaches for estimating crop yields with RS. The majority of these studies based on the relationship of VIs or other biophysical parameters, such as LAI to the yield with a focus on wheat, corn, rice and soybeans. One obvious reason is that these crops are found in quantity around the world, however an additional cause could be that there have been found better success in yield prediction against crops, like potato or sugar beet, which are grown below ground.

However, in literature some examples for potato are demonstrated. For instance, Bala and Islam (2009) developed linear regression models in Bangladesh between potato yield and NDVI, LAI and the fraction of Photosynthetically Active Radiation (fPAR) using coarse spatial resolution data imagery from Moderate Resolution Imaging Spectroradiometer (MODIS). Performing the analysis for two growing seasons (2006-2007), they found a strong correlation between NDVI, LAI and fPAR during the growing season with an average error of estimation was about 15% . Neale and Sivarajan (2011) compared potato yield for a total of 15 fields to the Soil Adjusted Vegetation Indices (SAVI), at three stages in the growing season. Also they integrated the entire area under the VI curve (AUC) and then finally regressed it with yield, in order to estimate the yield prediction for the whole growing season. Their results showed that the three integrated SAVI yield model developed using airborne and multispectral satellite (Landsat TM5) images resulted in good prediction for most of the fields, with a root mean square error (RMSE) of 0.29 and 0.24 kg/sq.m, respectively. The analyses in above mentioned studies were performed using VIs data across a wide range of spatial resolutions and regression models. The main drawback of these regressions is that cannot be utilized in a general manner because are only applicable to the

spatial and temporal extent of the specific study regions. This is caused by the variability of the relations between crop types, growth conditions and meteorological factors (Rudorff and Batista, 1990).

To date, studies at a scale that is relevant for management at field level are limited due to the availability of ground based yield data at a sufficient spatial resolution for discriminating the individual fields. In general, yield research at field level scale seems to have quite good accuracies with root mean square errors RMSE smaller than 10% comparing with studies that make use of low spatial resolution data (Lobell *et al.*, 2005). However, nowadays, more and more studies are performed at field scale level. For instance, Morel *et al.* (2014) compared sugarcane to integrated NDVI that was obtained from SPOT-4 and SPOT-5 time series images. Lyle *et al.* (2013), explored the variation, across three farms, in wheat yield with NDVI of individual images over 5 growing seasons. Although there are few examples for potato forecasting via RS , no studies performed at fields level scale in order to meet the scope of PA.

Finally, significant results of previous studies (Funk and Budde, 2009; Bolton and Friedl, 2013; Wang *et al.*, 2014b) indicated that crop phenology-tuned VIs present an enhanced correlation with yield. The authors attempt to relate the crop yield with several vegetation indices by developing empirical models specifying the start of photosynthetically growing season by using the onset of rainy season or the date of onset of VIs increase at each pixel. In summary, the results by Bolton and Friedl (2013) showed that yield forecast models based on phenologically adjusted VIs present better performance of approximately 10%, versus approach using a fixed calendar date to estimate remote sensing-based yield prediction models.

It is well known that yield is affected by several parameters. Previous studies have highlighted the strong dependency of agriculture vegetation on weather. For example, Prasad *et al.* (2006) developed crop yield prediction models based on corn and soybean crops, implementing NDVI and meteorological factors by using a piecewise linear regression method. Balaghi *et al.* (2008) also make use of rainfall and temperature for wheat yield forecasting by means of ordinary least squares regressions. Both studies indicated that weather parameters could be beneficial for yield prediction models. However, the models developed in the above studies were implemented in broader scales than individual fields. Both showed promising results indicating that incorporating weather parameters should have positive effect in yield estimation area.

The increased availability of survey datasets on weather, management practices, and time series of vegetations indices and the derivatives of them generate an issue for handling and smart using of big data for agriculture applications, such as yield estimation. Besides, the methods described in the previous paragraphs numerous methods have been reported and found suitable for crop yield forecasting such as, neural networks (Uno *et al.*, 2005) and decision tree type models (Fernandes *et al.*, 2011; Johnson, 2014).

2.2 Crop phenology

2.2.1 General features

Phenology is defined as the study of observing the changes in life stages of biological events in relation to the temporal occurrence (Menzel, 2003). According to White *et al.* (1997) climate variations cause changes in crop phenology as they are considered significant indicators on the impact of natural ecosystems. This is also confirmed by recently studies (Tao *et al.*, 2006; Heumann *et al.*, 2007) where, plant phenologies affected by factors such as soil and air temperature, photoperiod and precipitation vary depending upon the species, year and location. Agronomists consider any agricultural crop as a system that interact with all the factors of its physical environment.

From this point of view, variations on timing of growth stages of crop activity may be important for the agriculture science as can be an indicator of the impact of inter and intra-seasonal variations of the above mentioned factors. Previous studies demonstrated that this information assist to evaluate properly the crop condition and facilitate in supporting decisions and management practices (Pan *et al.*, 2015).

2.2.2 Potato phenology

Growth of a potato plant occurs in several stages: sprout emergence and development, plant establishment, tuber initiation, tuber filling, and tuber maturity (Figure 2). Timing of these growth stages can vary depending on environmental factors, such as elevation and temperature, soil type, adequate water levels in soil, cultivar selected, and geographic location (Coelho and Dale, 1980).

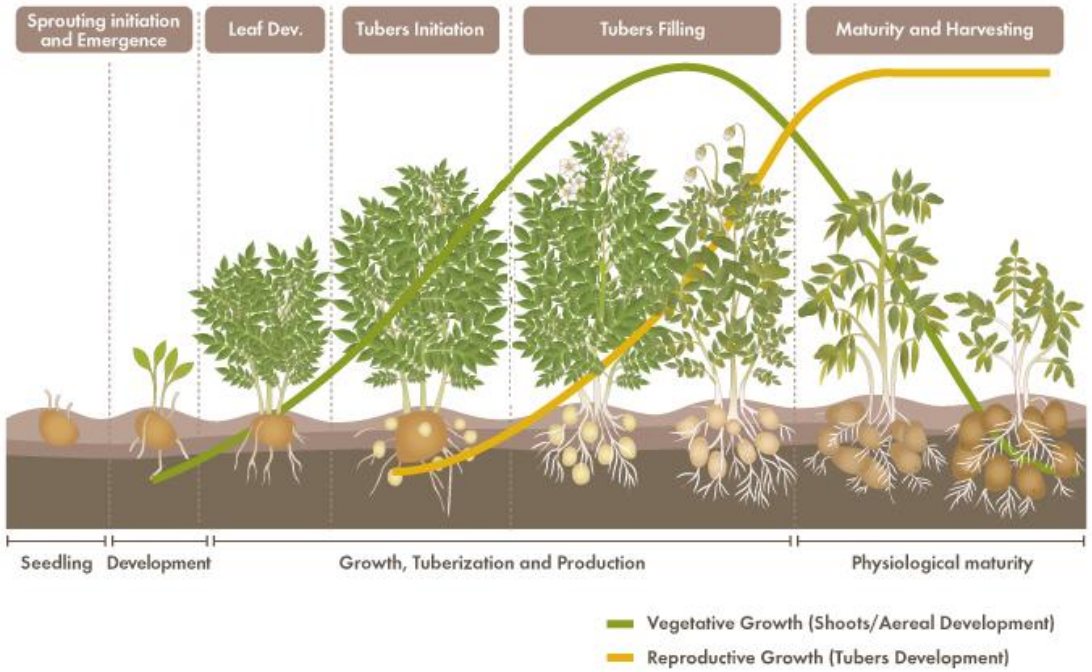


Figure 2: Growth stages of potato. The green line represents the vegetative growth coinciding with the aerial development, while the yellow line illustrates the reproductive growth coinciding with the development of the tubers, source: (Sqm.com, 2015)

The period of establishment followed by the vegetative phase in which the tuber initiation starts and then a yield formation during the tubers filling phase is taken place. Subsequently, at the start of the senescence characteristics and the end of the period of the potato cluster, the majority of the products of photosynthesis are transferred into the tubers with the duration of this period displaying an positively relation to tuber yield.

The effect of temperature on crop phenology was examined during the potato growing season by agronomists. However, field experiments (Sarquís *et al.*, 1996) in potato indicated that the magnitude of final yield was not directly related to temperature but was determined by an intricate interaction between all the meteorological factors. In general, the tuber initiation and the growing phase require temperatures between 18° to 20°C. Temperatures over 25° C increase the respiration rate therefore, the net assimilation rate is decreased, with dramatically effects in the photosynthetic process. Photoperiod influences the tuber filling rate, and it should be taking into account that this rate is significant high in the study area of this research (section 3.1) as it has few light hours during the day. Previous studies, showed that increases in solar irradiance led to significant potato yield improvements (Stutte *et al.*, 1996). Finally, it is important to avoid variation in the soil humidity levels during the crop growing period, particularly at the phase of tuber initiation, when good humidity level should be constant. This will influence the increment of stems, foliage density, weight and tubers number. Likewise, availability of soil moisture during the tuberization phase, eliminating the risk of diseases.

2.2.3 Monitoring crop phenological development using remote sensing

Remote sensing, due to the very good correspondence between signal and measures of vegetation and the repeated temporal sampling of satellite observations used for monitoring vegetation dynamics. Previously vegetation index time series analysis, in scope of crop phenology study, performed using data were obtained from sensors like AVHRR (Advanced Very High-Resolution Radiometer) (Jakubauskas *et al.*, 2002; White *et al.*, 2009) and MODIS (Zhang et al., 2003; Sakamoto et al., 2005) as their high temporal resolution benefits the monitoring of ecological process that occurred in crop plantation land (Wu *et al.*, 2010). However, their coarse spatial resolution is the intrinsic drawback of these sensors for applications in the sector of PA, such as is less capable of detecting small scale disturbances like those from agriculture activities scheduling (Estrella *et al.*, 2009; Begue *et al.*, 2014).

Studies (Lobell *et al.*, 2005; Lyle *et al.*, 2013) dealing with agriculture application, overcame this limitation by utilizing data from Landsat in order to achieve a finer spatial resolution. However, cloudiness conditions during different portions of the agricultural growing season display difficulties in the construction of time series. Specifically, yield forecasting applications require high frequency data during agricultural growing season (Becker-Reshef *et al.*, 2010) such as high spatial resolution in order to delineate small fields.

Few studies have been reported to resolve the issues of required high spatiotemporal resolution in agricultural applications. Some researchers introduced data fusion technologies in order to simulate high spatiotemporal resolution images (Hankui *et al.*, 2014; Gevaert and García-Haro, 2015), while others proposed interpolation techniques (Zhang *et al.*, 2003; Beck

et al., 2006). The first approach is out of the scope of this study, hence only the second one will be discussed below.

Numerous methods for interpolating and reconstructing VIs time-series have been proposed and several reviews on this topic are available (White *et al.*, 2009; Cong *et al.*, 2012). Only a brief summary of these approaches with their advantages and limitations is given here. They range from simple to more sophisticated techniques, mainly including: Harmonic Analysis of Time Series (HANTS) (Roerink *et al.*, 2000), curve fitting (e.g. double logistic and asymmetric Gaussian function) (Jonsson and Eklundh, 2002; Zhang *et al.*, 2003; Beck *et al.*, 2006; Julien and Sobrino, 2010), signal smoothing (e.g., Fourier analysis) (Mingwei *et al.*, 2008; Sakamoto *et al.*, 2010) and signal smoothing integrated with linear interpolation (Pan *et al.*, 2015).

Several of the aforementioned methods require representation of a time-series that is continuous and evenly interval (Jönsson and Eklundh, 2004; Zhu *et al.*, 2012). The eight-day composites from MODIS meet this requirement, hence several studies performed utilizing this kind of data. However, satellites (e.g., DMC or Chinese HJ-1A and HJ-1B satellites) with finer resolution than MODIS provide data having irregular availability over the year, especially in areas with frequent cloud cover, such as the Netherlands. Hence, when utilizing time-series data with irregular equidistantly spacing, the major problem is reconstructing the completeness of time-series dataset. Pan *et al.* (2015) claimed that methods to manipulate unevenly spaced time-series data (Baisch and Bokelmann, 1999) may not work well in non equidistantly spaced time-series data derived from Chinese HJ-1A and HJ-1B satellites which have similar characteristics with the DMC satellite which covers the Dutch territory.

In general the selection of method for creating an inter-annual curve fit, is not always straightforward (Hird and McDermid, 2009). It is recommended prior the processing to construct a complete time-series dataset, the researcher consider carefully the objectives of the study and the availability of satellite data, in order to select the most appropriate method fitting study's demands. Moreover, the researchers must be careful about the maintenance of original characteristics of time-series profile in order to avoid errors later in yield estimation (Lobell *et al.*, 2010).

Modeling the whole time series giving the advantage of obtaining a maximum amount of information in the VI data (Jönsson and Eklundh, 2004). Few studies taking this advantage determine the timing of vegetation greenup and senescence date and simultaneously extract several phenological metrics. Therefore, phenology monitoring and extraction of time series parameters are high significant over agricultural areas as they represent the impact of inter and intra-seasonal variations of climate.

In general useful information can be extracted and used in further analysis are spring of season (SOS), length of season (LOS), maximum NDVI value (maxNDVI), and cumulated NDVI over the season (cum-NDVI). These are illustrated in Fig. 3 and are the same as in Brown *et al.* (2010). In several studies SOS mentioned as greenup date, hence this term will be used hereafter in this research.

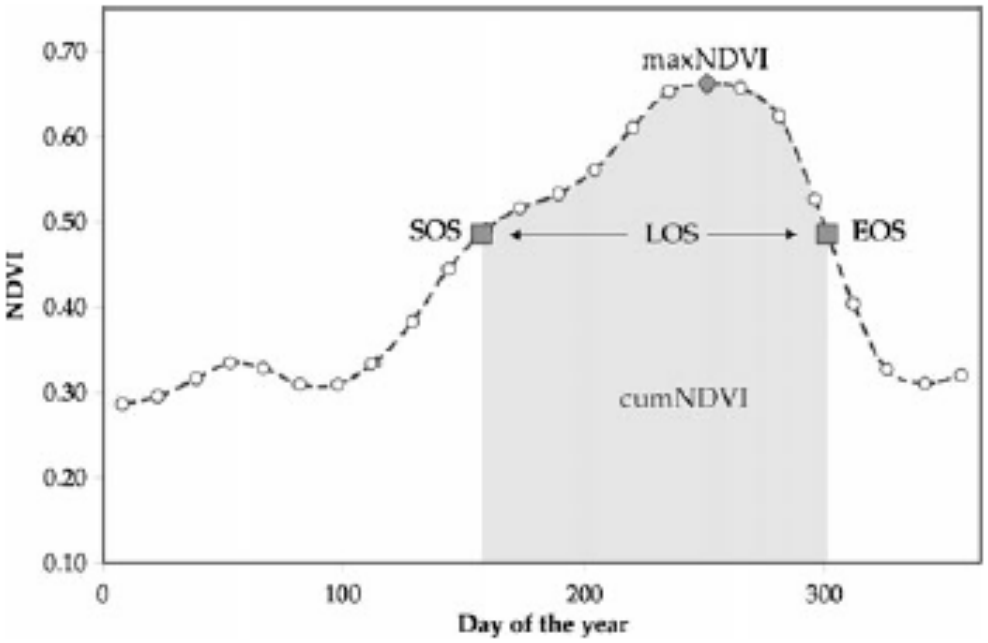


Figure 3: Phenological metrics, source: (Brown *et al.*, 2010)

2.3 Data mining approaches for data classification

Data mining provides the opportunity for knowledge discovery among large sets of data, in order to extract useful and important information. Data mining by artificial neural network, decision trees, and feature selection algorithms have been applied in agriculture, however it is a relatively new approach for forecasting the crop yield. Some of the techniques, such as the k-means (Shekoofa *et al.*, 2014), the k support vector machines (Löw *et al.*, 2013), decision tree (Fernandes *et al.*, 2011) have been applied in the field of agriculture for crop type and crop yield classifications.

A prerequisite of these techniques is the extraction of a specific dimensional features vector in order to reflect different aspects of the features with a class label attached. In a few studies (Deelers et.al, 2007, Sarangi et.al, 2013), the simple K means algorithm was used for this purpose. K-means clustering is a partitioning based clustering technique of grouping items into a specified number of cluster groups. There are two approaches to cluster center initialization either to select the initial values randomly (Sarangi et.al 2013), or to choose the first k samples of the data points (Deelers et.al, 2007).

After attaching a specific class label application of a classifier method is followed for prediction of the class label of the features input. The decision tree algorithm (Dancey *et al.*, 2007) has the capability to predict the value of a discrete dependent variable with a finite set from the values of a set of independent variables. The decision tree algorithm displays several advantages such as short computational time, ability to attain nonlinear mapping for feature selection. The decision tree approach is most useful in classification problem since it presents hierarchical ranking of important features and provides a clear image of effective factors(Lobell *et al.*, 2005; Fernandes *et al.*, 2011). For the purpose of this study J48 classifier (Quinlan, 1996) is selected hence, it is the only method that will be explained in this section. J48 classifier is a simple C4.5 decision tree for classification, the rationale is the creation of a

binary tree. With this technique, the constructed tree models the entire classification process. Once the tree is built, it is applied to each label in the entire database and results in classification for that label.

Finally, evaluation of the prediction performance is calculated and also its validity using cross validation technique or an independent evaluation dataset. Many studies dealing with classification used for this purpose a confusion matrix which illustrates the accuracy of the solution.

3 Study Area and Data Description

3.1 Study area

The study area spans between the northern Belgium and south part of the Netherlands incorporating, the Van den Borne Aardappelen farm (Figure 4). The study area was selected due to the farmer was able to provide a large amount of agronomic data for each field (e.g. yield, plant and harvest dates, agricultural practices) collected over two growing seasons (2013, 2014).

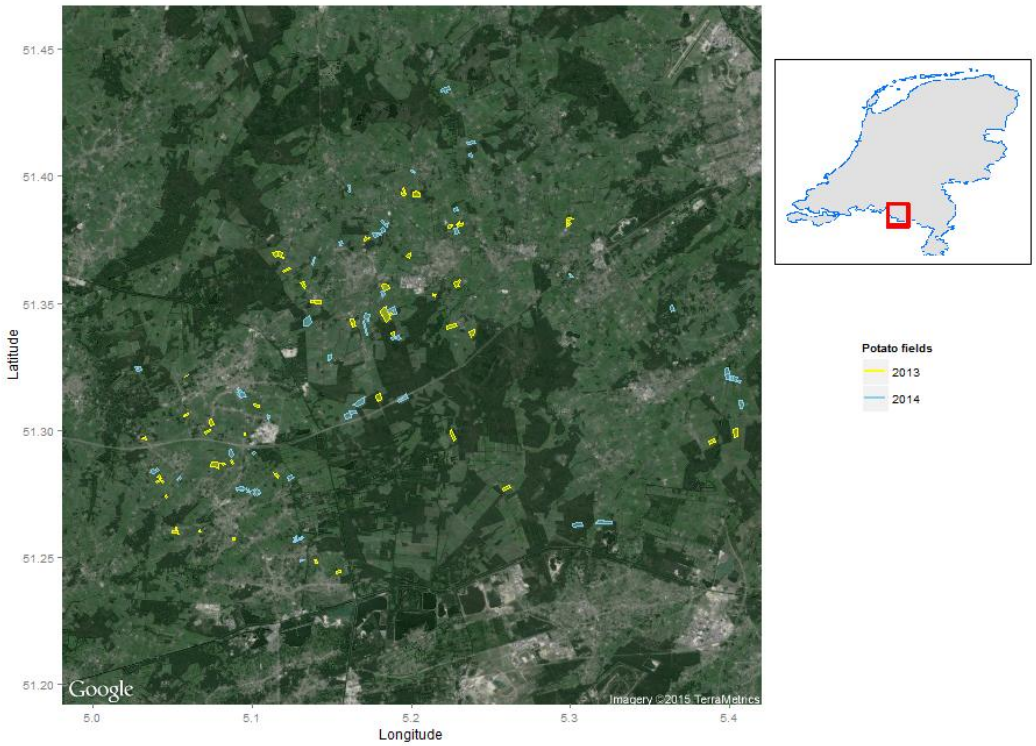


Figure 4: Location of the study area with potato fields on the border of Netherlands and Belgium.

The ground data set was composed of 122 fields spread over the two seasons with a mean area of the studied fields was 4.1 ha and four different potato cultivars were used: Fontane, Miranda, Ludmilla and Lady Anna. The fields used in this study were rain fed and irrigated as illustrated in Table 1

Table 1: Number and type of test fields for the two agricultural seasons

Agricultural year	Number and type of fields
2013	Non-irrigated:31
	Irrigated: 25
2014	Non-irrigated: 31
	Irrigated: 35

The farmer in order to calculate the production of each field makes use of a net to remove debris such as stones pieces of vegetation etc, then the remainder is weighted and is recorded as the actual yield of field. The weighted yields were similar in both seasons. Among all the fields, the weighted yields for 2013 ranged between 40.5 $t \cdot ha^{-1}$ and 84.23 $t \cdot ha^{-1}$ and the mean yield was 60.77 $t \cdot ha^{-1}$, while for 2014 the yields recorded between 40.9 $t \cdot ha^{-1}$ and 83.39 $t \cdot ha^{-1}$ with a mean value of 58.94 $t \cdot ha^{-1}$ (Figure 5). For rain fed and irrigated

fields there were no significant differences in yields, with them displayed an average value of 58.84 t·ha^{-1} and 60.82 t·ha^{-1}, respectively. A shapiro - test resulted a p-value equal to 0.2686 and 0.2005 for 2013 and 2014, respectively, as a result the yield for both years data are normally distributed.

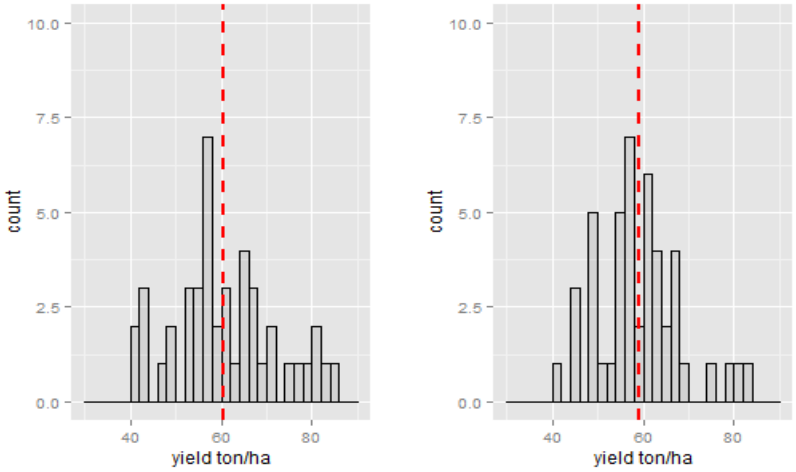


Figure 5: Yield statistics for 2013 (left) and 2014 (right)

3.2 Remote sensing data

Since, March 2012 the National Satellite Dataportal (NSD) of the Netherlands is active and provides Disaster Monitoring Constellation (DMC) images free available. DMC satellite has such capability to ensure sufficient repeat cycle, meeting the needs for higher temporal resolution during critical periods of crops growth cycle. Additionally, the spatial resolution of 22m, makes it appropriate for detailed analysis within fields for actual precision applications.

A total of 30 and 35 DMC images covering the study area were acquired for 2013 and 2014, respectively. Pre-processing procedures have already performed in satellite images by Alterra. Pre-processing procedures have included the conversion of raw data to reflectance, the removal of cloud effects and orthorectification. The NDVI was also computed for each available site image. Further details regarding the processing of satellite imagery can be found in www.groenmonitor.nl, in which time series of multispectral satellites vegetation imagery are provided since 2012.

The spatial variability in an agricultural field is inevitable in most cases due to various factors causing the variability. Some of the main causes of variability in crop growth are due to natural soil variability or impacts of erosion, land and crop management practices, and relief of the land. The other factors affecting the crop growth include fertilizer deficiencies causing soil nutrient variability, variability due to pest/disease attacks and water application non-uniformity during the crop growing season. An example of the variation of the NDVI values within a specific field is illustrated in Figure 6.

Figure 6: Variation of NDVI values within a specific field in three different DMC images

3.3 Meteorological data

The most important climatic variables for the vegetation are the daily maximum and minimum temperature, precipitation and global radiation Figure 7. All the meteorological data used in this study were obtained from the closest KNMI meteorological station in Eindhoven (KNMI 2014).

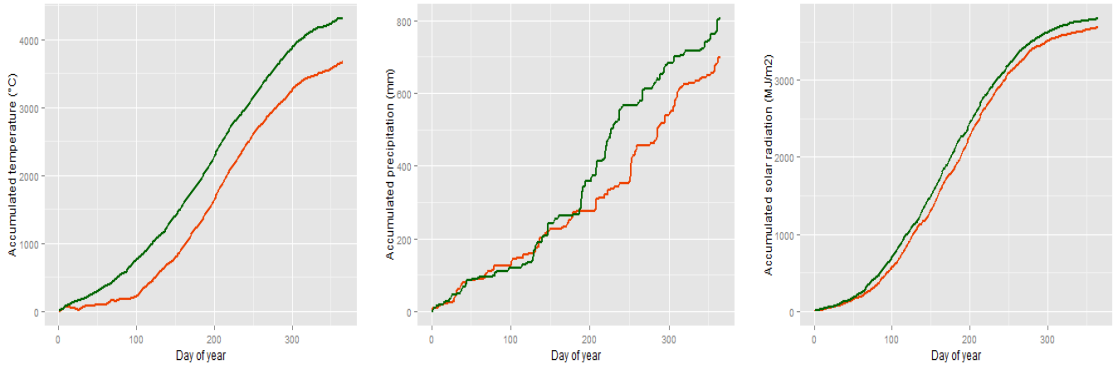


Figure 7: Accumulative temperature, precipitation and global radiation as recorded from KNMI meteorological station in Eindhoven The orange and green line correspond to 2013 and 2014, respectively

The meteorological data of the last 30 years (1984-2014) indicate that the mean averages temperatures display the minimum value in January (3.1 ^{o}C) while the maximum in July (18.2 ^{o}C). In addition, mean precipitation display a minimum value in April (43.5 mm) while the maximum in July (79.7 mm). In general, 2013 was cooler than 2014 however, both years were close to the average. The total annual rainfall of the two studied years does not explain the actual conditions observed during the respective growing seasons. The total annual rainfall amounted to 699 mm in 2013 and 806 mm in 2014. In 2014 most of the rain falling occurred between July and August whereas in 2013, it happened in August during the late growing season of potato crop. In 2013, rainfall occurring between planting and harvest (112-280 DOY) represented 45% (316 mm) of the annual total. By contrast, rainfall from 98 DOY to 281 DOY in 2014 represented only 63% (509 mm) of the annual total. In both years mean temperature rose gradually from the beginning of the growing season reaching a peak during July and then falling again towards the end. In details, in 2013 average temperature was 2°C below of 2014 during the period between respective planting and harvest dates. Finally, the cumulative global radiation in active growing period was 2687 and 2811 MJ/m^2 for 2013 and 2014, respectively.

3.4 Data pre-processing

Prior to the analysis, for each field an average NDVI time series profile was extracted by using a script initially developed in R Programming Language. For this reason, the data of field boundaries were utilized in the form of shape files, as provided from the farmer and the satellite data provided by WUR. A negative buffer distance of 20m was used to reduce the field boundaries in order to remove the tractor driving paths in the perimeter of the fields and also to eliminate the effect of mixed vegetation reflectance with neighboring fields. An average NDVI profile and standard deviation for all fields for both growing seasons are presented in Figure 8.

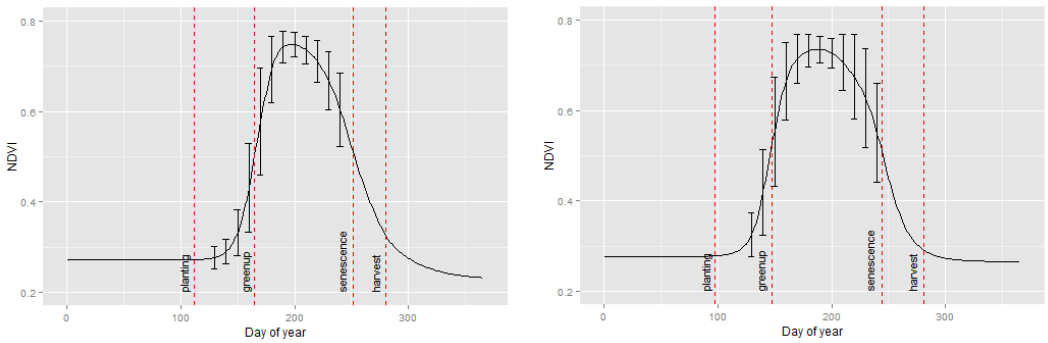


Figure 8: Average NDVI profiles and standard deviation for 2013 (left) and 2014 (right) cropping seasons and definition of estimated greenup and senescence boundaries and planting and harvesting dates as provided by the farmer

Another critical point is that several fields were covered by green vegetation before the cultivation of the potato. Therefore, in their temporal profiles high NDVI values observed ranging from 0.4-0.5 before the establishment and after the harvesting of potato crop.

This research only focused in the general characteristics of the main growth period of potato, as these explained by the unimodal shape of each field's NDVI curve. As a consequence, the aforementioned NDVI values, prior the establishment and after the harvesting of potato crop, were replaced with NDVI value of bare soil (Figure 9). The establishment date for each growing season defined as the average date of the planting dates as they provided by the farmer.

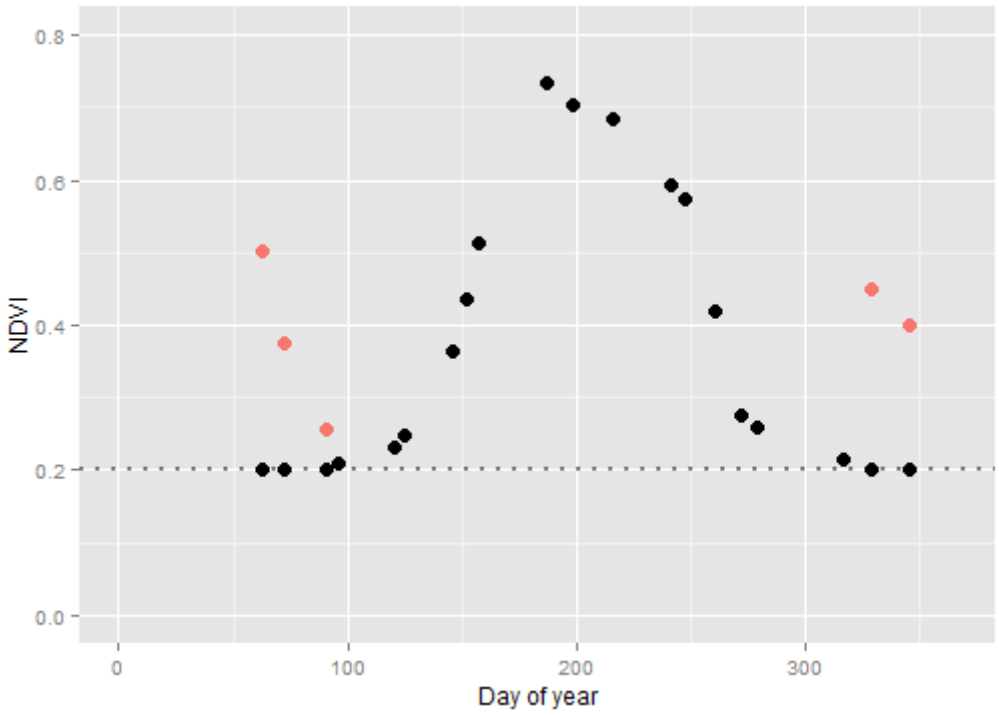


Figure 9: The dotted horizontal line corresponds with the bare soil NDVI, the solid pink represents high values before the establishment or after the harvesting of potato crop while the black solid symbols represent the NDVI values were adjusted

4 Methodology

This chapter describes the methods used in this research. Section 4.1 contains the methods used for the reconstruction of NDVI time series. In Section 4.2 the derivation of the phenological metrics is explained. These steps were necessary to perform later analysis as well as judge the reliability of results. Finally, the core analyses are explained in Section 4.3 to 4.5, following the logical order of the research questions in Chapter 1. Finally, Section 4.6 contain a short description about the yield maps. In general all methods presented here were developed around the average NDVI time series profiles which were obtained for each of the studied potato fields as described in Section 3.4.

4.1 Reconstructing of NDVI time series

As mentioned already in Section 2.1 there are several requirements and constraints to construct a complete time series dataset from remote sensing data. The use of NDVI time series of DMC data in this research need further processing to obtain an annual NDVI temporal profile for each field. Hence, two techniques from the literature were selected that had indicated successful application for describing vegetation dynamics: (i) a methodology including signal smoothing, and daily NDVI time-series interpolation (Pan *et al.*, 2015) and (ii) a method using a double logistic function (Beck *et al.*, 2006). From now on, these fitting approaches will be referred as SavGol and DL, respectively.

4.1.1 Savitzky Golay filtering and linear interpolation

The Savitzky and Golay (1964) smoothing algorithm has been used in several studies for data filtering and reconstruction of VI time series (White *et al.*, 2009; Chen *et al.*, 2011). A mathematical description of this process can be summarized as:

$$VI(t) = \frac{\sum_{n=-nL}^{nR} c_n VI_{i+n}}{n} \quad \text{(Eq.1)}$$

where, VI(t) is the fitted VI value, C_n is the filtering coefficient for the VI_i point, n is equal to the width of moving window to perform filtering and nL and nR correspond to the left and the right edge of the signal component, respectively.

For each point VIi, I implemented a least-square fit using a quadratic polynomial to all points within the moving window, and then I set VI(t) to be the value of that polynomial at position t. The idea of this polynomial is to preserve high moments within the data (Press, 1992). Subsequently, in order to complete the integrity of VI time series I used a linear interpolation method. Hence, a daily basis VI time series is obtained and be processed in the further steps.

4.1.2 Double logistic function

Among the models that have been developed and make use of different temporal template shapes, Beck et. al (2006) introduced a double logistic model. Their approach utilized in order to in this research to describe the unimodal crop growth period as follows:

$$VI_{(t)} = VI_0 + \Delta VI * \left[\frac{1}{1 + \exp\left(m_s(t - s)\right)} + \frac{1}{1 + \exp\left(m_a(t - a)\right)} - 1\right] \quad \text{(Eq.2)}$$

The model effectively has 6 parameters, two of which relate to the variation in the vegetation index (VI_0, the minimum value of the VI, and ΔVI, the difference between maximum and minimum VI). The other parameters relate to the shape of the ascending logistic function that models greening up senescence (m_s, s, m_a and a). The parameters s and a are related to the beginning and end of the photosynthetically active period while, m_s and m_a to the location of the inflexion point in the logistic curves. These starting parameters were specified, as a=90, s=320 and m_s=m_a=0.1 as introduced by Beck *et al.* (2006). In order to estimate the parameters of the logistic model and their prediction intervals of each field's NDVI time profile a nonlinear least-squares estimation was used (Bates and Watts, 1988) and executed in R (nls function) (R Development Core Team, 2011).

4.1.3 Evaluation of the methods

Outside the growing season, both methods should fit the observed NDVI values well, as it also include the bare soil NDVI as a parameter. Hence, the fit of the NDVI profiles was only quantified, during the growth cycle. The planting and harvesting DOYs, as provided by the farmer, were used to delimit this period for each field. When these information were not available, the median DOY value of the respective agronomic observation (planting or harvesting) of the year, was used as a replacement. The RMSE was calculated among the actual NDVI values from DMC images, O_i, and the estimated NDVI values after performing the interpolation approaches, P_i, as described in section 4.1.1 and 4.1.2, then it was used as a measure of model performance (Eq.3).

$$RMSE = \sqrt{\frac{1}{n} * \sum_{i=1}^{n} (P_i - O_i)^2} \quad \text{(Eq.3)}$$

4.2 Deriving NDVI time series metrics

The real curves of NDVI time-series are not regularly perfect, hence fitting precision have a directly influence in the acquisition of the phenological metrics. Comparing the RMSE between the SavGol and DL function the method that displayed the smallest RMSE was selected for the further steps of analysis. For this reason, in this section the daily NDVI modeled values after applying the DL method are utilized.

One of the advantages of using the double logistic function for describing an unimodal growing season is that the estimated parameters can be directly interpreted in terms of vegetation phenology (Fischer, 1994). In particular, the fitting parameters s and a (Section 4.1.2) provide the two inflection points of the fitted curve corresponding to the onset of NDVI increase (greenup) and NDVI decrease (senescence), respectively. In this research, these two dates of onset were used in order to demonstrate the start and the end of the actual potato growth process for each field. Subsequently, crop phenology parameters were extracted for the potato growth cycles for each field: greenup date, senescence date, maximum NDVI date, and length of growth season (summarized in Figure 10), all of which represent critical growth stages of potato. Finally, the integrated NDVI values were computed by integrals between greenup and senescence date of each field.

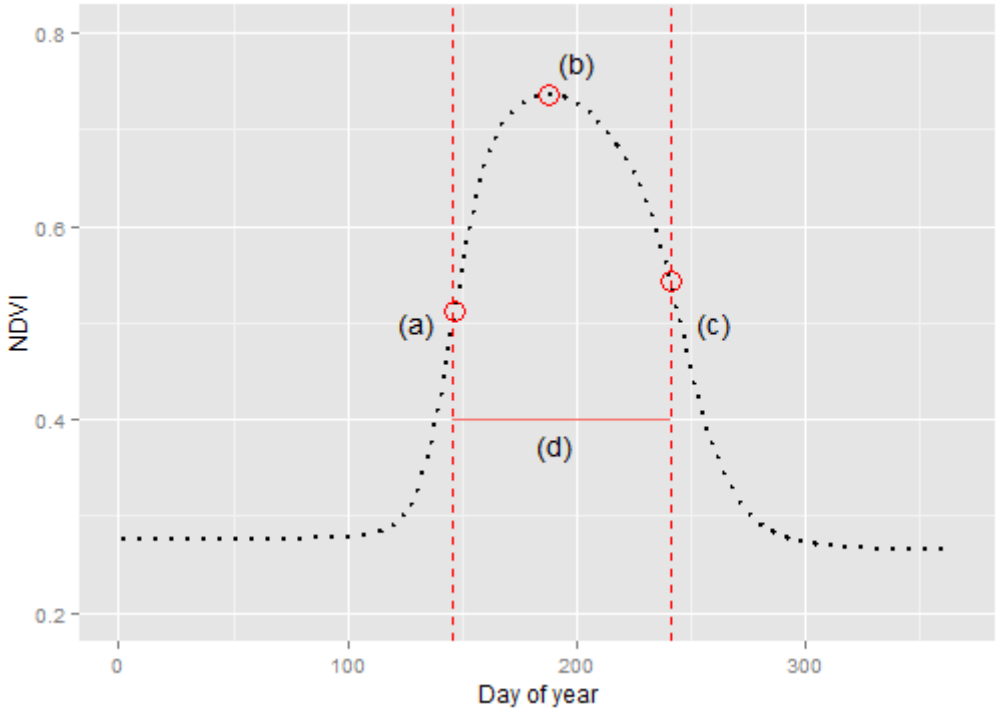


Figure 10: Phenology parameters of potato field in modeled NDVI time-series The points marked with red circle mean:(a) greenup date of (b) maximum NDVI (c) senescence date (d) length of growth duration (from greenup to senescence)

For validation, I based on the assumption that estimated greenup and senescence dates should be related with the relative planting and foliage killing, respectively. Therefore, I assessed the relationship among the phenological and agronomic dates by calculating the regression of determination (R^2).

4.3 Statistical Analyses

4.3.1 Estimate yield prediction models

The method that is proposed here, in order to forecast yield, is based on the findings of previous studies which indicated the importance of using "phenologically adjusted" spectral indices in yield forecast models (Bolton and Friedl, 2013; Mulianga *et al.*, 2013; Wang *et al.*, 2014b). Specifically, linear regressions models were compiled between potato yield measured on each field with NDVI derived information related to timing of crop phenology. The rationale behind this was that variability in the start of photosynthetical growing season at each field should have an instantaneous impact on the final yield prediction.

The strength of these relationships was tested at different times within the season as well for the entire potato growth season. For analyzing and forecasting yields within season, time series of averaged NDVI values were extracted for each 5 day intervals starting from the greenup date, as it estimated for each field in Section 4.2, and stopping 150 days after the greenup date. The result of this preprocessing step was a time series consisting of 31 NDVI values, with the starting point depends on the time of greenup of each field (Figure 11). Then the annual coefficients of determination between the NDVI and yields were computed in order to calculate a multi-year average coefficients of determination. The higher value of

R^2 indicates that the NDVI on the given date is a good predictor of final yield. Finally, linear regression models based on all the years were developed for the selected time step.

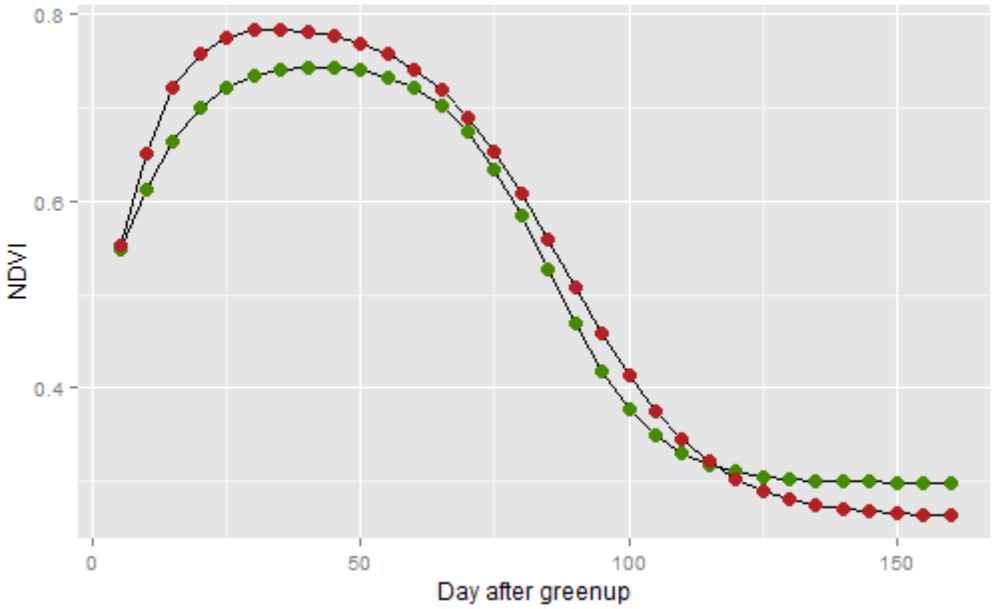


Figure 11: A schematic representation demonstrating the average NDVI values which were extracted at five day intervals, indicated by the solid symbols, starting from the greenup date until the 150 following days for two adjacent fields from two different growing seasons.

In addition, NDVI integrated over growing period has been found to be a relevant measure to study crop production (Mulianga *et al.*, 2013; Morel *et al.*, 2014). Therefore, the relation between measured yields and time integrated NDVI has been calculated also in this study. The integrated NDVI of the crop growing period, as estimated in Section 4.2 for each field, was adopted as well.

Predicted yields from the yield-NDVI models were then compared to the actual measured yield of each field. For this purpose an external validation was performed. This was accomplished by splitting the whole dataset in 70% for training the model in order to underlying the distribution between the years and testing the results with the remaining 30%. Two criteria were used to assess the performance of the models, the root mean square error (RMSE) and the mean absolute error (MAE) that can be computed using the following formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i' - y_i)^2} \quad \text{Eq.3}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i' - y_i| \quad \text{Eq.4}$$

where n is the number of districts used for validation, y'_i is the predicted yield, and y_i is the measured potato yields. RMSE and MAE also converted to percentages (i.e. %RMSE, %MAE) by dividing each by the mean of measured yield statistics (Noorian *et al.*, 2008) and used in

combination to measure the accuracy of the established model. RMSE and MAE values of <10% indicate consistency between the actual yield statistics and predicted yields, values of 10–20% indicate a good predicted result, and values of 20–30% or higher indicate large variations between the predicted and measured yields.

The extensive and varied set of field data were used in this study , allows for further analyses to enhance understanding of results from regression analysis. Hence, several analyses were performed by diving the fields into subgroups.

Sibley *et al.* (2014) indicated that irrigation has been shown to affect the relationship between NDVI and crop yield. A mixture of both irrigated and rain fed fields, allowed an evaluation of the performance under different water conditions in this study. So, the fields were aggregated, as irrigated and non-irrigated according to farmer's description, and linear regression models were developed for each water regime, independently. A second stratification was by grouping the fields according to their cultivated potato varieties. The objective here was to evaluate the performance of several potato varieties in final yield. Both approaches were developed, following the same procedure as described above.

Previous studies (Mkhabela *et al.*, 2011) utilized a fixed calendar day of the year (DOY) to estimate remote sensing based yield predictions models without taking into account the variations in crop phenology among fields. Hence, linear regression based on NDVI was established for a period spanning from mid spring (120 DOY) to late autumn (310 DOY) for each five day interval. Besides, the approaches that used adjusted NDVI values based on phenology metrics or fixed calendar DOYs several studies in field level utilized information related to planting and harvesting DOYs (Bala and Islam, 2009; Bégué *et al.*, 2010). However, the risk in these cases is that the availability of labor or farming machines could be strongly define the agricultural scheduling, having also critical influence on phenology cycle (Pan *et al.*, 2015). Thus, the models were re-estimated using as starting point the planting date of each field as recorded by the farmer and the integrated NDVI values were computed by integrals between the planting and the harvesting DOYs of each parcel, as recommended in previous studies. The aim here was to provide a comparison of the predictive performance when information related to crop phenology are used in yield prediction.

4.3.2 Influence of the Number of Satellite Images

As it was stated (Section 4.1.2), for the curve fitting procedure six parameters should be estimated, this is substantial, considering the number of observations per year per field. Hence, the methodology for yield estimation relies on the availability of cloud free images during the crop's growth cycle. The phenological characteristics of each crop are determined the important growth stages in their crop cycles as a consequence there are considered as important periods for acquisition of satellite observations. This section focuses on the upward and downward slope of the fitted curve which coincides with the parts of the growth cycle where the crop develops, grows and matures.

In general, fields whose time series consisted of too few observations within the growing period failed to be successfully modeled. These parcels enables to test the sensitivity of the proposed methods (Section 4.3.1) to the number of cloud free satellite images used in yield prediction process. For this purpose, the method of integrated values of the NDVI was

selected due to it cover and as it gives an overview of the growing season from emergence to plant kill date for the 122 studied fields.

Using the acquired green-up and senescence dates (Section 4.2) as center points, a time period spanning from -10 to 10 days was specified, in order to count the number of the satellite occurrences within this time period (Figure 12).

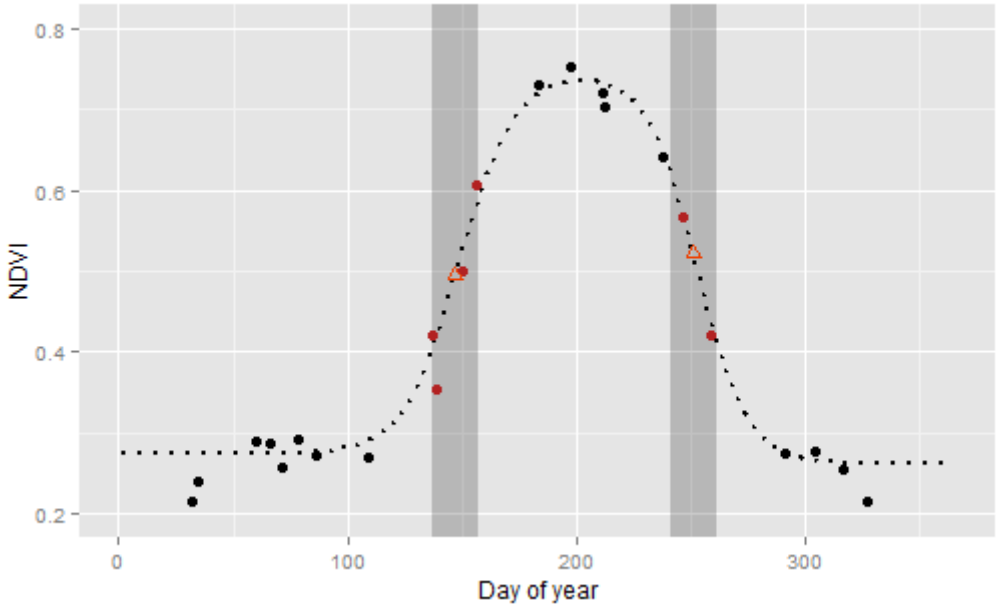


Figure 12: Dotted line represents the NDVI time profile, open symbols represent the estimated greenup and senescence points, solid symbols represent the DMC images and the shaded areas of a 20-days time frame

Subsequently, the fields were grouped into two classes according to the included number of observations within these periods per field. The first class contained all fields which were detected with no observations within 20-days time frame, while the second class contained all fields which were detected with at least one observation within the same period. Then, an empirical linear relationship was developed which takes into consideration the whole dataset of fields. To determine if the number of satellite images had a significant effect in prediction approach the coefficients of determination of the two classes were compared. Simultaneously, using the equation obtained when all the fields were grouped together, the yield of each field, as well as the absolute error were computed to perform an analysis of variance (ANOVA) for each class.

4.4 Influence of meteorological factors

The interaction of the meteorological factors with the crop responses is rather complex. However, by means of studies on determination techniques for assessing physiological crop responses to environmental factors under field conditions it is possible to come up with mathematical models to estimate crop production as a function of meteorological variables. Hence, another possible option for more accurate yield prediction is the incorporation of meteorological factors such as rain, temperature and global radiation who act as regulators by affecting the vegetative development and growth, such as the final yield (Section 2.2.2).

Prior the analysis, the meteorological factors were grouped by 5-day intervals in the same time windows as for each of the 5-day NDVI time steps starting from the greenup DOY of each field(4.3.1). The resulting table, contained field level averages for NDVI, precipitation, temperature and global radiation. In summary, for each field 128 variables were derived. Then, I separately analyzed each of the 5 day time steps by means of stepwise regression models to determine the relationship among yield and the independent variables. Therefore, a subset of explanatory variables, which best explained the potato yield was defined for each of the 5 day intervals. The probability significance threshold of candidate predictors were set to a = 5% and a = 10% for entry and exit in the stepwise regression model, respectively.

Here, the developed stepwise multivariate regression models utilized NDVI and meteorological input parameters and were tested in order to evaluate if the models have been improved compared to linear models where NDVI was used as a standalone yield prediction. I performed the judgment of the best fit based on the estimated R^2 values. However, models with high R^2 in this calibration step, do not necessarily have high predictive power. Therefore, I checked the prediction performance of the models using RMSE and MAE.

In a second approach I included the whole time series of the explanatory variables . It was fully anticipated that having the full season's information than only from a unique time step (5 day intervals) the modeling efforts would be improved due to the accumulation of information. Furthermore, the multivariate regression analysis was performed once again using the integrated NDVI values and the mean weather values in order to give an indication of prediction of yield for the whole growing season.

4.5 Qualitative crop yield classification by data mining techniques

To deal with data mining techniques, such as attribute selection and classification by decision tree, the observed yield data were discretized in three classes, Low-medium (LM), Medium (M) and Medium-high (MH) (Table 2). The discretization performed by K-means technique due to it gives more accurate result than others (Narenda, 2012). In several studies the centroids are specified randomly. In this study the selection of centroids performed based on the actual recorded crop yields from the studied growing seasons. The rationale behind this is that the magnitude of historic crop yields can provide an insight of yield productivity within a specific area. Previously, several studies (Lobell *et al.*, 2002; Lobell *et al.*, 2010) defined the maximum yield as high (e.g., 95^{th}) percentile of the yield distribution for several study regions and crop types. According to the authors, assumed that this approach is a good approximation of the most productive farm within a study area. Extending this approach, the low (e.g., 5^{th}) percentile of the historic yields express in general less efficient fields. Finally, the mean value coincide with the 50^{th} percentile also is calculated. Therefore, I set of the 5^{th} 50^{th} and 95^{th} percentiles of the historic yield data, as initial values of centroid for each cluster, respectively.

Table 1 shows the three classes of yields after discretization, as well as their respective lower and upper limits and number of occurrences. I defined the classes according to historical data of two previous years using k-means simple technique. The centroid of each class

specified as the 42.45, 58.28, 79.97 which are corresponded with the 5^{th} 50^{th} and 95^{th} percentiles of the historic yield data.

Table 2: Limits of classes for the field yield and number of occurrences

Class	**Number of occurrences**	**Lower limit**	Upper limit
		ton/ha	
Low medium	22	40	54
Medium	54	54	69
Medium high	14		>69

Then, NDVI and meteorological data, allowing for the generation of spectral and meteorological attributes for each 5 day intervals after the estimated greenup dates, as well as for the whole cropping season, resulting in 132 attributes for each field.

In this context, the Weka tool (Witten and Frank, 2005) is utilized, which allows to perform data mining to perform classifications using historical crop yields as input attributes. The accumulation of information for each field over the season can be described as a pooled data. In order to determine the more significant attribute the Wrapper's method was used, prior the classification.

Wrapper methods (Kohavi and John, 1997) consider the selection of a set of features as a search problem, where different combinations are prepared, evaluated and compared to other combinations. A predictive model us utilized to evaluate a combination of features and assign a score based on model accuracy. In this research the search process defined as a best-first search. Subsequently, the classification of yield attribute class carrying out by using J48 decision tree algorithm. J48 decision tree algorithm was applied in order to determine the relations and hierarchy of selected attributes.

A first classification was performed using the entire time series of 5 day intervals in order to evaluate the potato yield with a small number of diagnostic features in particular times during the growing season. Additionally, a second classification was performed incorporating also the integrated NDVI and the mean values of rain, temperature and global radiation for the entire growing season, in order to investigate the relevance for the whole period.

As a training set, a cross validation method with 10 folds was used. Accuracy assessment should be an important part of any classification, as it allows quantification of the classification errors. In this section, the accuracy assessment was performed with a confusion matrix in which I compared the classification with the predefined classes after applying K-means technique.

4.6 Yield maps

In the previous sections, approaches have been described for providing qualitative indications of estimated potato yields and for quantifying the expected yields (e.g. ton/ha) during and for the total crop cycle. Both approaches provide useful information about the yield variability among the fields that can be expressed in the form of colorized maps.

In the first place, the maps were developed using the functions derived from the best performing regression models, at specific time steps during the growing season, both purely

NDVI-based and mixed approach where additional meteorological predictor variables were utilized. The best models were assessed using relatively low RMSE of the prediction. This approach is commonly used in order to illustrate the spatial distributions of several crop yields in previous studies (Wang *et al.*, 2014a). Moreover, yield maps were extracted using the results of classification accuracy after applying the data mining techniques. Maps using information for the whole growing period of potato were also developed.

Maps using information for the whole growing season give an approximated picture of the expected yield. In contrast to the yield potato maps for the entire growing season, the within season signaling areas with a color ramp declares the fields that required additional attention and farming actions.

5 Results

The results are based on the methods described in the methodology. The datasets for which the results are presented in sections 5.1 through 5.2 are acquired through the procedure as explained in section 3.4. Sections 5.3 to 5.5 follow the order of the research questions given in Chapter 1.

5.1 NDVI Temporal Profile

The number of available DMC observations for the planted potato fields varied widely between 18 and 26. Especially when observations at the start of the growing season are missing, some of the time series consisted of too few observations to be successfully modelled with the DL and SavGol method. Hence, a total of 31 (10 and 22 for 2013 and 2014, respectively) fields from the total of 122 fields were excluded from the evaluation part (Figure 13). The availability on cloud free DMC images in certain periods, during key parts of the growing season, taking into account in order to remove the aforementioned fields. Complete details are provided in Section 4.3.2.

The RMSE of the two methods, DL and SavGol, estimated for each field within the growth cycle are presented in Table 3. Firstly, the RMSE for a total of 55 for 2013 and 67 for 2014 fields are presented. The average RMSE for the two years for the DL method is 0.0346 and it is ranging from 0.0018 and 0.1281. On the other hand, performance of SavGol method was slightly lower with a range 0.0286 to 0.1012 with a average value of 0.0482. Secondly, fields with not enough observations during key parts of the growing season were removed and RMSE was calculated again for fields that remained (45 for both 2013 and 2014). The adjustments showed positive agreements with the data of DMC for all fields when adequate observations existed to fit the curve. The average RMSE for the two years for the DL method is 0.0212 and it is ranging from 0.0018 and 0.0663. On the other hand, performance of SavGol method was slightly lower with a range 0.0247 to 0.0709 with a average value of 0.0438.

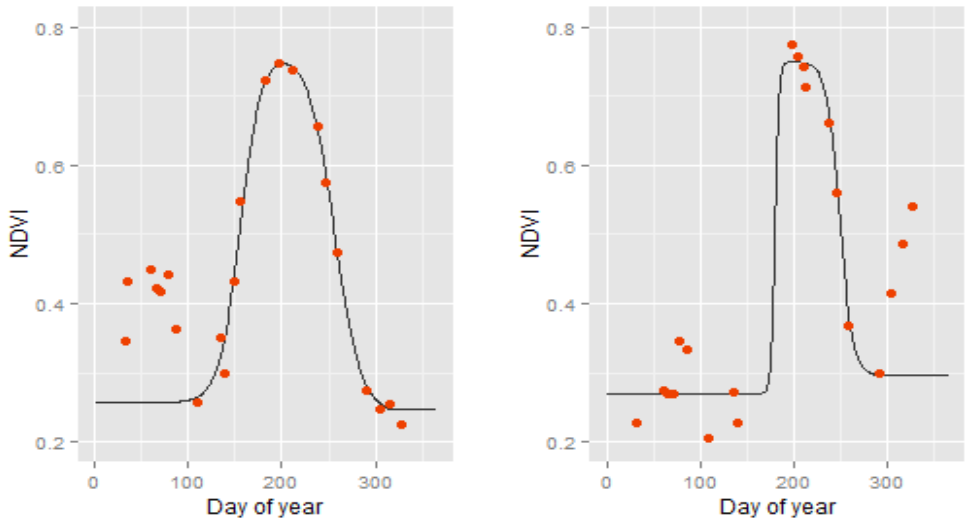


Figure 13: Model fitting results applying DL to available DMC observations for a field containing enough observations (left) and another one containing limited observations (right)

Table 3: RMSE for DL and SavGol method for reconstructing NDVI time series for 2013 and 2014. For each year the RMSE calculated for all the fields and after removing fields with not enough observations to adequately fit curve.

Method	Year	n	RMSE		
			min	max	mean
Double logistic	2013	55	0.0018	0.1997	0.0306
		45	0.0018	0.0405	0.0193
	2014	67	0.0112	0.1281	0.0386
		45	0.0112	0.0663	0.0230
Savitzky Golay	2013	55	0.0286	0.0851	0.0558
		45	0.0408	0.0709	0.0506
	2014	67	0.0247	0.1012	0.0406
		45	0.0247	0.0466	0.0369

Figure 14 shows NDVI modelled values with DL method compared with those of SavGol method, at an example field. It is illustrated that the Savitzky Golay method displays a limitation to mimic the shape of the annual NDVI curves as well as the double logistic function can.

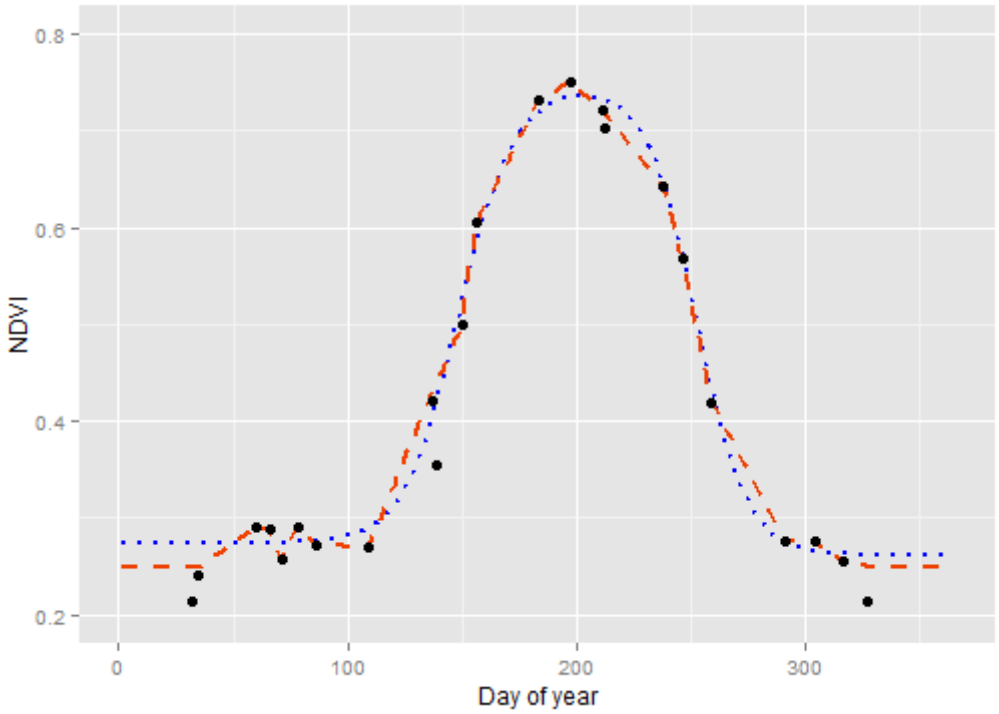


Figure 14: Annual NDVI time series for an example field in the study area using a double logistic function (blue dotted line) and Savitzky Golay with interpolation (red dashed line). Solid symbols represent the DMC images

5.2 Deriving of NDVI time series metrics

Box plots (Figure 15) shows the distribution of estimated greenup and senescence dates, compared with Van den Borne's calendar comments of the planting and foliage killing dates, as they are recorded for each growing season.

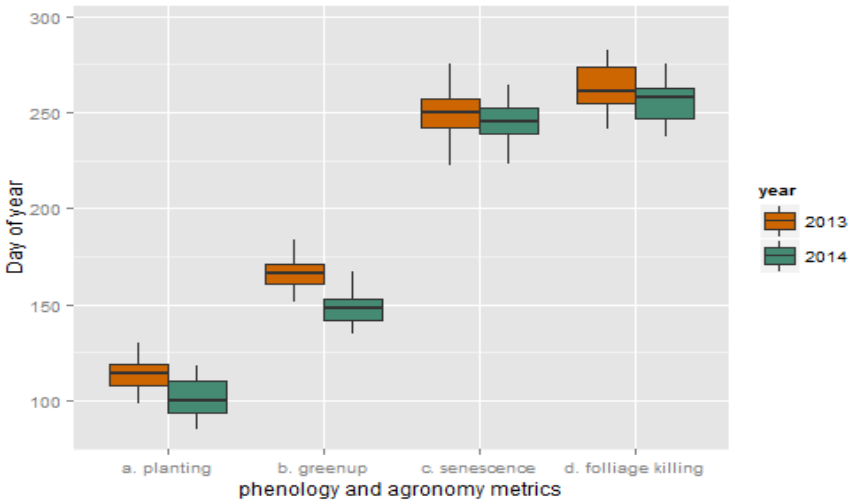


Figure 15: Box plots of seasonality extracted phenological metrics (greenup and senescence) based on DL modeled values in comparison with the recorded agronomic dates (planting and harvesting) as provided by the farmer for 2013 and 2014

The year 2014 had an earlier average greenup onset on 148 DOY than 2013 which had the average detected greenup on 165 DOY. On the other hand the average senescence DOYs were recorded as 244 and 251 DOY for 2014 and 2013, respectively. The inter annual variability in potato greenup dates was large for both years, ranging from DOY 151 to 184 for 2013 and from DOY 135 to 167 from 2014.

The greenup onset dates identified for potato fields showed that there were consistent with the relative planting. In particular, correlation was relative strong with 2013 showing the strongest correlation (R^2 = 0.80), followed by 2014 (R^2 = 0.72). The relationship between foliage killing and senescence dates had a high correlation (R^2 = 0.77) for 2014, but for 2013 was significant lower (R^2 = 0.14).

Using the generated crop phenology parameters, maps were created indicating the spatial distribution of the date of crop greenup/senescence, and duration of season length. These parameters indicate actual crop growth process in a field level. In Appendix A, the phenology parameters of potato in 2013 (Figure 30 a,b,c) and 2014(Figure 31 a,b,c) are presented.

5.3 Statistical Analyses

5.3.1 Estimate yield prediction models

Figure 16 displays the coefficients of determination after linear regression models were compiled for the relation between NDVI and yield, for both growing seasons. Both curves of the coefficient of determination indicated similar patterns and relationships, however, it is obvious that there is a time variation in these patterns between the two years. Various planting dates and contrasting meteorological conditions mainly accounted for this difference. Specifically, maximum correlation for 2013 occurred 70 days after greenup with NDVI showing a strong correlation (R^2=0.67) with potato yield. In 2014, NDVI shows a slightly lower correlation with yield (R^2= 0.63) peaked at 95 dates after greenup.

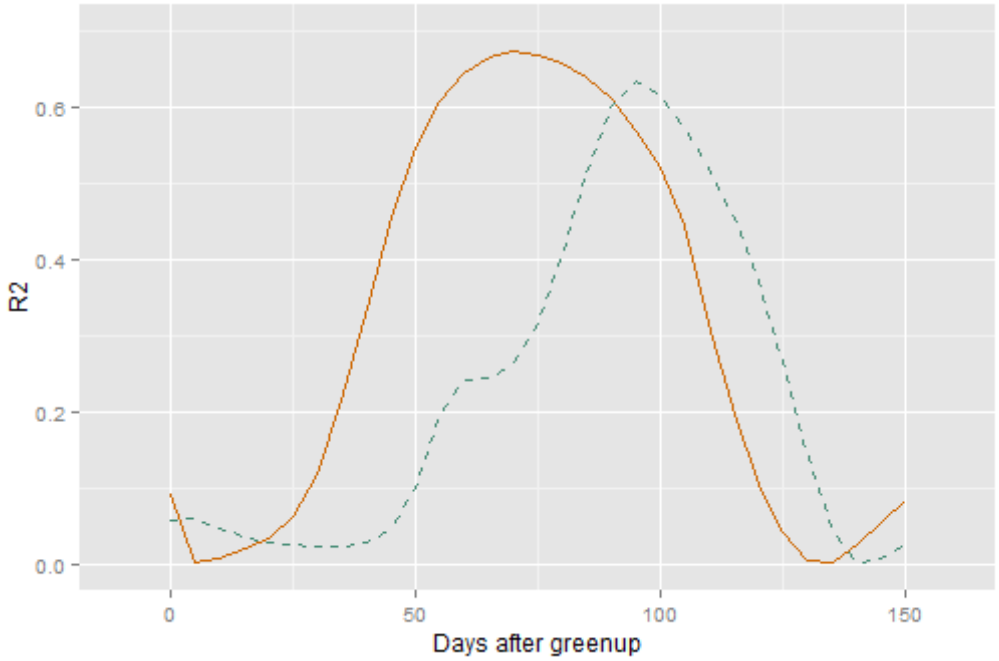


Figure 16: The coefficient of determination (R^2) for a simple regressions between yields and NDVI on each 5 day time step after the greenup date for the two growing seasons. Orange solid line and blue dashed line corresponds with 2013 and 2014 year, respectively

The results also showed a strong relationship between potato yield and integrated NDVI throughout the growing period, as specified by greenup and senescence DOY. In particular, a significant correlation was obtained, with R^2 ranging from 0.71 to 0.63 for 2013 and 2014, respectively.

The scatter plots during the peak correlation time for adjusted NDVI and integrated NDVI with yield are shown in Figure 17. Scatter plots seems to confirm a simple linear regression of potato yield versus NDVI at dates of maximum correlation such as for the whole growing season. Significant relationships of NDVI with yield ($p<0.001$), were observed both for integrated NDVI and for NDVI values at peak correlation time.

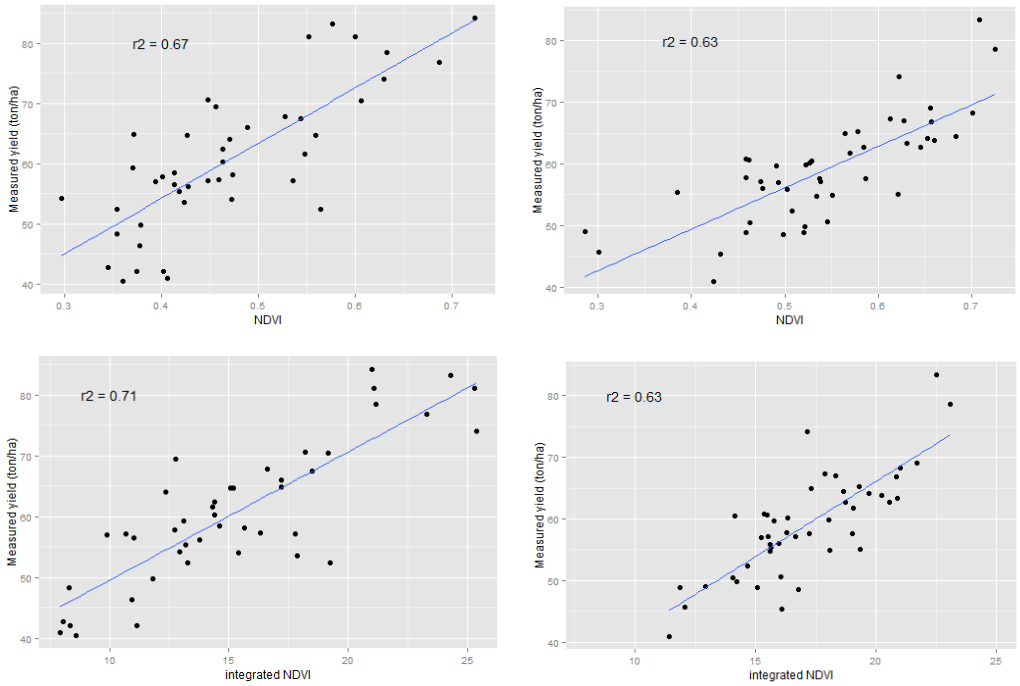


Figure 17: Relationships between NDVI and potato yield at the point of peak correlation (upper left) for 2013 , 70 days after greenup, (upper right) for 2014, 95 days after greenup. Integrated NDVI for 2013 and 2014 are presented in scatter plots in the lower left and right parts, respectively.

The average of coefficients of determination for all the studied years peaked 90 days after the greenup date, showing a significant correlation (R^2=0.61), indicating that the NDVI on the given date is a good predictor for the final field. Hence, this time step was selected and a linear regression model was developed considering all fields together.

Figure 18 shows the scatter plots, when the whole dataset is aggregated over all of the years (2013–2014), The correlation between yield and NDVI at peak time step reached a significant R^2 equal to 0.48 while for integrated NDVI is slightly higher (R^2= 0.65).

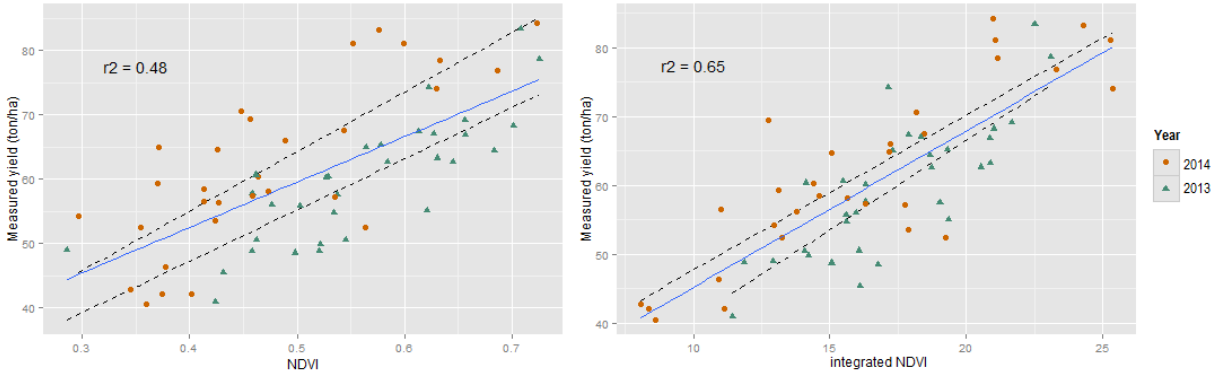


Figure 18: Relationship between measured yield and (left) NDVI 90 days after greenup and (right) integrated NDVI for all studied years and fields. Blue triangles and orange circles correspond with fields of 2013 and 2014, respectively. Best-fit regression lines and associated R^2 are indicated, along with the 1:1 line in blue

Table 4 provides an overview of the estimated linear regression models using the phenologically adjusted NDVI 90 days after greenup and the integrated NDVI, including R^2, R^2 adjusted, p-value, and the equation.

Table 4: Linear regression models for predicting potato yield with NDVI at the time of peak correlation and integrated NDVI.

Method	Equation	R^2	R^2 adjusted	p-value
Days after greenup	$Y = 69.6NDVI_{90} + 21.71$	0.48	0. 468	***
Integrated NDVI	Y = 2.27iNDVI + 22.5	0.65	0.6485	***

The predictive accuracy of the models estimated in validation phase resulting in RMSE and MAE values which provide an overview of how accurate can models predict yields using an external dataset. Comparison between the observed and predicted yield leads to acceptable results (Figure 19). For, model using the NDVI at peak correlation time, RMSE and MAE were 5 (8.36%) and 4.05(6.78%), respectively, while the estimated values using the integrated NDVI with the measured yield were 5.96 (9.97%) and 4.91 (8.23%). Both models demonstrate consistency between the actual and predicted yields. Linear models for all time steps including their prediction performance, RMSE and MAE, are presented analytically in Appendix B.

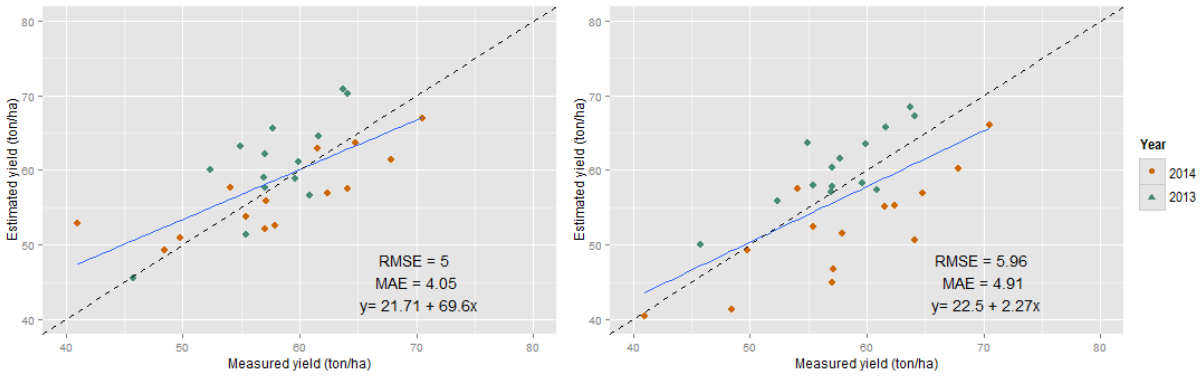


Figure 19: Comparison of measured yields with yields predicted from the linear regression models using an independent test of 27 fields, (left) using NDVI 90 days after greenup and (right) using the integrated NDVI. Blue and orange squares correspond with fields of 2013 and 2014, respectively. Best-fit regression lines (blue line) and RMSE, MAE and model's equation are indicated, along with the 1:1 line in blacked dashed.

The coefficients of determination between NDVI and potato yields observed lower by approximately 15% using a fixed DOY where variations in crop phenology not taken into account. In 2013, NDVI showed a moderate correlation with potato yield (R^2= 0.47) at DOY 245, while in 2014 was five days later reaching a higher correlation (R^2= 0.53). Figure 20 illustrates the annual coefficients of determination between NDVI and potato yield based on fixed calendar dates. For all fields over all years, results showed a maximum correlation (R^2= 0.5) between NDVI and yield at DOY 250. The timing of peak correlation coincides with the tuber filling phase for both years.

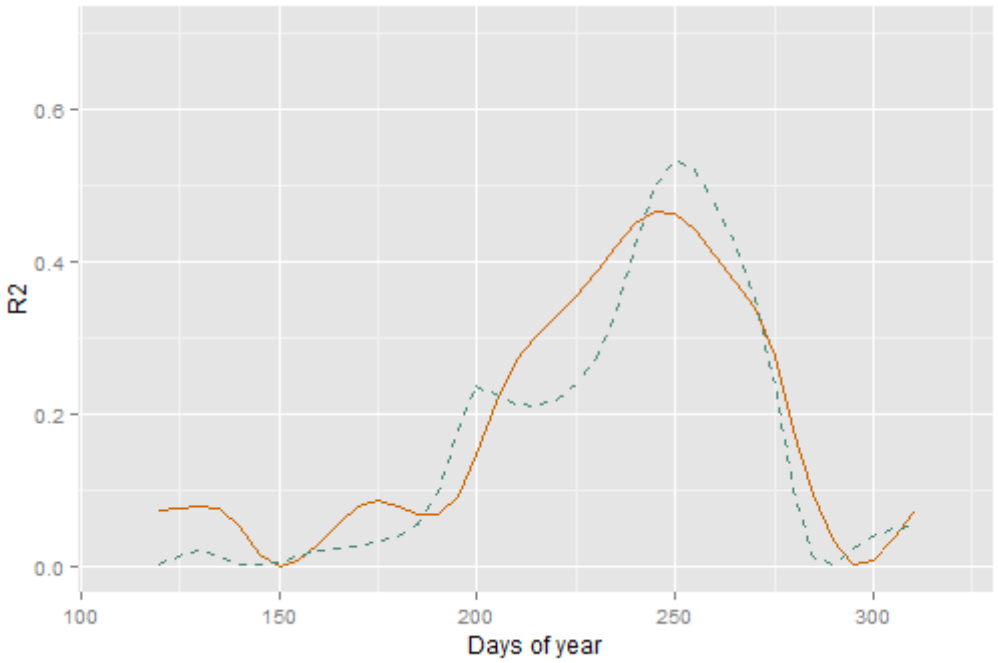


Figure 20: The coefficient of determination (R^2) for a simple regressions between yields and NDVI on each 5 day time step after a fixed calendar day (120 DOY) for the two growing seasons. Orange solid line and blue dashed line corresponds with 2013 and 2014 year, respectively

Finally, R^2 values were estimated using as starting point the planting dates for each fields as provided by farmer's dataset. From the regression analysis with potato yield against NDVI for the entire growing season indicating a variation of R^2 during the different growing seasons. Specifically, the R^2 values for the NDVI ranged from 0 to 0.55 and from 0 to 0.40 for 2013 and 2014, respectively. The peaks were attained 120 and 150 days after potato plantation, respectively. In general, dates when I enumerated the highest values of R^2 corresponds also with the end of tuber filling phase of potato crop as has been recorded using both greenup and a fixed calendar date as starting points in previous sections. Also, it is worth mentioning at this point, that in both years an early peak correlation was detected 50-55 days after plantation, giving a R^2 equal to 0.2 and 0.22 for 2013 and 2014, respectively. Although these correlation coefficients were significant low, this result is in agreement with results reported by Bala and Islam (2009) for potato crops in Bangladesh. In particular, they found that R^2 values of potato yield for the NDVI peaked 48 days after potato plantation, displaying a high correlation (R^2= 0.79). The annual coefficients of determination between NDVI and potato yield based on plantation dates of each field are illustrated in Figure 21.

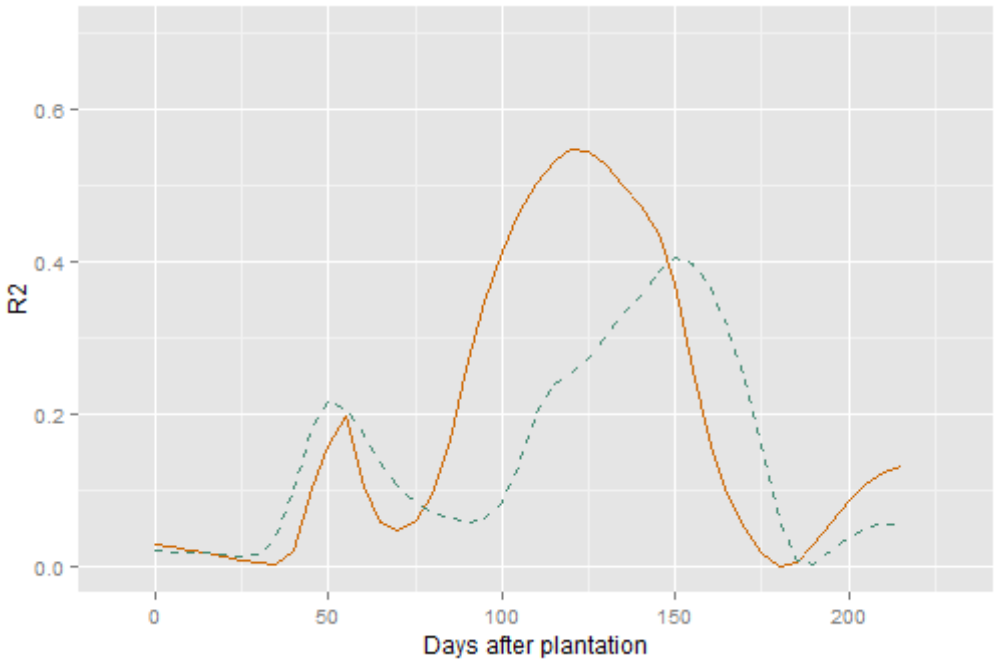


Figure 21: The coefficient of determination (R^2) for simple regressions between yields and NDVI on each 5 day time step after plantations dates of fields for the two growing seasons. Orange solid line and blue dashed line corresponds with 2013 and 2014 year, respectively

From this point on and for the rest of the analysis, only the NDVI values from each five day time step between 0 and 150 days after greenup were used. Subsequently, a variety of several factors that affect model performance were examined.

Firstly, the potato fields were further analyzed on influence of their irrigation regime. Surprisingly, NDVI perform almost equally in irrigated and rain fed fields when linear regression models were compiled between NDVI values 90 days after the greenup and integrated NDVI with yields. Scatter plots (Figure 22) confirm that both rain fed (R^2 = 0.52 and R^2 = 0.64, p< 0.001) irrigated fields(R^2 = 0.45 and R^2= 0.55, p<0.001) had a good correlation and were statistically significant, using both NDVI metric values. An adequate relationship between NDVI and yield for irrigated fields was expected and it has been reported before in literature (Sibley *et al.*, 2014). However, the strong correlation for the rain fed fields was unacceptable. In this part it should be taken into account the high recorded precipitation values during both growing seasons (Figure 7). Figure 22 illustrates that the relationship between NDVI and yield are not significantly varied as function of water regime when there are not obviously variations between statistic yield among irrigated and non irrigated fields.

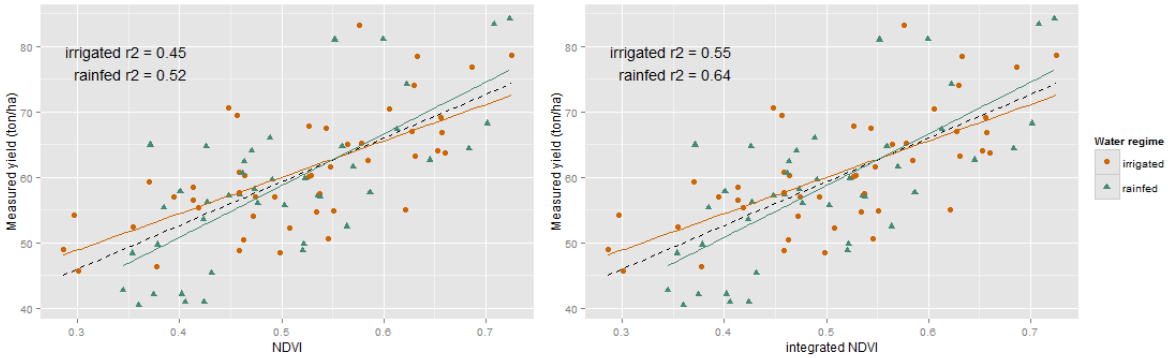


Figure 22: Scatter plots between measured yield and (left) phenologically adjusted NDVI 90 days after greenup and (right) integrated NDVI for all study years (2013-2014). Blue triangles and orange circles correspond with rain fed and irrigated fields, respectively. The best fits are also indicated with relevant colors as well as the best fit for all the fields with the dashed black line.

Figure 23 shows how the relationship between the NDVI metrics and yield can vary as function of variety. In particular, NDVI values for a Miranda variety were low compared to Fontane variety with close similarly yields, indicating that yield could be underestimated when all varieties modeled together. However, the relatively small sample of fields with Miranda variety (14) can resulted to non representative models.

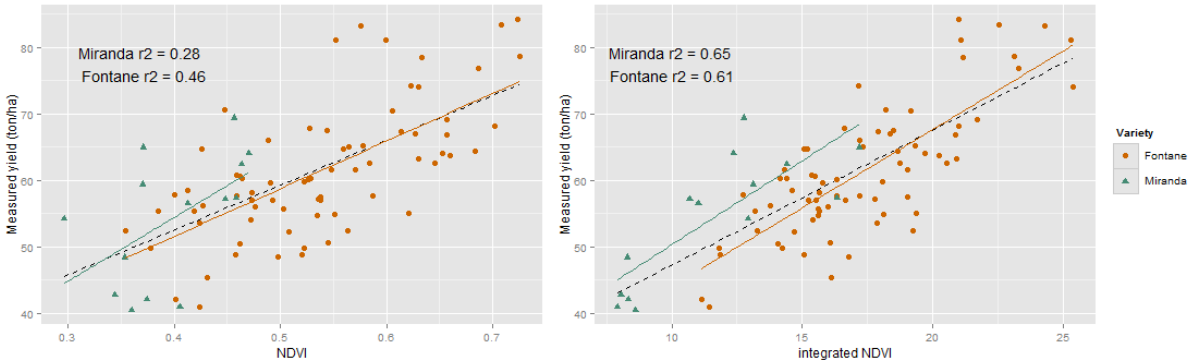


Figure 23: Scatter plots between measured yield and (left) phenologically adjusted NDVI 90 days after greenup and (right) integrated NDVI for all study years (2013-2014). Blue triangles and orange circles correspond with Miranda and Fontane variety, respectively. The best fits are also indicated with relevant colors as well as the best fit for all the fields with the dashed black line.

Based on the better understanding of what factors are correlated with crop yields, the next step was to put them into practice for forecasting. Therefore, Table 5 provides an overview of the estimated linear regression models using the phenologically adjusted NDVI 90 days after greenup and the integrated NDVI taking into consideration the several applied water regimes and crop varieties in the fields. In the case of Miranda variety no linear relationship considered since the number of fields was too small for a relationship to be determined.

Table 5: Linear regression models for predicting potato yield with NDVI at the time of peak correlation and integrated NDVI for various water regimes and varieties

Method	Model	Equation	R^2	R^2 adjusted	p-value
Days after greenup	Irrigated	$Y = 60.04NDVI_{90} + 30.04$	0.3468	0.3427	***
	Rain fed	$Y = 84.28NDVI_{90} + 17.19$	0.5	0.4842	***
	Fontane	$Y = 72.17NDVI_{90} + 22.53$	0.4201	0.408	***
Integrated NDVI	Irrigated	Y = 1.77iNDVI + 31.85	0.4893	0.4723	***
	Rain fed	Y = 2.3iNDVI + 22.95	0.6226	0.6092	***
	Fontane	Y = 2.53iNDVI + 17.96	0.63	0.6223	***

The efficiency of models obtained under the various water regimes and crop varieties was evaluated by comparing the predicted results with the actual measured crop yield statistics using again RMSE and MAE. Generally, comparison between the actual and estimated yields indicated satisfactory results (Table 6). The results achieved using the NDVI value 90 days after greenup indicated close prediction results. In irrigated fields, RMSE and MAE for the potatoes were 9.62% and 8.43%, respectively, while for rain fed both measurements were slightly higher (11.23% and 8.55%). In addition for Fontane variety the RMSE and MAE values of were <10% indicating consistency between the actual yield statistics and predicted yields. The results achieved using the integrated NDVI also showed good predictions results. In irrigated fields, RMSE and MAE for the potatoes were 8.79% and 7.76%, respectively, while those for rain fed were 9.12% and 7.5%. On the other hand Fontane also demonstrated satisfactory results with RMSE equal to 9.84% and MAE equal to 8.3%.

Table 6: RMSE and MAE between predicted and measured yield for water regime and variety models

Method	Model	RMSE	MAE
Days after greenup	Irrigated	5.86 (9.62)	5.13 (8.43)
	Rain fed	6.6 (11.23)	5.02 (8.55)
	Fontane	6.09 (9.91)	4.85 (7.90)
Integrated NDVI	Irrigated	5.35 (8.79)	4.72 (7.76)
	Rain fed	5.36 (9.12)	4.40 (7.50)
	Fontane	6.03 (9.81)	5.10 (8.30)

* percent RMSE and MAE is presented in parentheses

5.3.2 Influence of the Number of Satellite Images

From the 122 fields used in this section 32 an 90 were classified as class 1 and class 2, respectively. Fields in class 1 have no observations in the upward and downward slope while the number of available observations in these parts for class 2 ranging from 1 to 4 cloud free images. Appendix D, contains a map (Figure 34)in which is indicated the spatial variability of fields that contains cloud free images during the crucial pints and fields that have no observations. Observing the map, it is obvious that the fields that have no observations were relative close. When the whole dataset is aggregated over all of classes, the correlation between yield and integrated NDVI was low (R^2=0.3) but still significant.

Figure 24: Scatter plots between measured yield and integrated NDVI for all study years (2013-2014). Orange circles and blue triangles correspond with classes one and two, respectively. The best fits are also indicated with the dashed lines as well as the best fit for all the fields with the blue solid line.

Figure 24 shows a plot of linear relationship between the measured yield and integrated NDVI when both classes presented together. Each point represents one of the studied fields. According to the scatter plot there is some visual evidence that by way of increased scatter the model is less accurate on the class one comparing with class two. The analysis indicated that fields that belong to class one were not well correlated with the measured yield (R^2= 0.045), on the other hand fields which belong in class two were well correlated (R^2= 0.65).

Subsequently, a linear relationship was determined :

$$Y= 1.1583\ iNDVI + 40.863 \qquad \text{Eq.5}$$

Using Equation 5 to calculate the yield for each field an ANOVA was performed. The result shows that the number of available satellite images have a significant influence on the accuracy of the estimation of the yield. Particularly, for the class one p-value was equal to 0.67 while for class two estimated as $p<0.001$. The above strongly indicates the importance of cloud free satellite observations for the developed methodology of this study.

5.4 Influence of meteorological factors

The stepwise regression analysis was first applied separately for each of the 5 day intervals. The results are summarized in Table 7 and appear quite positive, with generally improved R^2 and R^2adj values compared with the approach when NDVI is used as unique predictor. Twenty regression models were developed, based on differing input parameters. Their relative importance can be judged via the partial R^2values, which are tabulated as well. From the wide range of candidate predictors, the stepwise regression always retained at most significant the NDVI. Quite significant is that rainfall and global radiation have always

positive influence on the final yield while temperature varying among the time steps. Only the regression models from 55 to 115 days after greenup are presented due to the rest indicated no significant effect of meteorological parameters.

Table 7: Regression models for the relationship between NDVI (from 30 to 125 days after greenup) and weather data (rainfall, temperature, and global radiation)

Days after greenup	NDVI	precipitation	temperature	radiation	Intercept	R^2	R^2adj
55	181.4057	0.95389			-67.2051	0.423853	0.404323
60	147.5244		-1.2924		-18.3541	0.437143	0.418063
65	107.5604				-13.2099	0.373438	0.362995
70	105.4637	0.94019			-5.64042	0.449736	0.431083
75	92.47041	0.99448	1.794223	1.22189	-7.70065	0.523054	0.489584
80	76.3708		1.034961		-1.6125	0.445005	0.426191
85	69.59916				21.707	0.47973	0.460392
90	82.0788			1.62878	37.54174	0.564168	0.549394
95	81.72176			1.65586	41.36123	0.581576	0.567392
100	79.7197		-1.43137		47.52461	0.522036	0.505834
105	93.22957		-1.44742		46.02087	0.492753	0.475558
110	89.57486				28.73944	0.35804	0.34734
115	105.0549		-1.06026		40.03229	0.326703	0.30388

The results from the second approach where whole time series of the explanatory variables and integrated NDVI among the mean meteorological values for the entire growing season used, are summarized in Table 8. The modeling efforts improved due to the accumulation of information and the R^2 are 0.75 and 0.69, respectively. Despite the fact that there are several independent variables, NDVI values are still the most significant in both cases.

Table 8: Regression models for using the whole time series of predictor variables such as the integrated NDVI with the mean weather values.

Equation	R^2	R2 adjusted
$Y = 66.56NDVI_{95} + 137.68NDVI_{95}+0.8temp_{10}+1.39rad_{10}-73.34$	0.75	0.73
Y = 2.53iNDVI + 8.82 rain + 44.60	0.69	0.68

The R^2 and R^2 adjusted values of Table 7 are repeated in Table 9, together with the estimated RMSE. Observing the Table 9 it can be concluded that good regression models have relatively low errors, however, also weaker regression models can return a low RMSE such as models 80-90 days after greenup. The regression models return a RMSE 4.2 (7%) and 4.51 (7.6%), respectively.

Table 9:Yield estimation root mean square errors for each 5 day interval. R^2 and R^2 adj are copied from Table 6 and results are sorted in the same way.

Days after greenup	R^2	R^2 adj	RMSE	RMSE(%)
30	0.08	0.06	7.07	11.82
35	0.12	0.11	6.93	11.59
40	0.19	0.17	6.77	11.33
45	0.27	0.26	6.73	11.26
50	0.35	0.33	6.70	11.21
55	0.42	0.40	6.34	10.61
60	0.44	0.42	6.53	10.93
65	0.37	0.36	5.49	9.18
70	0.45	0.43	6.34	10.61
75	0.52	0.49	6.60	11.05
80	0.45	0.43	5.15	8.62
85	0.48	0.43	4.97	8.31
90	0.56	0.55	5.91	9.89
95	0.58	0.57	5.00	8.37
100	0.52	0.51	5.64	9.44
105	0.49	0.48	5.64	9.44
110	0.36	0.35	6.00	10.04
115	0.33	0.30	6.10	10.21
120	0.20	0.18	6.76	11.31
125	0.11	0.10	7.00	11.71

5.5 Qualitative crop yield classification by data mining techniques

From the NDVI, precipitation, temperature and radiation values as derived for each field the most significant were selected prior the classification. The Wrapper's method was used to carry out the attribute selection. The aim was to determine which of the attributes had the strongest diagnostic features in order to obtain these most relevant attributes to classify field yield attribute class. The selected attributes for the first classification were the NDVI 60 and 65 days after greenup and also the precipitation values 60 days after the greenup. For the whole growing season integrated NDVI, NDVI 95 and the average values of the precipitation over the season.

Using the selected attributes, the J48 classification algorithm was applied based on a decision tree, in order to perform the first classifications. Figure 25 shows the decision tree for determination of yield class, which was resulted using the relative selected attributes as presented in previous paragraph. The decision tree includes NDVI values 60 days after greenup, precipitation values at the same time step and NDVI values 65 days after greenup as root and internal nodes and classes for nodes are Low Medium, Medium and Medium High.

NDVI 60 days after the greenup, is the attribute that contains much more information and for this reason it has been selected as the first split criteria. Observing the decision tree it can be noticed that if NDVI 60 days after greenup is less than 0.63 then there are 9 out of the total objects that fall in Low Medium (LM) class. Otherwise if rainfall values 60 days after greenup is smaller than 2.46 then there are 39 out of 90 objects that were split among the three classes. The procedure is continued and when the NDVI 65 days after greenup is greater than 0.704 26 fields classified mainly in the higher classes M, respectively. By contrast, when NDVI 65 days after greenup is smaller than 0.704 a new node is created in order to classify the rest of the fields.

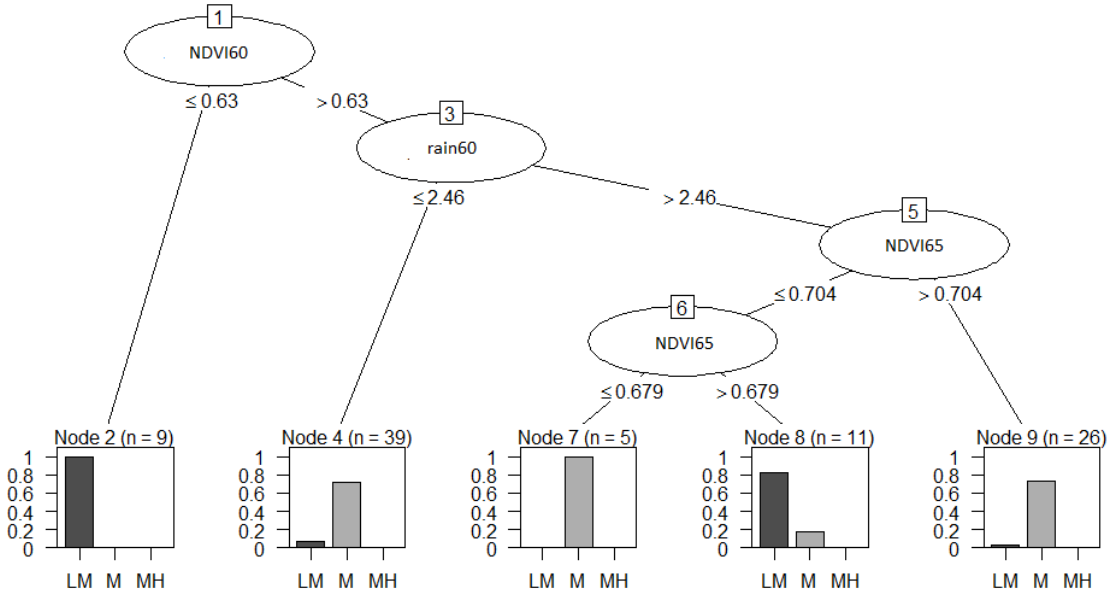


Figure 25: Decision tree for yield classification using NDVI 60 and 65 days after greenup and the average value of precipitation 60 days after the greenup.

Figure 26 shows the decision tree when attributes for the entire growing season are included. In this case, the decision tree includes the integrated NDVI, the average precipitation values during growing season and NDVI values 95 days after greenup as root and internal nodes and classes for nodes are Low Medium, Medium and Medium High.

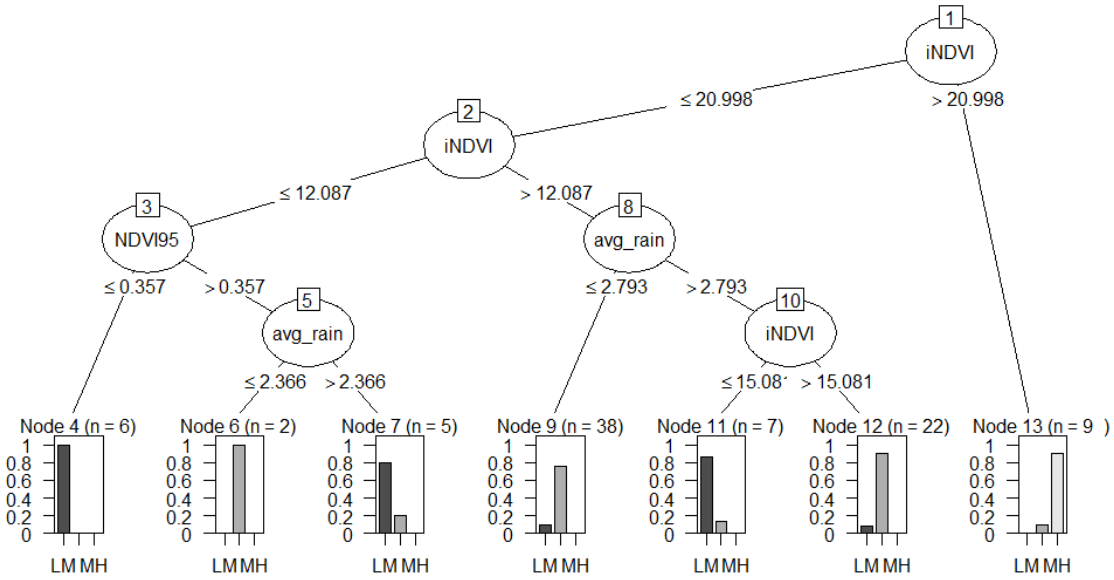


Figure 26: Decision tree for yield classification using integrated NDVI, NDVI95 days after greenup and the average value of precipitation for the entire growing season.

Observing the decision trees (Figure 25 and Figure 26), it is clear that the NDVI 60 days after greenup and the integrated NDVI attributes which are located on top of the trees play the most important role in yield classification, respectively in both cases.

Table 10 and Table 11 present the confusion matrices and accuracies of the performed classifications. The diagonal values represent the correctly classified instances. In the first classification from a total of 90 instances, the J48 classifier rated 61 correctly, corresponding to 67.78% accuracy. Here, is important to note that only the MH class fail to be classified correctly. Table 11 shows the classification results for the second approach corresponding which express the entire growing season. There was a improvement regarding the classification of yield within the growing season. Therefore in this case the classifier rated correctly 65 instances, corresponding to 72.2% accuracy.

Table 10: Confusion matrix and classification accuracy for yield classification using NDVI60,NDVI65 and rain60 attributes.

	LM	M	MH	User's Accuracy [%]
LM	13	9	0	59.1
M	6	48	0	88.9
MH	0	14	0	0
Producer's Accuracy [%]	68.42	67.60	0	
Overall Accuracy[%]: 67.78				

Table 11: Confusion matrix and classification accuracy for yield classification using integrated NDVI,NDVI95 and average rain over the season attributes.

	LM	M	MH	User's Accuracy [%]
LM	7	15	0	31.82
M	4	49	1	90.74
MH	0	5	9	64.29
Producer's Accuracy [%]	63.6	71.01	90	
Overall Accuracy[%]: 72.2				

5.6 Yield maps

The resulting map using the NDVI values 90 days after the greenup consisted of one layer, whereby each field in the map had a value for the predicted potato yield (Figure 27). After generating the yield map, it can be observed that yield variability, ranging from 42.6 to 67.6 ton/ha. The above could be a valuable source of information for the farmer from the point of view of expected yield. In particular, the yield map documented the spatial distribution of crop yield, hence provide a kind of priority list with the fields that more action and attention are required. In Appendix C, also, a potato yield map generating by the equation making use of NDVI and meteorological variables is presented (Figure 32). For the entire growing season a map (Figure 33) was generated using the empirical equation of integrated NDVI.

Moreover, a map making use of qualitative information is also presented in this section (Figure 28). In particular, the map indicates which of the fields are classified correctly by the J48 classifier during the growing season. The map was generated using the results from the confusion matrix (Table 10). Following the same steps and making use of confusion matrix presented in Table 11, a map providing information for the whole growing period was generated (Figure 29).

Figure 27: Predicted yield (ton/ha) maps for potato, based on linear regression models using NDVI values 90 days after greenup

Figure 28: Yield map as resulted after classification with J48 classifier using NDVI 60 and 65 days after greenup and the precipitation values 65 days after greenup

Figure 29: Yield map as resulted after classification with J48 classifier, using integrated NDVI, average precipitation values over the growing season and the NDVI values 95 days after greenup

6 Discussion

In this chapter different aspect of this research are critically reviewed, considering the research questions proposed in section 1.3.

6.1 NDVI Temporal Profile

The current study made use of two data methods for reconstructing annual NDVI time series profile: the DL method (Beck et.al 2006) and a method using Savitzky–Golay filters smoothing approach combined with a linear interpolation based on the available images (Pan et.al, 2015).

The accuracy of curve fitting depends on a variety of factors including the quality and temporal resolution of DMC data. Missing data during key parts of the growing season strongly influence the result of curve fitting (Figure 13). Limitations imposed by cloud cover can be seen in Table 3, where the RMSE for both DL and SavGol method improved when fields with missing observations during critical periods of growth cycle were removed. The results were similar to a study conducted by Zhang et al. (2006) where they concluded that it does not make sense to apply algorithms for noise reduction in cases where are there are more than two consecutive missing 16-day MODIS composites during snow free periods.

The DL function describes the NDVI data better than the SavGol method during the growth cycle for both study years (RMSE (DL) < RMSE (SavGol), $p < 0.001$ in all cases) (Table 3). Figure 14 illustrated that the Savitzky Golay method displays a limitation to mimic the shape of the annual NDVI curves as well as the double logistic function can. The previous studies (Beck *et al.*, 2006; Julien and Sobrino, 2010) indicated that the use of a double logistic function is appropriate for observing unimodal NDVI curves, hence the function as outlined in these papers can also describe well the growth cycle of the potato crop. In addition, double logistic is demonstrated as more able and efficient to preserve the high NDVI values than the Savitzky Golay method which tended to overestimate them during the growing season. These findings are in accordance with those of (Hird and McDermid, 2009), where they compared several techniques for noise reduction of NDVI time series.

The DL function showed significant small RMSE in previous studies (Julien and Sobrino, 2010) working fine for regional vegetation, the findings of this research show that it is also applicable for crop monitoring in agricultural areas at field level.

6.2 Deriving of NDVI time series metrics

Assessment of field-level phenology detection results showed that DMC time series data can provide a good detection of phenological metrics (Figure 15). An evaluation demonstrated that the greenup date at field level were consistent with the relative planting for both years. The unexpected low correlation between senescence and foliage killing dates for 2013 generate some undesirable inconsistencies. One possible explanation may arise by taking into account two important factors. First of all, in this research the senescence date was estimated as the inflection point in the downward slope of the annual NDVI curve. From this definition the senescence date corresponds to the time when the growing rate of the crop is

minimum, which is before the complete maturation of the crop. This is why the senescence date is earlier than the foliage killing dates. Additionally, it should be considered that foliage killing procedure is highly dependent on the agricultural scheduling of the farmers. Hence, the agricultural scheduling are eventually resulting in the end of crop season from the agronomist point of view.

A closer assessment of the results, indicates that the greenup onset dates of 2013 were much later than the relative dates of 2014 (Figure 15). Subtle shift in timing of crop phenology reveals annual variations of crop growth and that can be due to the variation of climatic conditions or the schedule of the agricultural activities.

As explained in section 3.3 there are differences among the meteorological factors of two studied years (Figure 7) hence, the contrasting meteorological conditions result to specify the agricultural scheduling having also a directly effect to the phenology metrics.

The earlier planting dates in 2014 may prolonged the rest period until more favorable conditions for growth occur, while in 2013 a rapid increase of NDVI values occurred from the emergence point (greenup) until the time when NDVI reach the maximum value. Based on the above and observing the results in section 5.2, it can be concluded that growing periods estimated quite satisfactory for both years. Mapping results in Appendix A indicate also the different range among the growth periods for 2013 and 2014.

Additionally, the magnitude of the variations (~30days) among the greenup dates of the fields highlights the important role of using phenologically adjusted VIs in order to fulfill agricultural applications through time series remote sensing data, as also observed in previous studies (Bolton and Friedl, 2013; Mulianga *et al.*, 2013; Wang *et al.*, 2014b).

To conclude, the results were acceptable ranging within a specific range of values which doesn't exceed the limits of whole crop season, as delimited by planting and foliage killing dates. The detection phenology of metrics (green up, senescence and growing period) was satisfactory in field level using the daily NDVI time series of 2013 and 2014 deriving from the DL method. DMC provide the necessary spatial, spectral and temporal resolution for precision farming applications overcoming limitations of large satellites (Sandau *et al.*, 2010).

6.3 Statistical Analyses

6.3.1 Yield prediction models

The main objective of this study is to provide recommendations for the optimal time period for the estimation of yield during the growing season, such as for the entire season prior to harvest. Additionally, aim of the study was to test how integrated NDVI can be used for potato yield assessment before the harvest. In general, yield prediction based on remotely sensed vegetation indices like NDVI indicates the amount of existing biomass and the vigor of the crops. However, forecasting below ground crop yields such as potato displayed some difficulties, due to the indirect relation between the belowground yield and spectral data related to aboveground green biomass (Hayes and Decker 1996).

The idea of using 5 day time intervals for investigate the best time step within the growing season was first applied in the study conducted by Bolton and Friedl (2013). Although, studies focus on potato have not been carried out following the proposed methodology, the results are in agreement with several studies where phenologically adjusted time series were used for other crops yield forecasting (Bolton and Friedl, 2013; Wang *et al.*, 2014b). Comparing results of Figure 16, Figure 20 and Figure 21, It can be summarized that the particular models based on phenologically adjusted NDVI values can provide substantial benefits for yield prediction models, result in a general improvement of 15% in the correlation coefficient values.

The results showed that phenologically adjusted NDVI 90 days after greenup and integrated NDVI resulted in a good correlations (Table 4 and Figure 19) with the yield and gave comparable RMSE at the field level (<10%). The MAE values demonstrated a similar trend but were slightly lower than the RMSE values(Figure 19Figure 21). However, NDVI shows a limitation to capture the biomass variations when the full canopy was developed and also influenced by the soil background reflectance. In general, during the end of tubers filling phase and extending to early maturity phase, the leaves tends to be yellowish due to the degradation of the amounts of chlorophyll. It can be concluded that these changes in the color of leaves influence the reflectance in visible bands with a result to better capture of the yield variations. The above seems to be confirmed by the fact that annual NDVI curves for both growing seasons displayed a high standard deviation in their NDVI values in period following the maximum NDVI until the foliage killing dates (Figure 8).

This piece of information could be useful in precision agriculture point of view where the farmers can assess to delay the last application of fertilizers in the fields, in order to distinguish those fields that really need additional management for increasing the final yield. Although the results show a possibility of using remote sensing data to correlate and forecast field level yield within season, the contrasting meteorological conditions between a two year-study don't provide an ideal scenario to achieve accurate results regarding the exact time for early yield predictions. A more effective research should contain images and yield data for more than two growing seasons.

It is also important to note that the NDVI values acquired during leaf development and extending to tuber initiation phase resulted in achieving low accuracy for estimation of yield, 0.2 and 0.22 for 2013 and 2014, respectively (Figure 21). These results are in disagreement with the findings of Bala and Islam (2009), who concluded that the use of a single image acquired 48 or 64 days after the plantation were well correlated and can substantially predict the final yields with a moderate accuracy (RMSE > 15%). However, the statistical significance of the obtained regression model was not investigated in their research. In this sense it is necessary to take a closer look into the differences in methodologies to find possible reasons of deviations.

Knowing that 50% of the studied fields were irrigated combined with the high rainfall levels recorded during the two growing seasons, the effects of soil reflectance could be significant due to the varying surface soil moisture levels among the fields. An enhanced yield prediction can theoretically be achieved by developing models using vegetation indices that

can reduce the effect of background soil reflectance. SAVI as introduced by Huete (1988) can be an alternatively and it had been already tested in previous studies (Neale and Sivarajan et. al 2011) capturing better the yield variations.

Previous studies (Neale and Sivarajan et. al 2011) indicated that capturing the final yield in potato crops is mainly related to the duration of the green leaf area and not only to the peak leaf area index. This condition can be seen in the good relationship which is obtained when the yield is related to the integrated area under the NDVI curve (Figure 19).

The findings of this study are in agreement with others (Bégué *et al.*, 2010; Morel *et al.*, 2014) which previously have shown that early season NDVI values alone were not accurate estimators for crop yields and especially for tuber crops. The above rationale suggests that a large aboveground biomass, early in the season is not a determinant of the final yield and vice versa. Earlier studies (Gomez-MacPherson and Richards, 1995) also concluded that vegetation development prior to the tuber filling phase determines mainly the plant structure but not definitely the final production. Several factors occur later in the season such as meteorological factors or potentially diseases could have an impact in yields to a greater extent.

It was concluded from the results (Table 5)that there existed a good correlation for both rain fed and irrigated fields, in this research. On the other hand, (Sibley *et al.*, 2014), indicated that irrigated yield variations can more successfully captured than rain fed yields for maize. A noticeable point was the equal performance in the final yields for irrigated and non irrigated here as explained in Section 3.1, while Sibley *et al.* (2014) explained their results mainly due to the relatively poor performance in rain fed fields. Inconsistencies between the two results lead to question the performance of stratification approach in general. It should be taken account that farmers should make more efficient application of water in their fields. Yuan *et al.* (2003) confirmed that water stress is one of most crucial factor that affects the final yield and is related not only to applied irrigation regimes but also to the soil water storage. Hence, soil type of the fields need to be considered as an important factor in estimating the yield accurately in future studies. At the same time, the relatively small sample of fields cultivated with Miranda variety didn't permit extraction of clear conclusions on whether the stratification of potato varieties give a better estimation or not.

6.3.2 Influence of the Number of Satellite Images

The second objective in this research was to explore the influence of available cloud free images in yield estimation accuracy. In general, limited references in the literature citing the influence of the number of satellite images in yield forecasting. Morel *et al.* (2014), have been shown that a minimum of 5 satellite images during the growing season have to be acquired in order to correctly describe the dynamic of the NDVI for utilizing in yield prediction applications. Their results are in agreement with the results from the ANOVA. In addition, the presence of fields with no satellite observations in crucial points of the growing season seems to affect in the relationship between NDVI and measured yield (R^2 = 0.3) (Figure 24).

In the light of the above, the importance of cloud free satellite observations for the developed methodology of this study was indicated. The problem of obtaining satellite imagery in cloudy locations often is cited as an obstacle for precision farming applications (Zhang and Kovacs, 2012). Previous studies indicated that the over mentioned problem can be partially overcome by using multiple sensory platforms (Gevaert *et al.*, 2015) or by utilizing Unmanned Aerial Vehicle (UAV) imagery during the cloudy days. Forthcoming satellite systems, such as Sentinel-2, with a 10m spatial resolution, and a high visiting frequency, will provide a better access to farm level information.

6.4 Influence of meteorological factors

The third objective of this research was to analyze if the predictive power of regression models can be improved by the concomitant use of NDVI together with bio-climatic indicators. Although the coefficients of the regression models are restricted to the scope of the current study, the parameters included in the stepwise regression are important (Table 7). In summary, NDVI factor still explains the most of the yield variability, although incorporating of meteorological factors led to improved regression models compared with those on specific time steps using only NDVI. For instance the stepwise multivariate regression models utilized NDVI and global radiation 90 days after greenup resulted a higher R^2 in comparison with the regression model where NDVI was the unique predictor. This could indicate that crop status around 85-95 days after greenup plays an important role in potato yield; information which could be useful for PA. The results reinforce that the modeling efforts in which meteorological data were incorporated are generally strong with more significant correlation to final yield (Prasad *et al.*, 2006; Balaghi *et al.*, 2008) but the predictive power wasn't always significant stronger. Furthermore, it is important to note that all regression models presented in this section were able to predict the yield to a RMSE of between 4.97 and 7.07 ton/ha. For the whole growing season the integrated NDVI is the most significant predictor. However the regression model indicated the strong relation between irrigation and in general availability of water with the potato crop due to its crop characteristics.

Furthermore, the 5 day intervals of the meteorological explanatory variables does not necessarily result in the highest correlations and enhanced results can be obtained when the data are first integrated in more relevant periods during vegetative and growth development of potato. The purpose of this study was to examine the most optimal time step for accurate forecasting within the growing season. Further analysis can be conducted on the relationships of potato yields to when the data are first integrated in more relevant periods during vegetative and growth development of potato.

In summary, the unexplained variance may be due a plethora of other factors , mainly errors in the basic inputs (yield statistics, weather and remote sensing data) and effects not covered by the regression models (diseases, soils, cultural practices, etc.).

The multiple stepwise regression approach implemented in this study, apparently displayed a drawback for precision agriculture application due to the weather parameters derived from the closest KNMI meteorological station in Eindhoven. Recent studies presented approaches to overcome the limited spatial coverage of the meteorological stations.

Specifically, Carrer *et al.* (2012) derived solar radiation, while Kloog *et al.* (2012) estimated near surface temperature, utilizing satellite data. Although, the low resolutions of several dozen meters still constitute an impediment of the explicitly use of weather spatial data, the results can be noticeably improved. Likewise, via the use of advanced technologies daily weather records can be obtained by stations located near the farm (Hadders *et al.*, 2009). Hence, weather measurements can be provided at a finer scale to characterize the local weather conditions, accurately.

Finally, it must be noted that many meteorological indicators, especially if they are derived from satellites as well, are not really independent from vegetation indices(Johnson, 2014). In a study conducted by Prasad et al. (2006) multiple regression models were develop by introducing environmental information and NDVI, as independent variables. In this study the authors checked the multicollinearity among the variables using the variance inflation factor.

6.5 Qualitative crop yield classification by data mining techniques

In general, the results presented moderate level of accuracy, however, useful conclusions can be extracted taking a closer look. The results obtained using Wrapper's method with J48 decision tree algorithm are in agreement with previous studies from the scope of the overall accuracy (Fernandes *et al.*, 2011). Overall, the accuracies of both classifications were over 65%. Hence, it can be resulted that this approach can provide a good basis for qualitative assessment of yield for several crop types and in various region of interests.

For the entire growing season, integrated NDVI has been found again as the most appropriate measure to characterize the crop production (Mulianga *et al.*, 2013; Morel *et al.*, 2014). Also the selection of NDVI 95 days after greenup is consistent with previous finding where a strong relation 85-95 days after greenup between NDVI and yield is illustrated. However, a more interesting and a new knowledge is the finding of the classification approach during the growing season. NDVI 60 and 65 days, respectively, selected as the most significant attributes. As summarized in earlier studies (Johnson, 2014), using entire time series of NDVI in data mining techniques can provide useful information for some dates which are not return a high correlation coefficient as single optimal dates.

Subsequently, the results indicated in both cases a relationship between the precipitation and yields due to it is a general dependency that crops need rainfall to thrive. The absence of temperature and global radiation selection is probably related with the method used to discretize the yield. Three signaling classes obtained from K-means method may not have been adequate to relate with the numeric meteorological attributes.

There was a significant worsening regarding the classification of yield LM, whose hit percentage dropped from 59.1 to 31.82 (Table 10 and Table 11). Classification of M class yield improved, increasing from 88.9 to 90.74 its accuracy percentage, when the entire growing season is studied. To conclude, for classification of MH class I observed the most significant change in the results. Ranging from 0 to 64.29 percentage in user's accuracy it is obvious that the yield of potatoes mainly developed during the maturation phases hence early predictions can be dramatically failed. Excluding the MH class in the first classification I

observed a general balance and coherence of results. However, confusion matrices (Table 10 and Table 11) indicated that there was more confusion between neighboring classes (LM and M or M and MH) than between remote classes (LM and MH). In future studies, subject of sentence is missing should be taken into account of other discretization methods or more classes. The most important result is that during the growing season and especially using attributes 60 and 65 after the greenup we can achieve a significant accuracy for LM class, 59.81%. All the above can inform the farmers to make appropriate crop management such as rate and timing of fertilization and irrigation, for optimizing the yield quantity and quality (Wu *et al.*, 2007).

7 Conclusions & recommendations

The aim of the current study was involved to develop and validate methods for prediction and classification of potato yields, using imagery from high spatiotemporal DMC data in order to inform precision farming applications.

Firstly, reconstructing NDVI time-series remote sensing data has been proven feasible, using the double logistic function. The last one describes the NDVI data better than smoothing-interpolation approach, as quantified by the RMSE (SavGol>DL). It has also been shown that removing fields with missing observations during critical periods of growth cycle result in an enhancement of performance of curve fitting methods.

When enough observations were available then relevant phenology parameters could be derived from NDVI time series profiles and be used to develop potato yield prediction models, during and over the entire growing season. Meanwhile, a fixed calendar date and planting date were also used to propose a yield prediction model and it was compared to the above models indicating that inclusion of information related to crop phenology significantly improved model performance. The advantages in yield forecasting depends on the available cloud free data during growing season. The correlation coefficient between yield and NDVI increased from R^2=0.3 to R^2=0.65, when fields with adequate number of satellite observation used in yield estimation approach.

Yield prediction models based on the NDVI as a stand-alone predictor resulted in almost the same accuracy as yield prediction models based on both NDVI and meteorological factors, during the cropping season. In both case reliable estimation achieved at the beginning of maturity phase, with RMSE values of 8.36% and 8.31% at model validation, respectively. A significant correlation was observed between the estimated yields obtained from NDVI-based models (p-value < 0.001).

However, the most promising results were obtained by incorporating the whole growing seasons' worth of data in a decision tree model. As a consequence, an earlier estimation was achieved, approximately 55-60 days after green-up, during the tuber filling phase. Although, the results were more effective for LM and M classes with 59.1% and 88.9% accuracy, respectively, the physical dimension of this information will help inform decisions of farmers on agronomic management and especially for effectual irrigation and fertilization during the formation of the potato bulk.

Future tasks to improve the framework in this study are summarized below:

There is a need for further investigation on the soil physical and chemical properties at field level. Soil properties such as pH, electrical conductivity, water holding capacity have a significant effect on the final yield. All these properties should be taking into consideration in order to stratify relevant zones that could considerably improve the prediction for final crop.

Further research could focus on the investigation of using other vegetation indices, like WDVI and SAVI which can be calculated using DMC's satellite bands and test if provide a better estimation of the final yields. Future Earth Observing satellite systems, such as

Sentinel-2 (ESA), with higher spectral resolution will give access to estimation of a plethora of VIs.

With the abundance of availability of several data and especially for meteorological data the interrelation of the different input variables should be considered and corrected when integrating bio-climatic and spectral indicators into multiple regression models. This is not taking into account in this research.

Moreover, a limitation of the approach is that for fitting the DL function a true periodic length of satellite data is required. As author, I recommend for possible next studies to try to fit a logistic function that describes the upward slope of annual NDVI curve in order to perform analyses also in the middle of the season. Hence, different starting parameters should be selected for curve fitting procedure.

Finally, the perspectives of UAV imagery offering high spatial data and enabling to flight during cloudy days, hence they should be tested as an alternative to overcome the obstacle of cloud contamination during the agricultural seasons.

References

Atzberger, C., 2013. Advances in remote sensing of agriculture: Context description, existing operational monitoring systems and major information needs. Remote Sensing 5, 949-981.

Baisch, S., Bokelmann, G.H.R., 1999. Spectral analysis with incomplete time series: an example from seismology. Computers & Geosciences 25, 739-750.

Bala, S.K., Islam, A.S., 2009. Correlation between potato yield and MODIS-derived vegetation indices. International Journal of Remote Sensing 30, 2491-2507.

Balaghi, R., Tychon, B., Eerens, H., Jlibene, M., 2008. Empirical regression models using NDVI, rainfall and temperature data for the early prediction of wheat grain yields in Morocco. International Journal of Applied Earth Observation and Geoinformation 10, 438-452.

Beck, P.S.A., Atzberger, C., Høgda, K.A., Johansen, B., Skidmore, A.K., 2006. Improved monitoring of vegetation dynamics at very high latitudes: A new method using MODIS NDVI. Remote Sensing of Environment 100, 321-334.

Becker-Reshef, I., Vermote, E., Lindeman, M., Justice, C., 2010. A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data. Remote Sensing of Environment 114, 1312-1323.

Bégué, A., Lebourgeois, V., Bappel, E., Todoroff, P., Pellegrino, A., Baillarin, F., Siegmund, B., 2010. Spatio-temporal variability of sugarcane fields and recommendations for yield forecast using NDVI. International Journal of Remote Sensing 31, 5391-5407.

Begue, A., Vintrou, E., Saad, A., Hiernaux, P., 2014. Differences between cropland and rangeland MODIS phenology (start-of-season) in Mali. International Journal of Applied Earth Observation and Geoinformation 31, 167-170.

Blackmore, S., 1994. Precision farming: an introduction. Outlook on Agriculture 23, 275-280.

Bolton, D.K., Friedl, M.A., 2013. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. Agricultural and Forest Meteorology 173, 74-84.

Brisson, N., Gate, P., Gouache, D., Charmet, G., Oury, F.X., Huard, F., 2010. Why are wheat yields stagnating in Europe? A comprehensive data analysis for France. Field Crops Research 119, 201-212.

Brown, M.E., de Beurs, K., Vrieling, A., 2010. The response of African land surface phenology to large scale climate oscillations. Remote Sensing of Environment 114, 2286-2296.

Carrer, D., Lafont, S., Roujean, J.L., Calvet, J.C., Meurey, C., Le Moigne, P., Trigo, I.F., 2012. Incoming solar and infrared radiation derived from METEOSAT: Impact on the modeled land water and energy budget over France. Journal of Hydrometeorology 13, 504-520.

Cassman, K.G., 1999. Ecological intensification of cereal production systems: Yield potential, soil quality, and precision agriculture. Proceedings of the National Academy of Sciences of the United States of America 96, 5952-5959.

Cassman, K.G., Dobermann, A., Walters, D.T., Yang, H., 2003. Meeting cereal demand while protecting natural resources and improving environmental quality. Annual Review of Environment and Resources, pp. 315-358.

Chahbi, A., Zribi, M., Lili-Chabaane, Z., Duchemin, B., Shabou, M., Mougenot, B., Boulet, G., 2014. Estimation of the dynamics and yields of cereals in a semi-arid area using remote sensing and the SAFY growth model. International Journal of Remote Sensing 35, 1004-1028.

Chen, J., Huang, J., Hu, J., 2011. Mapping rice planting areas in southern China using the China Environment Satellite data. Mathematical and Computer Modelling 54, 1037-1043.

Coelho, D.T., Dale, R.F., 1980. An Energy-Crop Growth Variable and Temperature Function for Predicting Corn Growth and Development: Planting to Silking. Agronomy Journal 72, 503-510.

Cong, N., Piao, S., Chen, A., Wang, X., Lin, X., Chen, S., Han, S., Zhou, G., Zhang, X., 2012. Spring vegetation green-up date in China inferred from SPOT NDVI data: A multiple model analysis. Agricultural and Forest Meteorology 165, 104-113.

Dancey, D., Bandar, Z.A., McLean, D., 2007. Logistic model tree extraction from artificial neural networks. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 37, 794-802.

Doraiswamy, P.C., Sinclair, T.R., Hollinger, S., Akhmedov, B., Stern, A., Prueger, J., 2005. Application of MODIS derived parameters for regional crop yield assessment. Remote Sensing of Environment 97, 192-202.

Estrella, N., Sparks, T.H., Menzel, A., 2009. Effects of temperature, phase type and timing, location, and human density on plant phenological responses in Europe. Climate Research 39, 235-248.

FAO, 2009. High Level Expert Forum - How to Feed the World in 2050.

Fernandes, J.L., Rocha, J.V., Lamparelli, R.A.C., 2011. Sugarcane yield estimates using time series analysis of spot vegetation images. Scientia Agricola 68, 139-146.

Fischer, A., 1994. A simple model for the temporal variations of NDVI at regional scale over agricultural countries. Validation with ground radiometric measurements. International Journal of Remote Sensing 15, 1421-1446.

Funk, C., Budde, M.E., 2009. Phenologically-tuned MODIS NDVI-based production anomaly estimates for Zimbabwe. Remote Sensing of Environment 113, 115-125.

Funk, C., Verdin, J.P., Husak, G., 2007. Integrating observation and statistical forecasts over sub-Saharan Africa to support Famine Early Warning. 87th AMS Annual Meeting.

Gallego, J., Carfagna, E., Baruth, B., 2010. Accuracy, Objectivity and Efficiency of Remote Sensing for Agricultural Statistics. Agricultural Survey Methods. John Wiley & Sons, Ltd, pp. 193-211.

Gevaert, C., Suomalainen, J., Tang, J., Kooistra, L., 2015. Generation of Spectral–Temporal Response Surfaces by Combining Multispectral Satellite and Hyperspectral UAV Imagery for Precision Agriculture Applications. Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of PP, 1-7.

Gevaert, C.M., García-Haro, F.J., 2015. A comparison of STARFM and an unmixing-based algorithm for Landsat and MODIS data fusion. Remote Sensing of Environment 156, 34-44.

Godfray, H.C.J., Beddington, J.R., Crute, I.R., Haddad, L., Lawrence, D., Muir, J.F., Pretty, J., Robinson, S., Thomas, S.M., Toulmin, C., 2010. Food security: The challenge of feeding 9 billion people. Science 327, 812-818.

Gomez-MacPherson, H., Richards, R.A., 1995. Effect of sowing time on yield and agronomic characteristics of wheat in south-eastern Australia. Australian Journal of Agricultural Research 46, 1381-1399.

Gonzalez-Sanchez, A., Frausto-Solis, J., Ojeda-Bustamante, W., 2014. Predictive ability of machine learning methods for massive crop yield prediction. Spanish Journal of Agricultural Research 12, 313-328.

Grassini, P., Thorburn, J., Burr, C., Cassman, K.G., 2011. High-yield irrigated maize in the Western U.S. Corn Belt: I. On-farm yield, yield potential, and impact of agronomic practices. Field Crops Research 120, 142-150.

Gregory, P.J., Ingram, J.S.I., Andersson, R., Betts, R.A., Brovkin, V., Chase, T.N., Grace, P.R., Gray, A.J., Hamilton, N., Hardy, T.B., Howden, S.M., Jenkins, A., Meybeck, M., Olsson, M., Ortiz-Monasterio, I., Palm, C.A., Payn, T.W., Rummukainen, M., Schulze, R.E., Thiem, M., Valentin, C., Wilkinson, M.J., 2002. Short communication: Environmental consequences of alternative practices for intensifying crop production. Agriculture, Ecosystems and Environment 88, 279-290.

Hadders, J., Hadders, J.W.M., Raatjes, P., 2009. Agri Yield Management: Practical solutions for profitable and sustainable agriculture based on advanced technology. Precision Agriculture 2009 - Papers Presented at the 7th European Conference on Precision Agriculture, ECPA 2009, pp. 397-401.

Hankui, Z., Chen, J.M., Bo, H., Huihui, S., Yiran, L., 2014. Reconstructing Seasonal Variation of Landsat Vegetation Index Related to Leaf Area Index by Fusing with MODIS Data. Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of 7, 950-960.

Hatfield, J.L., Prueger, J.H., 2010. Value of using different vegetative indices to quantify agricultural crop characteristics at different growth stages under varying management practices. Remote Sensing 2, 562-578.

Heumann, B.W., Seaquist, J.W., Eklundh, L., Jönsson, P., 2007. AVHRR derived phenological change in the Sahel and Soudan, Africa, 1982-2005. Remote Sensing of Environment 108, 385-392.

Hird, J.N., McDermid, G.J., 2009. Noise reduction of NDVI time series: An empirical comparison of selected techniques. Remote Sensing of Environment 113, 248-258.

Huete, A.R., 1988. A soil-adjusted vegetation index (SAVI). Remote Sensing of Environment 25, 295-309.

Jakubauskas, M.E., Legates, D.R., Kastens, J.H., 2002. Crop identification using harmonic analysis of time-series AVHRR NDVI data. Computers and Electronics in Agriculture 37, 127-139.

Johnson, D.M., 2014. An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. Remote Sensing of Environment 141, 116-128.

Jonsson, P., Eklundh, L., 2002. Seasonality extraction by function fitting to time-series of satellite sensor data. Geoscience and Remote Sensing, IEEE Transactions on 40, 1824-1832.

Jönsson, P., Eklundh, L., 2004. TIMESAT - A program for analyzing time-series of satellite sensor data. Computers and Geosciences 30, 833-845.

Julien, Y., Sobrino, J.A., 2010. Comparison of cloud-reconstruction methods for time series of composite NDVI data. Remote Sensing of Environment 114, 618-625.

Kissinger, K., Herold, M., De Sy, V., 2012. Drivers of Deforestation and Forest Degradation, A Synthesis Report for REDD+ Policymakers.

Kloog, I., Chudnovsky, A., Koutrakis, P., Schwartz, J., 2012. Temporal and spatial assessments of minimum air temperature using satellite surface temperature measurements in Massachusetts, USA. Science of the Total Environment 432, 85-92.

Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artificial Intelligence 97, 273-324.

Kooistra, L., Clevers, J., Beza, E., Van Vliet, P., Van Den Borne, J., Van Der Velde, W., 2012. Opportunities for Sentinel-2 in an integrated sensor approach to support decision making in precision agriculture. European Space Agency, (Special Publication) ESA SP.

Laborte, A.G., de Bie, K.C.A.J.M., Smaling, E.M.A., Moya, P.F., Boling, A.A., Van Ittersum, M.K., 2012. Corrigendum to "Rice yields and yield gaps in Southeast Asia: Past trends and future outlook" [Eur. J. Agron. 36 (2012) 9-20]. European Journal of Agronomy 43, 96.

Lin, M., Huybers, P., 2012. Reckoning wheat yield trends. Environmental Research Letters 7.

Lobell, D.B., 2013. The use of satellite data for crop yield gap analysis. Field Crops Research 143, 56-64.

Lobell, D.B., Asner, G.P., Ortiz-Monasterio, J.I., Benning, T.L., 2003. Remote sensing of regional crop production in the Yaqui Valley, Mexico: Estimates and uncertainties. Agriculture, Ecosystems and Environment 94, 205-220.

Lobell, D.B., Ortiz-Monasterio, J.I., Addams, C.L., Asner, G.P., 2002. Soil, climate, and management impacts on regional wheat productivity in Mexico from remote sensing. Agricultural and Forest Meteorology 114, 31-43.

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

Lobell, D.B., Ortiz-Monasterio, J.I., Lee, A.S., 2010. Satellite evidence for yield growth opportunities in Northwest India. Field Crops Research 118, 13-20.

Löw, F., Michel, U., Dech, S., Conrad, C., 2013. Impact of feature selection on the accuracy and spatial uncertainty of per-field crop classification using Support Vector Machines. ISPRS Journal of Photogrammetry and Remote Sensing 85, 102-119.

Lyle, G., Lewis, M., Ostendorf, B., 2013. Testing the temporal ability of landsat imagery and precision agriculture technology to provide high resolution historical estimates of wheat yield at the farm scale. Remote Sensing 5, 1549-1567.

Matson, P.A., Parton, W.J., Power, A.G., Swift, M.J., 1997. Agricultural intensification and ecosystem properties. Science 277, 504-509.

Menzel, A., 2003. Plant phenological anomalies in Germany and their relation to air temperature and NAO. Climatic Change 57, 243-263.

Mingwei, Z., Qingbo, Z., Zhongxin, C., Jia, L., Yong, Z., Chongfa, C., 2008. Crop discrimination in Northern China with double cropping systems using Fourier analysis of time-series MODIS data. International Journal of Applied Earth Observation and Geoinformation 10, 476-485.

Mkhabela, M.S., Bullock, P., Raj, S., Wang, S., Yang, Y., 2011. Crop yield forecasting on the Canadian Prairies using MODIS NDVI data. Agricultural and Forest Meteorology 151, 385-393.

Morel, J., Todoroff, P., Bégué, A., Bury, A., Martiné, J.F., Petit, M., 2014. Toward a satellite-based system of sugarcane yield estimation and forecasting in smallholder farming conditions: A case study on reunion island. Remote Sensing 6, 6620-6635.

Mulianga, B., Bégué, A., Simoes, M., Todoroff, P., 2013. Forecasting regional sugarcane yield based on time integral and spatial aggregation of MODIS NDVI. Remote Sensing 5, 2184-2199.

Noorian, A.M., Moradi, I., Kamali, G.A., 2008. Evaluation of 12 models to estimate hourly diffuse irradiation on inclined surfaces. Renewable Energy 33, 1406-1412.

Pan, Z., Huang, J., Zhou, Q., Wang, L., Cheng, Y., Zhang, H., Blackburn, G.A., Yan, J., Liu, J., 2015. Mapping crop phenology using NDVI time-series derived from HJ-1 A/B data. International Journal of Applied Earth Observation and Geoinformation 34, 188-197.

Panda, S.S., Ames, D.P., Panigrahi, S., 2010. Application of vegetation indices for agricultural crop yield prediction using neural network techniques. Remote Sensing 2, 673-696.

Pinter Jr, P.J., Hatfield, J.L., Schepers, J.S., Barnes, E.M., Moran, M.S., Daughtry, C.S.T., Upchurch, D.R., 2003. Remote sensing for crop management. Photogrammetric Engineering and Remote Sensing 69, 647-664.

Prasad, A.K., Chai, L., Singh, R.P., Kafatos, M., 2006. Crop yield estimation model for Iowa using remote sensing and surface parameters. International Journal of Applied Earth Observation and Geoinformation 8, 26-33.

Quinlan, J.R., 1996. Improved use of continuous attributes in C4.5. Journal of Artificial Intelligence Research 4, 77-90.

Ramírez, D.A., Yactayo, W., Gutiérrez, R., Mares, V., De Mendiburu, F., Posadas, A., Quiroz, R., 2014. Chlorophyll concentration in leaves is an indicator of potato tuber yield in water-shortage conditions. Scientia Horticulturae 168, 202-209.

Reynolds, C.A., Yitayew, M., Slack, D.C., Hutchinson, C.F., Huetes, A., Petersen, M.S., 2000. Estimating crop yields and production by integrating the FAO Crop Specific Water Balance model with real-time satellite data and ground-based ancillary data. International Journal of Remote Sensing 21, 3487-3508.

Roerink, G.J., Menenti, M., Verhoef, W., 2000. Reconstructing cloudfree NDVI composites using Fourier analysis of time series. International Journal of Remote Sensing 21, 1911-1917.

Rudorff, B.F.T., Batista, G.T., 1990. Spectral response of wheat and its relationship to agronomic variables in the tropical region. Remote Sensing of Environment 31, 53-63.

Sakamoto, T., Wardlow, B.D., Gitelson, A.A., Verma, S.B., Suyker, A.E., Arkebauer, T.J., 2010. A Two-Step Filtering approach for detecting maize and soybean phenology with time-series MODIS data. Remote Sensing of Environment 114, 2146-2159.

Sakamoto, T., Yokozawa, M., Toritani, H., Shibayama, M., Ishitsuka, N., Ohno, H., 2005. A crop phenology detection method using time-series MODIS data. Remote Sensing of Environment 96, 366-374.

Sandau, R., Brieß, K., D'Errico, M., 2010. Small satellites for global coverage: Potential and limits. ISPRS Journal of Photogrammetry and Remote Sensing 65, 492-504.

Sarquís, J.I., González, H., Bernal-Lugo, I., 1996. Response of two potato clones (S. tuberosum L.) to contrasting temperature regimes in the field. American Potato Journal 73, 285-300.

Savitzky, A., Golay, M.J.E., 1964. Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry 36, 1627-1639.

Shekoofa, A., Emam, Y., Shekoufa, N., Ebrahimi, M., Ebrahimie, E., 2014. Determining the Most Important Physiological and Agronomic Traits Contributing to Maize Grain Yield through Machine Learning Algorithms: A New Avenue in Intelligent Agriculture. PLoS ONE 9, e97288.

Sibley, A.M., Grassini, P., Thomas, N.E., Cassman, K.G., Lobell, D.B., 2014. Testing remote sensing approaches for assessing yield variability among maize fields. Agronomy Journal 106, 24-32.

Sqm.com, (2015). Potato. [online] Available at: http://www.sqm.com/en-us/productos/nutricionvegetaldeespecialidad/cultivos/papa.aspx [Accessed 21 May 2015].

Stutte, G.W., Yorio, N.C., Wheeler, R.M., 1996. Interacting effects of photoperiod and photosynthetic photon flux on net carbon assimilation and starch accumulation in potato leaves. Journal of the American Society for Horticultural Science 121, 264-268.

Tao, F., Yokozawa, M., Xu, Y., Hayashi, Y., Zhang, Z., 2006. Climate changes and trends in phenology and yields of field crops in China, 1981-2000. Agricultural and Forest Meteorology 138, 82-92.

United Nations, 2013. World Population Prospects: The 2012 Revision.

Uno, Y., Prasher, S.O., Lacroix, R., Goel, P.K., Karimi, Y., Viau, A., Patel, R.M., 2005. Artificial neural networks to predict corn yield from Compact Airborne Spectrographic Imager data. Computers and Electronics in Agriculture 47, 149-161.

Vitousek, P.M., Mooney, H.A., Lubchenco, J., Melillo, J.M., 1997. Human domination of Earth's ecosystems. Science 277, 494-499.

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

Wang, M., Tao, F.L., Shi, W.J., 2014b. Corn yield forecasting in northeast china using remotely sensed spectral indices and crop phenology metrics. Journal of Integrative Agriculture 13, 1538-1545.

White, M.A., de Beurs, K.M., Didan, K., Inouye, D.W., Richardson, A.D., Jensen, O.P., O'Keefe, J., Zhang, G., Nemani, R.R., van Leeuwen, W.J.D., Brown, J.F., de Wit, A., Schaepman, M., Lin, X., Dettinger, M., Bailey, A.S., Kimball, J., Schwartz, M.D., Baldocchi, D.D., Lee, J.T., Lauenroth, W.K., 2009. Intercomparison, interpretation, and assessment of spring phenology in North America estimated from remote sensing for 1982-2006. Global Change Biology 15, 2335-2359.

White, M.A., Thornton, P.E., Running, S.W., 1997. A continental phenology model for monitoring vegetation responses to interannual climatic variability. Global Biogeochemical Cycles 11, 217-234.

Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann Publishers Inc.

Wu, J., Wang, D., Rosen, C.J., Bauer, M.E., 2007. Comparison of petiole nitrate concentrations, SPAD chlorophyll readings, and QuickBird satellite imagery in detecting nitrogen status of potato canopies. Field Crops Research 101, 96-103.

Wu, W.-b., Yang, P., Tang, H.-j., Zhou, Q.-b., Chen, Z.-x., Shibasaki, R., 2010. Characterizing Spatial Patterns of Phenology in Cropland of China Based on Remotely Sensed Data. Agricultural Sciences in China 9, 101-112.

Yang, C., Anderson, G.L., 2000. Mapping grain sorghum yield variability using airborne digital videography. Precision Agriculture 2, 7-23.

Yuan, B.Z., Nishiyama, S., Kang, Y., 2003. Effects of different irrigation regimes on the growth and yield of drip-irrigated potato. Agricultural Water Management 63, 153-167.

Zhang, C., Kovacs, J.M., 2012. The application of small unmanned aerial systems for precision agriculture: A review. Precision Agriculture 13, 693-712.

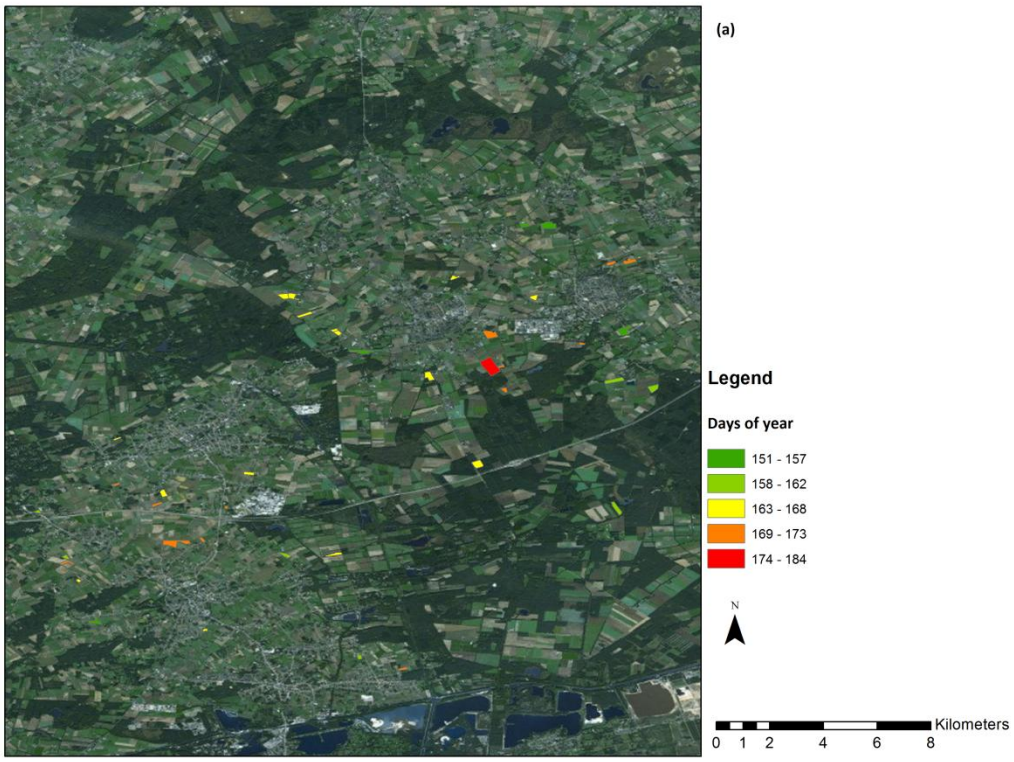
Zhang, X., Friedl, M.A., Schaaf, C.B., 2006. Global vegetation phenology from Moderate Resolution Imaging Spectroradiometer (MODIS): Evaluation of global patterns and comparison with in situ measurements. Journal of Geophysical Research: Biogeosciences 111, G04017.

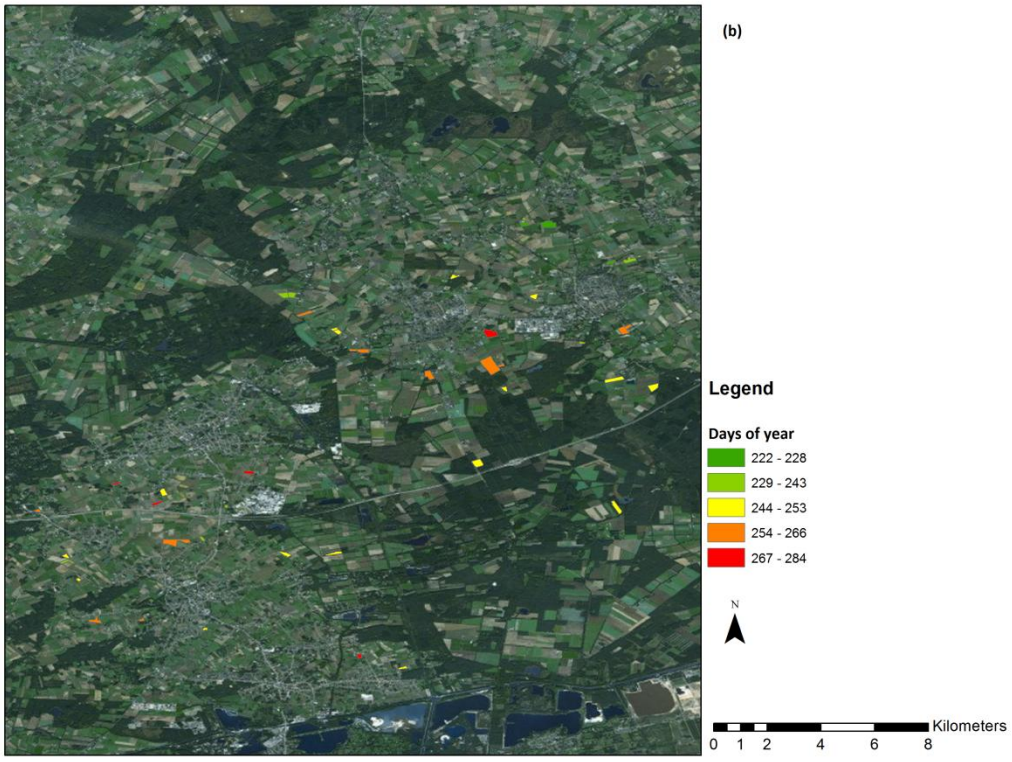
Zhang, X., Friedl, M.A., Schaaf, C.B., Strahler, A.H., Hodges, J.C.F., Gao, F., Reed, B.C., Huete, A., 2003. Monitoring vegetation phenology using MODIS. Remote Sensing of Environment 84, 471-475.

Zhu, W., Pan, Y., He, H., Wang, L., Mou, M., Liu, J., 2012. A changing-weight filter method for reconstructing a high-quality NDVI time series to preserve the integrity of vegetation phenology. IEEE Transactions on Geoscience and Remote Sensing 50, 1085-1094.

Appendices

Appendix A: Seasonality parameters maps





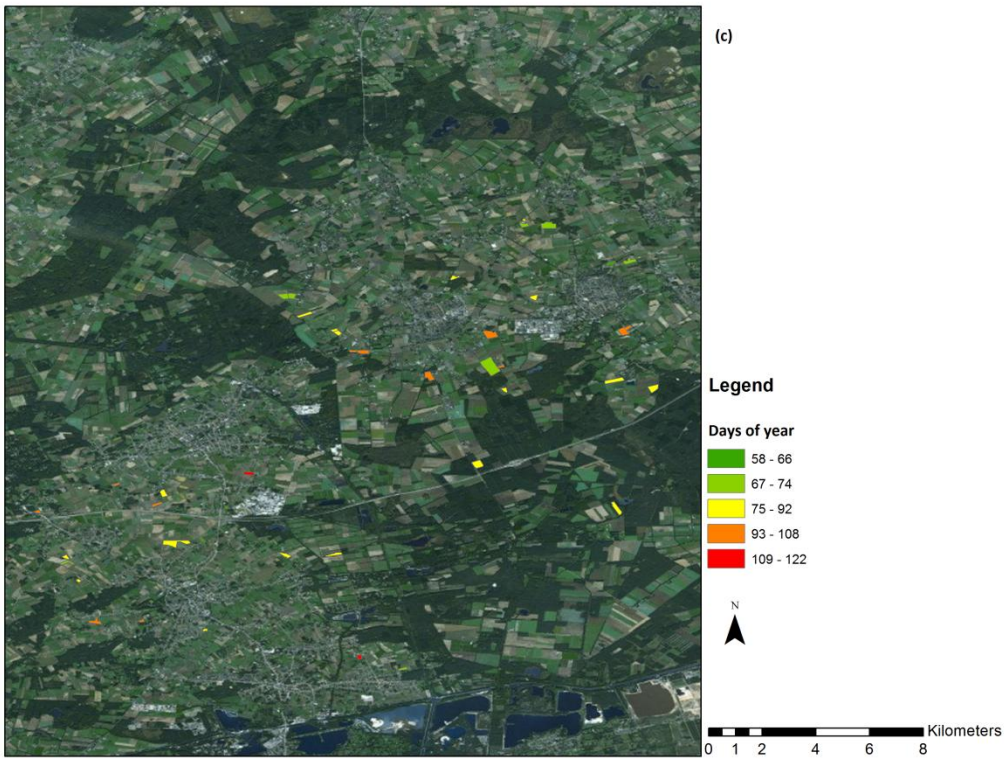
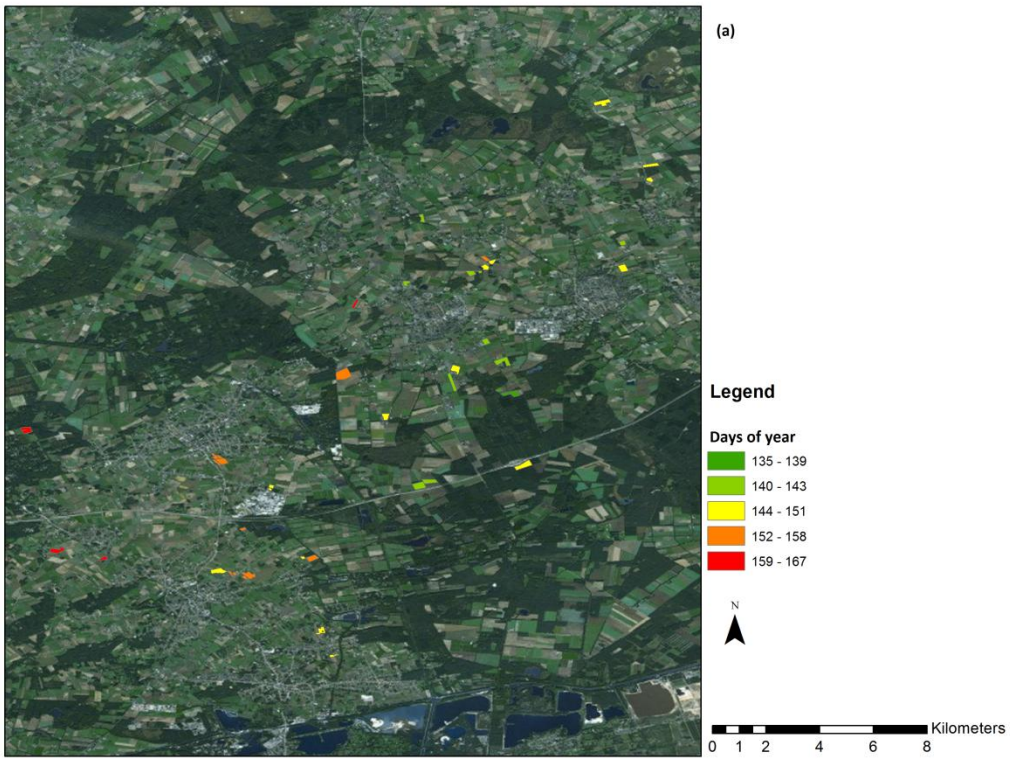
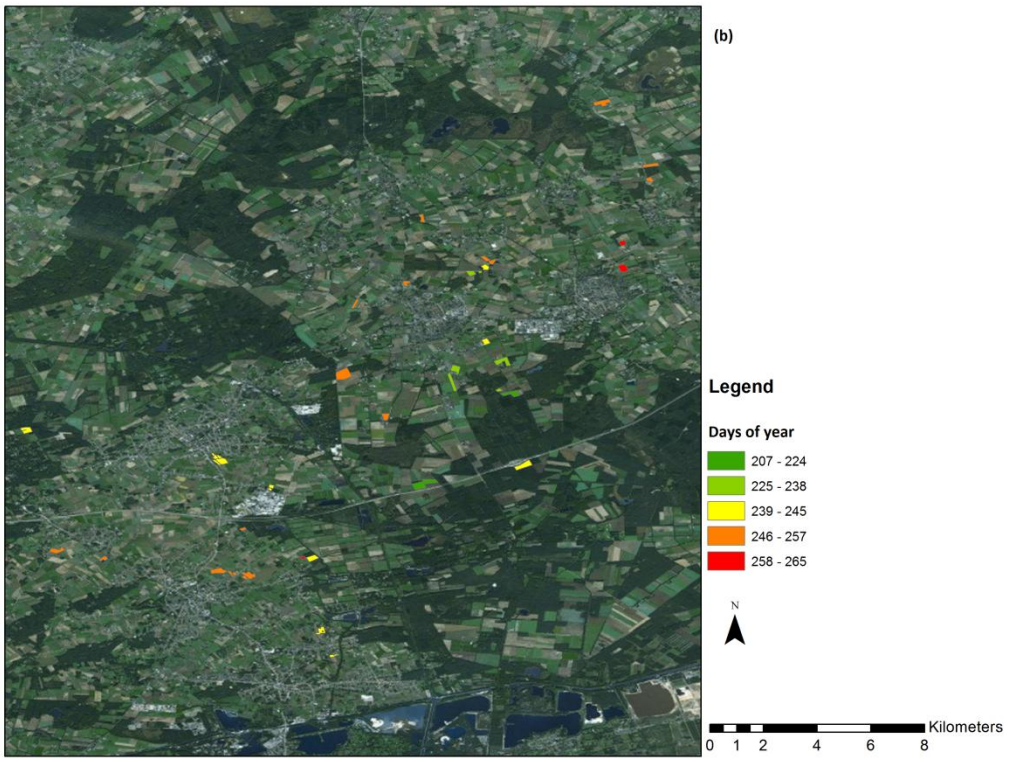


Figure 30: Seasonality parameters extraction for potato in 2013: (a) green-up date, (b) senescence date, (c) length of growth duration (from green-up to senescence)





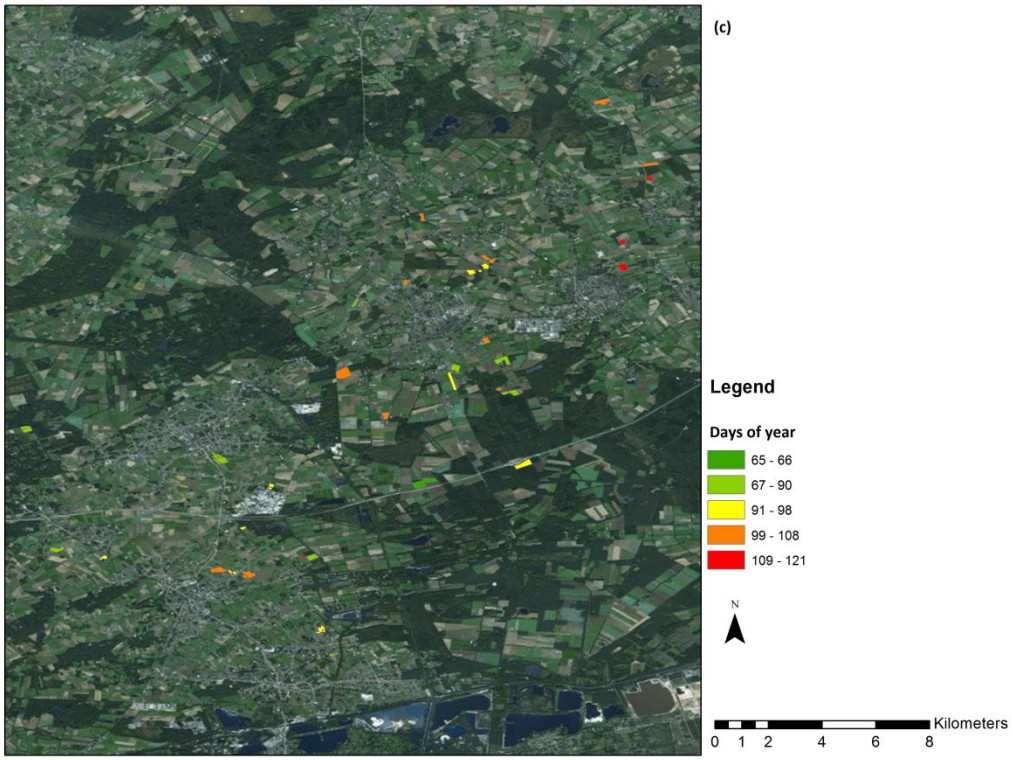


Figure 31: Seasonality parameters extraction potato in 2014: (a) green-up date, (b) senescence date, (c) length of growth duration (from green-up to senescence)

Appendix B: Linear Model Summary

Table 12: Quantitative relationships between potato yields (Y) and NDVI. Models were developed using a training set of 63 fields and validated using an independent test set of 27 fields

Model calibration			Model Validation			
Fitting model	**R^2**	**R^2_{adj}**	**RMSE**	**MAE**	**RMSE(%)**	**MAE(%)**
$Y = 3.55NDVI_5 + 62.6$	0.03	0.01	7.39	5.68	9.43	12.26
$Y = 51.98NDVI_{10} + 27.98$	0.04	0.02	7.33	5.62	9.32	12.16
$Y = 56.15NDVI_{15} + 22.2$	0.04	0.03	7.25	5.51	9.14	12.03
$Y = 59.45NDVI_{20} + 18.02$	0.05	0.03	7.19	5.43	9.01	11.93
$Y = 66.17NDVI_{25} + 12.07$	0.06	0.04	7.15	5.41	8.98	11.87
$Y = 79.41NDVI_{30} + 1.7$	0.08	0.06	7.07	5.45	9.04	11.73
$Y = 102.41NDVI_{35} - 15.7$	0.12	0.11	6.93	5.46	9.06	11.50
$Y = 136.27NDVI_{40} - 40.93$	0.19	0.17	6.77	5.29	8.78	11.24
$Y = 174.66NDVI_{45} - 69.1$	0.27	0.26	6.73	5.40	8.97	11.17
$Y = 200.83NDVI_{50} - 87.55$	0.35	0.33	6.70	5.54	9.19	11.12
$Y = 199.82NDVI_{55} -85.3$	0.38	0.37	6.35	5.29	8.78	10.54
$Y = 174.15NDVI_{60} -64.71$	0.39	0.38	5.81	4.91	8.15	9.64
$Y =138.71NDVI_{65} -37.2$	0.37	0.36	5.49	4.75	7.88	9.11
$Y = 107.56NDVI_{70}-13.21$	0.36	0.35	5.30	4.58	7.60	8.80
$Y = 86.66NDVI_{75} + 3.19$	0.37	0.36	5.12	4.35	7.22	8.50
$Y = 75.16NDVI_{80} + 12.98$	0.40	0.39	5.00	4.13	6.85	8.30
$Y = 70.38NDVI8_5 + 18.45$	0.44	0.43	4.97	4.10	6.80	8.25
$Y = 69.6NDVI_{90} + 21.71$	0.48	0.47	5.00	4.05	6.78	8.36
$Y = 70.52NDVI_{95} + 24.29$	0.49	0.48	5.10	4.15	6.89	8.46
$Y = 72.71NDVI_{100} + 26.42$	0.48	0.47	5.32	4.35	7.23	8.83
$Y = 77.2NDVI_{105} + 27.62$	0.43	0.43	5.63	4.57	7.58	9.34
$Y = 83.92NDVI_{110} + 28$	0.36	0.35	6.00	4.77	7.92	9.96
$Y = 89.57NDVI_{115} + 28.74$	0.28	0.27	6.40	4.94	8.19	10.62
$Y = 94.65NDVI_{120} + 29.38$	0.20	0.18	6.76	5.14	8.53	11.22
$Y = 97.66NDVI_{125} + 30.33$	0.11	0.10	7.01	5.18	8.59	11.63
$Y = 90.05NDVI_{130} + 34.05$	0.03	0. 02	7.04	5.18	8.59	11.68
$Y = 59.57NDVI_{135} + 43.73$	0.00	0.02	6.98	5.31	8.81	11.58
$Y = 5.62NDVI_{140} + 59.11$	0.02	0.00	7.11	5.62	9.33	11.80
$Y = 49.11NDVI_{145} + 73.9$	0.06	0.05	7.36	5.89	9.77	12.21
$Y = 84.36NDVI_{150} + 83.01$	0.10	0.08	7.54	6.03	10.00	12.51

Appendix C: Yield maps

Figure 32: Predicted yield (ton/ha) maps for potato, based on linear regression models using NDVI and meteorological values 85 days after greenup

Figure 33: Predicted yield (ton/ha) maps for potato, based on linear regression models using the integrated NDVI

Appendix D: Spatial distribution of fields with cloud free images

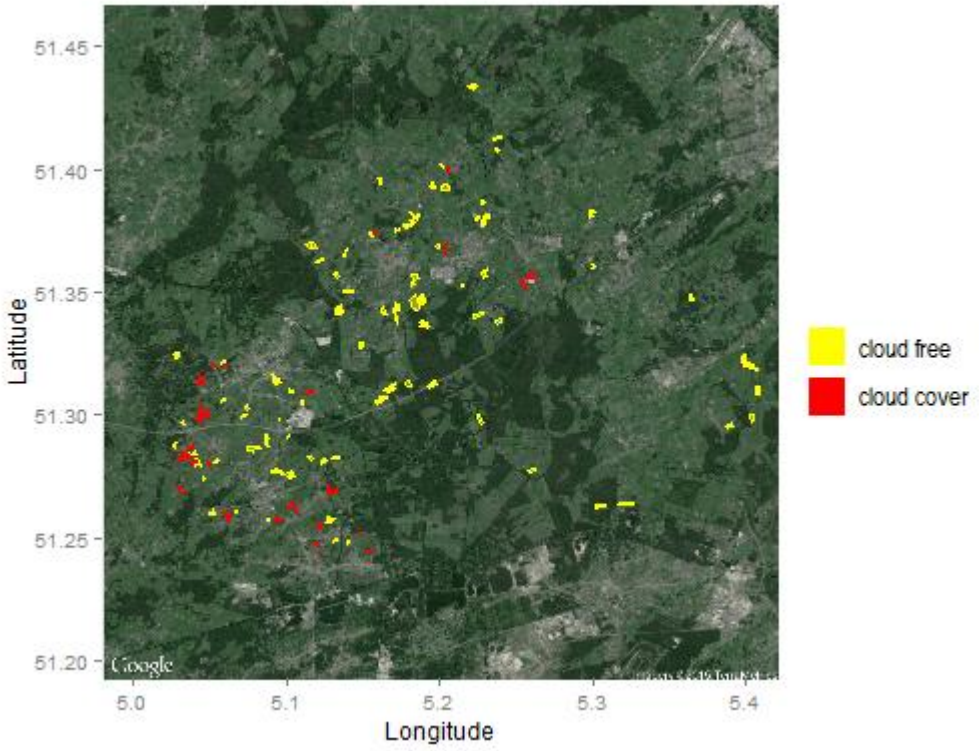


Figure 34: Based on the availability of cloud free images during the growing period fields were illustrated with yellow have from 1 to 4 images into the crucial points, while fields with re have no observations