UNIVERSITY OF COPENHAGEN
DEPARTMENT OF GEOSCIENCES AND NATURAL RESOURCE MANAGEMENT

Assessing the ability of Sentinel-2 derived vegetation indices to explain inter-field yield variation in the context of index insurance

– A case study of paddy rice in Haryana and Odisha, India

Master thesis

Nicolai Bloch Pedersen & Kristian Wille Østergaard

Supervisor: Rasmus Fensholt

Submitted on: 02-07-2020

Name of department:	Department of Geosciences and Natural Resource Management
Authors:	Nicolai Bloch Pedersen (qzv877) & Kristian Wille Østergaard (lmw255)
Title and subtitle:	Assessing the ability of Sentinel-2 derived vegetation indices to explain inter-field yield variation in the context of index insurance – A case study of paddy rice in Haryana and Odisha, India
Topic description:	Based on two study sites in India, Haryana and Odisha, this study aims to contribute to the technical aspect of improving the indices for index insurance, more specifically on how field level yield can be estimated through Sentinel-2 derived VI variables and which design options are more suitable to create these variables.
Supervisor:	Rasmus Fensholt
Submitted on:	02-07-2020
Study points:	30 ECTS
Pages:	~65 (of 2400 characters)

Table of contents

Abstract

Smallholder agriculture in the Global South is characterised by high degree of risk, which disincentivises investment in productivity gains and limits rural development. Index Insurance aims to overcome the limitations of traditional insurance to insurance farmers against exposure to climatic extremes. Based on two study sites in India, Haryana and Odisha, this study contributes to the technical aspect of improving the indices, more specifically on how field level yield can be estimated through Sentinel-2 derived VI variables and which design options are more suitable to create these variables. The study shows that the best variables alone can explain 20% of the inter-field grain yield variation and that the best combination of variables can explain 53%. Furthermore, the main findings of the study suggest that it is beneficial to test different triggering measures and that including variables from phenologically tailored phases and isolating the rice varieties significantly improves the results. Additional research is needed before the approach is suitable for individualised index insurance but compared to alternative data sources the method will likely pose an effective and scalable way to identify yield gaps and to specifically target policy interventions.

Acknowledgements

This study was made possible by the yield data shared by the International Food Policy Research Institute (IFPRI). Berber Kramer, who has been our contact person in IFPRI, has in addition to the providing data also contributed with guidance regarding the objectives of the study and the data management. For this we are grateful.

Appreciation for Rasmus Fensholt, our teacher and supervisor, who has endured us through many years, always ready to help. Thank you.

This project has benefited significantly from the expertise and insights on Index Insurance gained though conversation with Elinor Benani, working in the UC Davis Department of Agriculture and Resource Economics, and Daniel Edward Osgood working in the International Research Institute for Climate and Society, thank you.

Thanks to Daniel Alexander Rudd for sharing his knowledge about Google Earth Engine.

We will end on a quote from Yahya Hassan, a young Danish poet who passed away this year:

"*Det er ikke fordi, jeg med disse udtalelser tror, at jeg har opfundet den dybe tallerken. Jeg har bare fyldt den med suppe, og nu skal der altså smages på den.*" - Yahya Hassan, 2015 – RIP

List of abbreviations

AGB	Above Ground Biomass
AOI	Area of Interest
CCE	Crop Cut Exercises
CVS	Cross-Validation Score
DL	Double Logistic smoothing
DOY	Day of Year
DY	Dynamic
EVI	Enhanced Vegetation Index
EOS	End of Season
FI	Fixed
FRS	Flowering and Reproductive Stage
GEE	Google Earth Engine
GCVI	Green Chlorophyll Vegetation Index
IFPRI	International Food Policy Research Institute
LAI	Leaf Area Index
MWLR	Moving Window Linear Regression
NDVI	Normalized Difference Vegetation Index
NIR	Near-Infrared
RF	Random Forest
RFECV	Recursive Feature Elimination Cross-Validation
R^2	Coefficient of determination – R squared
RMS	The Ripening and Maturity Stage
SOS	Start of Season
TIR	Thermal Infrared
UAI	Unit Area of Insurance
VI	Vegetation Indices
VS	Vegetation Stage
WS	Whole Season

Reading guide

The thesis will begin with a glossary, giving the reader an introduction to the concepts that will be used in the study. The motivation behind the topic of index insurance will then be described, followed by an exploration of the existing research field. How this study is placed in the research field will then explained, followed by the composition of the research, culminating in the specific research question and the aim of the study.

Theoretical knowledge of rice phenology will then be presented and the two the study sites and the used data will be introduced. This will be followed by some considerations of the scientific methodology applied in the study. The method of the analyses will then be presented, followed by a presentation of the results. These will then be evaluated in the discussion, along with a comparison of the results to similar studies and a discussion about the implications of the study for the research field. The discussion will result in a list of recommendations for further research of the topic and lastly the main findings of the study will be concluded.

Glossary

This list of concepts will be useful to be familiarised with, when reading this study.

Inclusive insurance: Inclusive insurance is a school of insurance that focuses on affordable and fair insurance products, providing insurance to lower-income population segments, typically in the Global South (Cheston, 2018).

Adverse selection: Adverse selection describes a situation where the information between a buyer and seller is unequal. In insurance, adverse selection entails a higher demand for insurance from farmers that know they are more at risk. The insurance companies then need to adjust for this when assessing their exposure and determining the premiums (Investopedia, 2020).

Morale hazard: Morale hazard describes how being insured can change the behaviour of the insured towards more risk-taking behaviour. In agriculture, insurance can lead farmers towards practices that are more likely to suffer losses (BD, 2020; IRMI, 2020; Greatrix et al., 2015).

Index insurance: Index insurance is a relatively new approach for insurance where the insurance pay-outs are determined by an objective index, and thus detached from the experienced losses. The index could for example be based on precipitation, vegetation indices, wind, or temperature. For the insurance to be reliable, the indices should be closely related to the agricultural production losses (GIFF, 2020; Greatrix et al., 2015).

Basis risk: Basis risk is a term in index insurance, that describes the risk of a mismatch between the experienced loss and the insurance pay-out. It is an inherent challenge in index insurance as the index is decoupled from the experienced losses (GIFF, 2020; Greatrix et al., 2015).

Design options: In this context, design options cover the different possibilities when designing an index for index insurance. When creating an index, it must be decided which variable to base the index on, how that variable is aggregated to one value, which period is covered and how the start of the period is determined. The design will often depend on the objectives of the index (IFAD, 2017).

Input-based variables: In the context of index insurance, input-based variables are variables that directly influence the crop production, such as precipitation and temperature. Index insurances based on input-based variables builds on the assumption that the variables are drivers of crop changes and therefore a good indicator for the yield (IFAD, 2017).

Output-based variables: Output-based variables are in this context, variables that can be used as proxies for the yield. The variables do not influence the crop growth but are merely estimations of the changes in vegetation productivity. Vegetation indices are an example of an output-based variable (IFAD, 2017).

Smoothing type: Smoothing the data is an important tool working with satellite dataset. The reflection from the vegetation is frequently altered or blocked, typically due to aerosols, clouds or changing illumination patterns. This produces noise in the dataset. In order to address this, the raw dataset is processed through a smoothing technique that produces a more representative data set (USGS, 2020).

Vegetation indices: Vegetation indices (VIs) are an indicator of the 'greenness' of the vegetation, derived from its reflective properties. Vegetation indices are created from the values of specific spectral bands, combined in different mathematical formulas. In this study two vegetation indices are used: Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) (Lillesand et al., 2015; Pasimeni et. al., 2019).

VI variables: In the study there will occasionally be referred to VI variables. This will refer to the variables created on the bases of vegetation indices from satellite data.

1 Introduction

1.1 Motivation

Smallholder agriculture is an important foundation for employment and food security in many countries in the Global South. It is however characterised by a high degree of risk, especially from exposure to climate variability and adverse weather events. Events, such as droughts and floods can force farmers to utilise short-term strategies to cope with the immediate crisis. These responses can however reduce the development of the farmers´ livelihood in the long term, as they often involve deterioration of productive assets. Large investments in smallholder agriculture are necessary to accelerate rural development and to meet the increasing food demand from growing populations. However, the vulnerability context of the agricultural sector disincentivise investments in production gains, keeping people in many rural areas in the Global South trapped in a state of food insecurity and persistent poverty (Carter et al., 2017; Hansen et al., 2017; IFAD, 2017; Miranda & Farrin, 2012). Climate change which in many places is expected to increase the frequency and severity of extreme events, will further strengthen this poverty trap (IPCC, 2018; GIZ, 2016).

Increasing agricultural productivity and empowering smallholder farmers is widely considered an effective way to increase resilience and reduce poverty and hunger (Ivanic & Martin, 2018; Lobell et al., 2018; UN, 2019). Inclusive insurance can contribute to this by providing a safety net, preventing farmers from falling into poverty after a shock, and by making it more attractive for investments. Agricultural insurance is however not well developed in the Global South, and traditional indemnity insurance suffers from several challenges, such as lack of trust between insurance provider and policyholder and proportionally high verification cost. Index insurance has been proposed as a solution to these challenges. By basing the claim and verification process on an objective, automated index, the verification costs and the mistrust can be reduced (GIZ, 2017; Miranda & Farrin, 2012; Greatrex et al. 2015; Carter et al. 2017; Platteau et al. 2017; Liu & Myers, 2016; The World Bank Group, 2018; GIIF, 2019).

1.2 Research field

1.2.1 Challenges for index insurance

There are several challenges that needs to be overcome to increase the scalability and socioeconomic impact of index insurance in the Global South. These challenges can be grouped

in two; delivery challenges and technical challenges. The delivery challenges are related to how index insurance policies can be designed to add the most value for the policyholders, how the products can be scaled, keeping the cost low and how demand can be increased by raising awareness and building trust. The technical challenges are related to how the specific indices can be created to increase the agreement between the experienced loss by the farmer and the detected loss by the index i.e. how the basis risk can be reduced (IFAD, 2017; GIZ, 2016; Carter et al., 2017; Greatrex et al. 2015). The focus of this study will be on the latter, more specifically on how satellite data can contribute to overcome the technical challenges and increase the index accuracy.

The research field of technical challenges is quite diverse, and research varies in terms of the overall objective for the index, the variables used, the scale, the specific design options and the resolution of the used satellite products (IFAD, 2017). To place this study in the research field it can be useful to categorise the diverse field in two schools of research. The boundaries are however unrestrictive, and a range will likely be more beneficial understanding of the field than categories, as many studies fall somewhere between the two.

1.2.2 First school of the technical challenges

The first school is characterised by a goal to insure against the most severe events, caused generally by a single peril. The essential aspect is whether the index is able to accurately capture which years that has been the worst, as experienced by the farmers. It is less important to be able to estimate the precise yield each year. The index is typically based on input-based measures, such as rainfall or soil moisture and will often only insure against a single weather-related peril. If a farmer is hit by a pest attack, it will not be captured by the index nor compensated from the insurance. The Unit Area of Insurance (UAI) in which all policyholders are assumed to be similar, are typically large (sub-county to district), also as a consequence of low-resolution data sources. The large UAI are a source of basis risk, as specific local conditions are averaged out. A farmer who has been victim of a flood might not be compensated if the majority of the other farmers in the region are unaffected by the shock (IFAD, 2017; Rosema et al., 2014; Enenkel et al., 2018; Osgood, et al., 2018). Studies which focus on output-based measures, such as NDVI, can also be considered part of this school when the size of the UAI´s are large (sub-district to region). Due to the large UAI´s the insurance will not typically cover

multiple perils, as individual losses from e.g. pest attack will be lost when focusing on the average of a large area, even though the loss could have been detected in the VI at the farmer´s field. (Klisch & Atzberger, 2016; Flatnes et al., 2019; Chantarat et al., 2013; Makaudze & Miranda, 2010; Son et al., 2013).

1.2.3 Second school of the technical challenges

Recognising the heterogeneity of crop losses, the aim of the research in the second school is to create indices that are able to insure a village or even individual farmers against multiple perils on a seasonal basis i.e. to be able to estimate the yield of the farmers after each season and compensate if the yield is less than average. For this, high-resolution data is necessary in order to differentiate between villages or single fields. Output-based data, such as vegetation indices are often used in this school of research. As these indices directly reflects the crop growth, they will typically insure against multiple perils, ranging from pest and diseases to the different climatic conditions. If the accuracy of the indices is adequate, basis risk will be less severe when the index is individualised. But this specified insurance approach increases the risk of morale hazard and adverse selection, two inherent challenge of insurance. These challenges impose high demands on the overall design of the insurance policies. Several innovative solutions are being developed, but that is a topic belonging in the delivery challenges category and will not be further elaborated in this study (Burke & Lobell, 2017; Hufkens et al., 2019; Lobell et al., 2019; Lambert et al., 2017; Jain et al., 2016; Azzari et al., 2017; Guan et al., 2018).

Index insurance products are not the only application for field level yield estimations obtained through satellite images and therefore not the only objective of the research in this school. Accurate estimates of farm level yield can also be used to identify productivity gaps, enabling specific targeting of policy interventions, such as fertilizer and seed supply or access to microcredit. As such intervention are often otherwise implemented as one-size-fits-all, the farm level yield estimates will likely increase the effectiveness of such interventions. Furthermore, the estimates can be used to assess the results of implemented initiatives. These initiatives can all contribute to increased agricultural production and thus rural development and food security. The estimates can also be used as verification data for when developing insurance products on a larger scale, as an alternative to other sources of yield data that are often

expensive and unreliable (Lambert et al. 2017; Hufkens et al., 2019; Guan et al., 2018; Burke & Lobell, 2017; Lobell et al., 2018; Lambert et al., 2018).

1.3 Focus of the study

This study is placed in the second school of the research field and aims to contribute to how high-resolution satellite data can be utilised to accurately estimate farm level yield and to ultimately, develop effective insurance mechanisms for farmers in the Global South.

To create an index insurance, a prerequisite is that farm level yield can be approximated by objective data. If a reliable relationship between farm level yield and a satellite derived VI variable can be established, this can be used as a foundation for the index insurance. The VI variable would then be referred to as the index. Insurance pay-outs for a specific farmer would then be dependent on the VI value observed by the satellite. Pay-outs would be made if the index estimates a poor harvest for the specific farmer, irrespective to the farmers experienced loss. The exact threshold and price for the insurance policy would typically be determined by historical data. In this study it is however only the relation between yield and the index that will be in focus (Miranda & Farrin, 2012; GIIF, 2019; Greatrex et al. 2015).

The study will not make yield estimations as such but only create and assess VI variables´ ability to explain yield variation. How well the VI variables can explain the inter-field yield variation is a direct measure of well they can estimate farm level yield, and the two terms will therefore be used interchangeably throughout the study.

1.4 Research composition

The aim of the study is to assess the ability of vegetation indices to explain inter-field yield variation in paddy rice for two study sites in India and to assess which VI design options are most suitable for this.

Numerous VI variables will be created through processing of satellite images using different combinations of the selected design options. These will then be systematically assessed.

Three statistical analyses will used to test the VI variables: A linear regression, returning the correlation between the individual variables and the yield. A multiple regression, providing the

correlation between the yield and multiple explanatory VI variables[1], and lastly a Random Forest classification yielding an accuracy assessment of how well groups of VI variables were able to classify the samples according to their yield. Assessing the design options across three statistical analyses is expected to increase the robustness of the results. This approach can be referred to as methodological triangulation, which has been found to be beneficial when working with comprehensive data sources (Bekhet & Zauszniewski, 2012). The multiple regression and RF classification both run on groups of variables. It can therefore also be tested how well different VI variables can supplement each other and together explain the yield variation. The two tests also make it possible to correct for biases in the data, by including suspected bias-creating variables.

Subsequently, it will be analysed which aspects create the most uncertainty in the results: The smoothing types and VIs will be compared, the effect of correcting for biases will be assessed and the amount of uncertainty from the mismatch between above ground biomass and grain yield will be evaluated.

Before analysing the output-based VI variables, it will first be assessed which independent variables that affect the yield. This will be done with a similarly method, using linear regression, multiple regression and RF classification to assess the ability of the variables to estimate yield, but using input-based variables, such as climatic, socioeconomic and spatial variables instead of output-based. The aim of this is to get a better understanding of the factors determining the farm-level yield.

[1] In this study, the VI variables will frequently be referred to as explanatory or independent variables, as they are used to explain the variation in yield. This does however not imply a causal relation as it is the VIs that are dependent on the yield and not the other way around. When using the input-based variables to explain yield variation, a causality is expected and the yield is therefore also the dependent variable in reality.

1.4.1 Research questions

The composition of the study can be summarised in following research questions, on which this thesis will be based:

How much of the inter-field yield variation can be explained by Sentinel-2 derived vegetation indices and which design options give the best results?

1. Which input-based variables influence the farm level yield?

2. How well does the VI variables explain the inter-field yield variation?

3. Which design options results in the VI variables most suitable to estimate yield?
   3.1. Vegetation index: NDVI or EVI
   3.2. Smoothing type: MWLR or DL
   3.3. Triggering measure: Peak, integral, mean, length, SoS or EoS
   3.4. The period: Phenologically tailored phases or only for the whole season
   3.5. Seasonality: Dynamic or fixed seasonality

4. What aspects of the index creating process create the largest sources of uncertainty?
   4.1. The smoothing type or VI choice
   4.2. The bias creating variables
   4.3. The imperfect correlation between grain yield and total yield

1.5 Ambition of the study

The recent launch of the Sentinel-2 satellite offers new possibilities for yield estimations of smallholder fields. The data provides sufficiently high spatial resolution to distinguish between individual fields, and the temporal resolution makes it possible to extract parameters representative of the entire crop season thus providing more information about how the crops have developed (Lambert et al. 2017; Lambert et al., 2018; Jain et al., 2016).

Reliable ground truth data is often mentioned in the literature as a limiting factor for creating indices and assessing their accuracy. Farmer surveys of historical yield can be very uncertain

and systematically biased. When used to create and assess index products it can be difficult to determine how much of the error is due to inaccurate indices and how much is from inaccurate yield data. The very comprehensive yield dataset used in this study is made from crop cuts exercises (CCEs) of over 500 fields. CCEs have in previous studies been shown to be more reliable and basing the analyses on this data therefore allows for higher confidence when assessing the indices and specific design options (IFAD, 2017; Jain et al., 2016; Lobell et al., 2018; Lobell et al., 2019; Guan et al., 2018).

The yield dataset used in this study also provides information about the phenological stage of the crops at different times during the season. This provides the unique opportunity to, based on empirical information, divide the crop season according to different phenological stages and to access the effect of this specified information on the yield estimations.

We are not aware of other studies that systematically assess the design options across multiple variables and statistical tests. In addition, comparable studies generally only include one triggering measure for the whole season. The benefit of combining different VI variables and of including variables from phenologically tailored phases is therefore not well explored in the literature.

In the absence of a substantial theoretical body, this study aims to contribute to this important and emerging area of research with novel insights on how the VI variables can be designed to increase the accuracy of farm level yield estimations.

2 Theory

Having covered the essential terms and concepts of index insurance in the glossary and introduction, the theoretical section will be short, only providing an introduction to the phenology of paddy rice.

2.1 Rice phenology

A thorough understanding of the rice phenology is essential to this study as an important aim is to create variables specifically designed to capture the condition of the vegetation in certain crop phases. The developing stages of the paddy rice crop will therefore be presented here.

There are several ways to classify the rice crop stages. For this study the most simple, consisting of three phases, is sufficient. The three phases are:

(1) The Vegetation Stage (VS), which has a duration of 50-100 days, is characterized by the formation of shoots and leaf development, and hereby an increase; in height, tillers and leaf area. This initiates a gradual increase in above ground biomass (AGB) as seen in Figure 1 (Guan et al., 2018; Dong & Xiao, 2016).

(2) The Flowering and Reproductive Stage (FRS) has a duration of 30-35 days. In this stage the growth of the reproductive parts is initiated and the AGB continue its gradual increases (Figure 1). At this stage the rice experiences the fastest plant height increase, booting initiates with panicle production and flowering begins. The FRS is a significant phase for the rice production, as the formation of flower buds determines the number of grains, a decisive measure for the grain yield (Guan et al., 2018; Dong & Xiao, 2016).

(3) The Ripening and Maturity Stage (RMS) has a duration of 30-35 days in which the AGB reaches its maximum (Figure 1). When entering this stage, the number of grains is fixed equivalent to the number of flowers produced in the FRS. In the RMS, the filling of the grains begins, which leads to an increase in grain size and weight. The RMS ends as the leaves and the grains have gradually turned golden yellow and the rice is ready for harvest (Guan et al., 2018; Dong & Xiao, 2016; Hufkens et al., 2019; CGIAR, 2013).

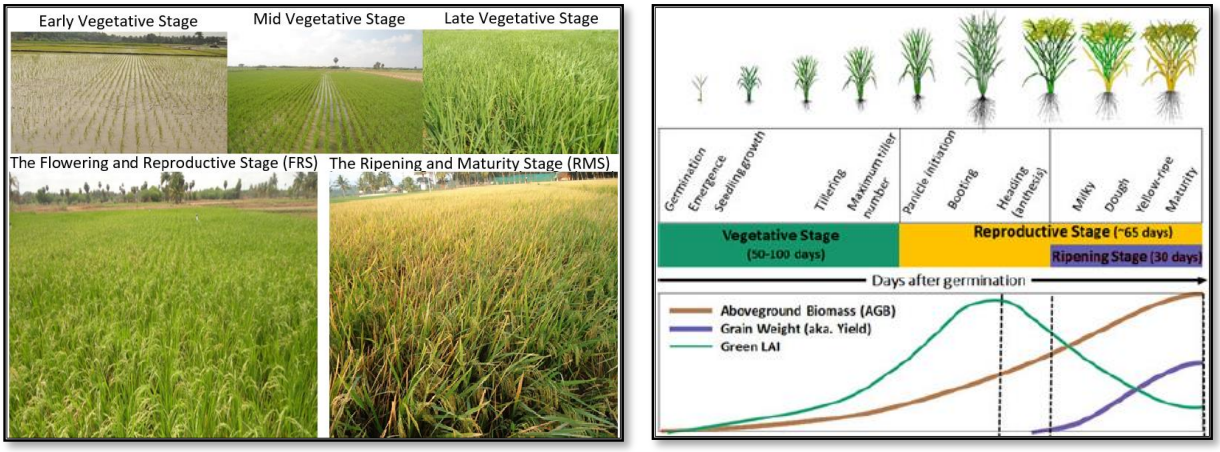


Figure 1: Overview of paddy rice crop stages (Guan et al., 2018; Dong & Xiao, 2016).

3 Study sites

This study is done for two sites in India: Haryana in North West India and Odisha in East India (Figure 2). Using two study sites gives the possibility to compare the results and assess the impact of aspects that differ between the sites. It is also the expectation that the results will be more robust when assessed over two study sites.

Figure 2: Map of study sites.

Haryana has a semi-arid climate with high temperatures and a condensed precipitation period. Odisha has a tropical savanna climate. It generally receives more precipitation, distributed over a longer period and the annual temperatures are more stable (Figure 3).

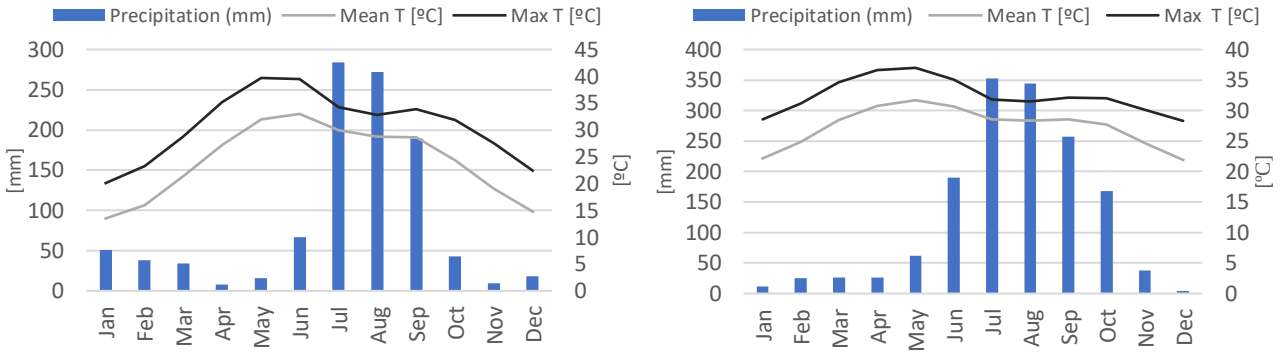


Figure 3: Climographs from the two study sites. ***Left****: Haryana (CLIMATE-DATA.ORG, 2020a).* ***Right****: Odisha (CLIMATE-DATA.ORG, 2020b)*

The annual per capita income of the around 25 million people living in Haryana is 236 thousand rupees. In Odisha, which is populated by almost 44 million people, the annual per capita income is 96 thousand rupees (***Table 1***) (GOH, 2020; GOO, 2020; Statista, 2020a; Statista, 2020b).

Table 1*: Information about the two study sites (GOH, 2020; GOO, 2020; Statista, 2020a; Statista, 2020b).*

	Haryana	***Odisha***
Location	North West India (30° 43′ 48″ N, 76° 46′ 48″ E)	East India (20° 17' 46'' N, 85° 49' 28 '' E)
Population	25,35 million	43,73 million
Climate	Hot semi-arid climate	Tropical savanna climate
Per capita income	236 thousand Indian rupees in financial year 2019	96 thousand Indian rupees in the financial year 2019

The participating farmers have access to irrigation and are mainly practicing conventional paddy rice production for selling and exporting. The average field size is around 0.20 Ha (2032 m2) for Haryana, and 0.05 Ha (465m2) in Odisha[2]. The average grain yield in 2019 was 24.6

[2] According to the manually drawn polygons around the fields with yield data. The method of this will later be described.

quintals per acre (6.08 t/Ha) in Haryana and 17.2 quintals per acre (4.25t/Ha) in Odisha. For both study sites, the variance was however rather high (Figure 4).

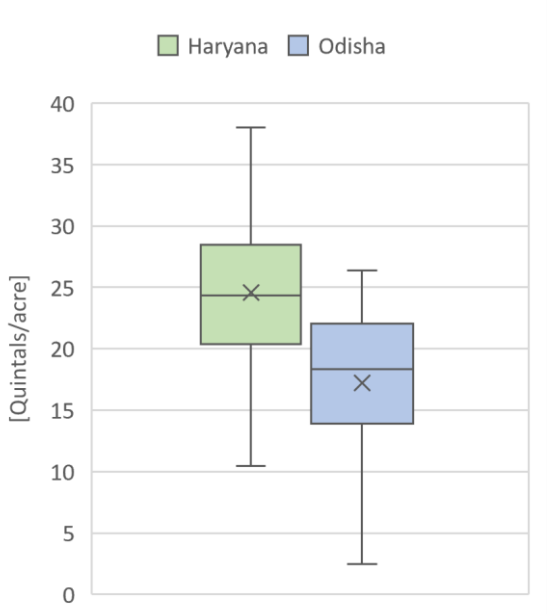


Figure 4: Distribution of grain yield for fields in Haryana and Odisha

An overview of the soil types and rice varieties for the two study sites can be found in the appendix (Figure 50 & Figure 51).

4 Data

The data sources used in this study will be introduced here, starting with an overview of what is included in the yield data provided by IFPRI. This is followed by a presentation of the specifications of the three sources of satellite data.

4.1 Yield data

4.1.1 Data from IFPRI

"*This data was provided by IFPRI. IFPRI bears no responsibility for the analyses or interpretations of the data presented here*"

Farm level yield data from the Indian states, Haryana and Odisha was made available to us by Berber Kramer from IFPRI. The dataset contains information from field surveys of the late 2019 rice harvest and from smartphone images of the fields during the entire length of the late 2019 crop season.

The dataset contains geo-localised information from CCEs of 317 fields in Haryana and 105 fields in Odisha (Figure 5). Of these fields, the farmers in Haryana and Odisha had on instruction taken 766 and 718 smartphone pictures of the fields during the crop season. A manual classification of the vegetative stage of crops in each smartphone picture was also included in the dataset.

The survey data includes: Grain yield (standardised at 14% moisture level) and biomass yield, collected in $9m^2$ or $25m^2$ CCE. (See appendix for further details about the CCE process: Figure 54). A varying subset of the fields also had information on rice variety, soil type, money spent since last image, farmer reported cause of damage, observed cause of damage (on smartphone pictures) and farmer reported input use.

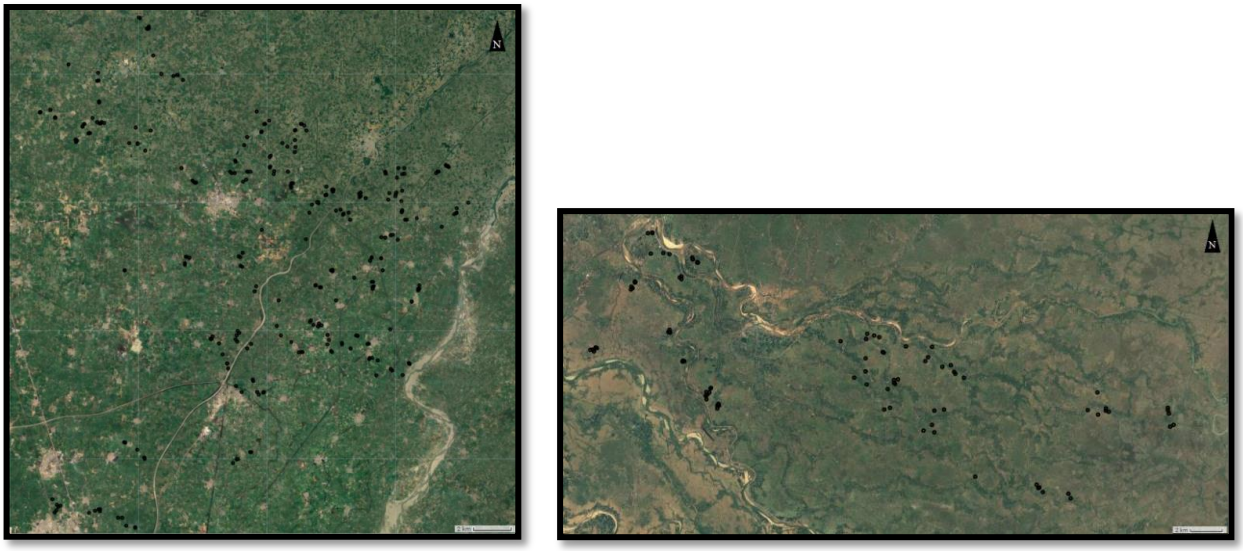


*Figure 5: The spatial distribution of the yield data for Haryana (**Left**) and Odisha (**Right**).*

4.2 Satellite data

4.2.1 Sentinel-2 MSI: MultiSpectral Instrument, Level-2A

Sentinel-2 is a part of the Copernicus mission, which comprises a constellation of two twin Sentinel-2 satellites. Sentinel-2A was launched and set in orbit in June 2015 and Sentinel-2B in March 2017, both at an altitude of 786km (ESA, 2020a). They operate simultaneously, with a 180° angle in a sun-synchronous orbit, which guarantees consistency of the illumination direction. This minimises the potential impact of differing shadows and ground illumination levels and is thus a vital feature when assessing time-series of images (ESA, 2015). The combination of the two satellites generates a revisit frequency of 5 days at the Equator i.e. a temporal resolution of 5 days (ESA, 2015). The Sentinel-2 produces 13 different spectral bands at different spatial resolutions. This study uses 3 bands to calculate the VIs, the Blue (B2), the Red (B4) and the Near-Infrared (B8), all with a spatial resolution of 10 meters.

The Sentinel-2 product used is "Sentinel-2 MSI: MultiSpectral Instrument, Level-2A". The Level-2A data has been pre-processed, ensuring that the images are ortho-rectified and bottom of the atmosphere reflectance (ESA, 2020b).

4.2.2 MOD11A2.006 Terra Land Surface Temperature and Emissivity 8-Day Global 1km

This dataset is derived from the Terra satellite which was launched in 1999 and is part of the collection "Terra MODIS" (NASA, 2020). It provides land surface temperature and emissivity

with a spatial resolution of 1km and a temporal resolution of 8 days, averaged from the daily MOD11A1 values within the 8-day period (USGS, 2020).

4.2.3 CHIRPS Daily: Climate Hazards Group InfraRed Precipitation with Station Data

CHIRPS Daily was created in 1981 to produce rainfall maps, specifically in areas where surface data is limited. Like the National Oceanic and Atmospheric Administration's (NOAA's) rainfall predictions, it builds on approaches using thermal infrared (TIR) reflectance to estimate precipitation (USAID, 2020). It provides a daily global precipitation dataset, with a 0.05° spatial resolution (Funk et. al, 2015).

5 Scientific methodology

5.1 A quantitative case study

This project is a quantitative case study. A set of research questions creates the foundation for a quantitative analysis. The outcome of the analyses is assessed, and conclusions are drawn from the results. The method is then evaluated and recommendations for further research are made.

A thorough understanding of the research field gained from recent scientific literature and through conversations with field experts, allowed us to identify very specific and unexhausted research questions that could potentially contribute with new knowledge to the field.

Several modifications to the research design were made in the research process to improve the quality of the results.

5.2 Choice of software

The study has been carried out using Google Earth Engine (GEE). This ensures high transparency as the scripts contains all information of the analysis. While creating the scripts for the analysis we simultaneously create a pipeline for replicating the analysis for other study sites, as the scripts only need few adjustments to work at other locations. (Azzari et al., 2017; Lobell et al., 2015).

The strong processing power and replicability of GEE allowed us to create many different VI variables for the two study sites. This made it possible to evaluation the design options across multiple variables, thus increasing the robustness of the results. Examples of the GEE scripts can be found in the appendix (11.11 GEE script: Data preparation and the temporal aggregation p.118 & 11.12 GEE script: Spatial aggregation p. 134).

6 Methods

In this section, the methods of the study will be presented: It will be explained how the yield data has been processed and how the VI variables were created from the satellite data by calculating the VIs, smoothing the timeseries and then temporally and spatially aggregating the timeseries. It will then be presented how the created variables were evaluated in the three statistical analyses. The section starts with a short overview of the entire analysis.

The overall process has been to create variables from satellite data using many different design options, assessing the ability of these variables to explain the yield variation through three statistical tests and then systematically assess the results to isolate the suitability of the different design options.

To create the VI variables from the satellite data several steps were needed. First the VIs had to be calculated for the images in the timeseries and the effect of clouds had to be smoothed out. A time period for when to extract the values then had to be defined along with a method to do so. The timeseries then had to be temporally aggregated to a single image. For each step in this process there are several different options i.e. the design options. In this study, two different VIs were created and two different smoothings were applied. Variables were extracted for four different periods, using two different ways of defining the phases and six different temporal aggregations. Images were then created for almost all the possible combinations of the chosen design options to be able to compare each design option against its alternative across multiple different variables.

The prepared yield data points were associated with a field though manual plotting of the fields. All the created images were then spatially aggregated using the field plots to obtain a single value for each field. This result in a list of variables, which have a single VI value associated with each yield data value. The ability of the variables to explain inter-field yield variation either individually or in groups of variables was then assessed through three different analyses; linear regression, multiple regression and RF Classification.

In addition, the three statistical tests were also one for different sources of input-based data to get an understanding of the decisive factors for rice yield. Different supplementary variables from the yield dataset were also utilised to in efforts to correct for biases in the yield values.

Lastly it was assessed which aspects of the process that produces the most uncertainty. A graphical representation of the workflow can be seen below (Figure 6).

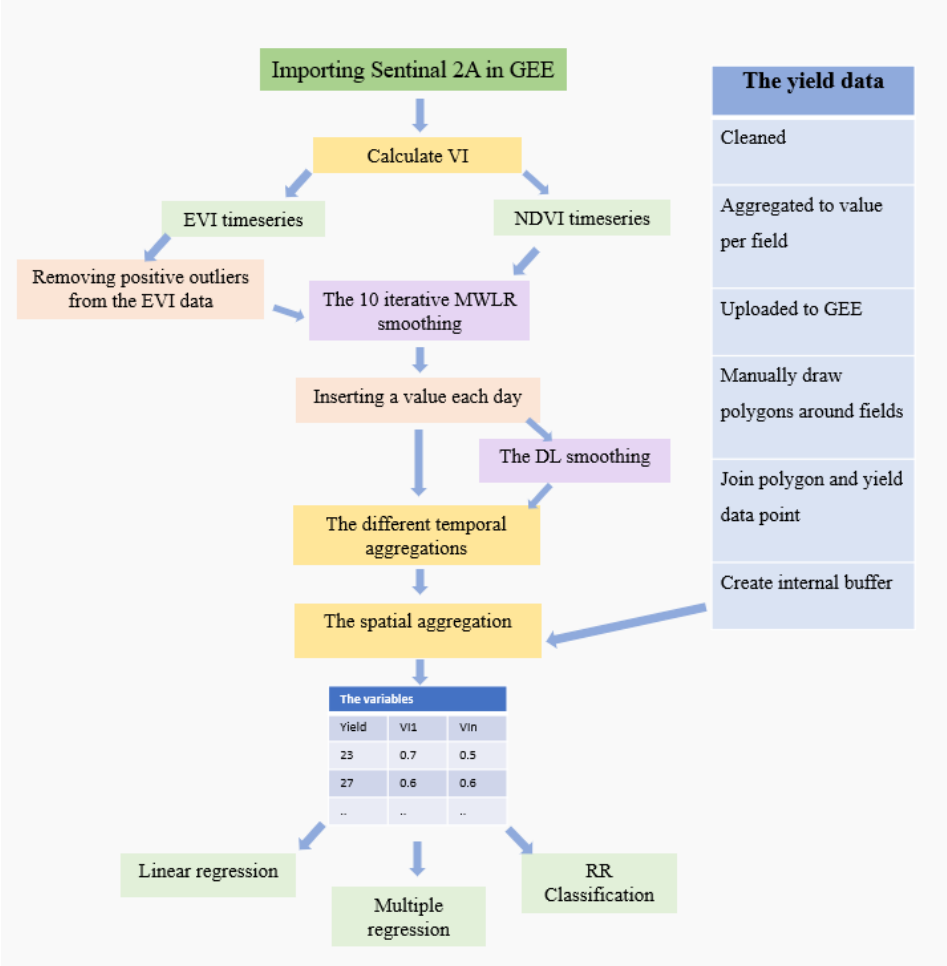


Figure 6: Workflow of the analyses.

6.1 Yield data preparation

A few pre-processing steps were done to prepare the yield data for the analysis.

First the data was cleaned by removing certain troublesome values[3].

[3] Values were deleted if they:
- Had coordinates that places it outside the study-area
- Had a harvest index above 1
- Had a CCE GPS accuracy above 20m (i.e. over 20 meters possible error in the location estimation)
- Were located in Odisha and had grain yield above 35 quintals per acre. On recommendation of the data provider.

Initially, each row of the data represented a picture taken by a farmer. This was converted to a dataset with a row per field, by exploiting that pictures of the same field were listed with an identical CCE yield and therefore could be grouped by a pivot table.

Then, a new variable was added by taking the sum of the grain yield and biomass yield. This represents the total AGB and will henceforth be referred to a "total yield". The total yield is what is measured by the satellite and is therefore expected to correlate better with the VIs. If what the satellite measures (total yield) does not correlate well with the variable we are trying to predict (grain yield), it is an early indicator that it will be difficult to explain the grain yield variation through satellite-derived VIs (Guan et al., 2018).

In Odisha the grain yield correlates well with the total yield (R^2 = 0.74) but in Haryana much less of the grain yield variation can be explained by the total yield (R^2 = 0.22) (Figure 7).

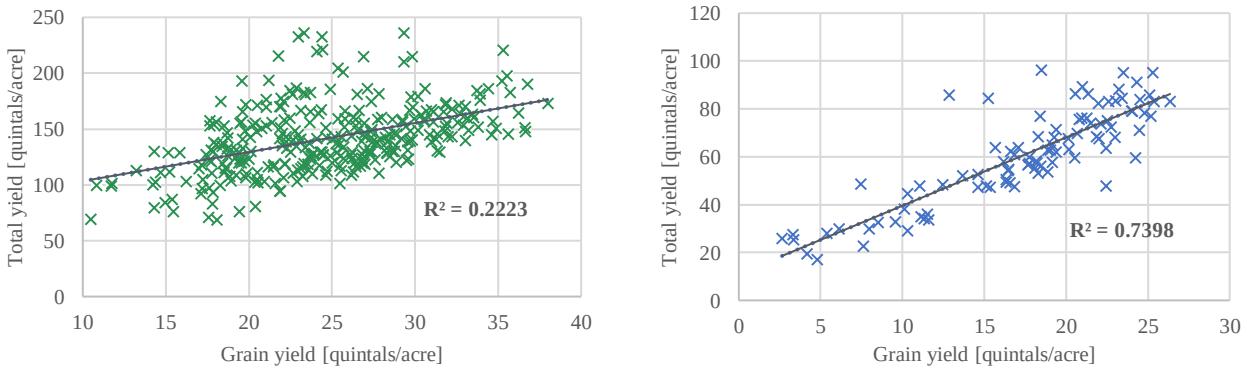


*Figure 7: Relation between total yield and grain yield for Haryana (**Left**) and Odisha (**Right**). Throughout the thesis, the graphs from Haryana will be in green nuances and blue for Odisha.*

The unexpectedly low correlation for Haryana could have multiple explanations. If the critical phase of flowering and reproduction is mistimed it could result in low grain yield but still giving a high total yield. The Haryana samples could also include different groups with differing relation between grain yield and total yield. This could for example be caused by different rice varieties. Lastly, the low correlation could be related to the differing fertilizer application rate.

6.2 Satellite data processing

In this section it will be presented how the VI variables were created from the Sentinel-2 data.

6.2.1 Creating the vegetation indices

From the pre-processed bottom-of-the-atmosphere satellite bands, NDVI and EVI were calculated with the following formulas (F1 & F2) (Lobell et al., 2019; Son et al., 2013).

NDVI:

$$NDVI = \frac{\rho_{NIR} - \rho_{RED}}{\rho_{NIR} + \rho_{RED}} = \frac{B8 - B4}{B8 + B4} \tag{F1}$$

EVI:

$$EVI = 2.5 * \frac{\rho_{NIR} - \rho_{RED}}{\rho_{NIR} + 6 * \rho_{RED} - 7.5 * \rho_{BLUE} + 1} = 2.5 * \frac{B8 - B4}{B8 + 6 * B4 - 7.5 * B2 + 1} \tag{F2}$$

Both indices are frequently used in similar studies (Burke & Lobell, 2017; Guan et al., 2018; Lambert et al. 2017; Lobell et al., 2019; Lobell et al., 2019).

In the preliminary comparisons of yield and NDVI measures, the yield spanned over a large interval while the differences in NDVI where relatively small. A potential explanation for this is that NDVI has a tendency to saturate at high biomass levels. EVI was therefore included as a supplement, as it is less prone to saturation (Son et al., 2013).

6.2.2 Smoothing the time series

On Figure 8, timeseries of NDVI and EVI can be seen. The irregularity of the VI´s over the season is due to cloud contamination. The clouds and cloud shadows consistently result in a lower NDVI value, while they can affect EVI in both directions.

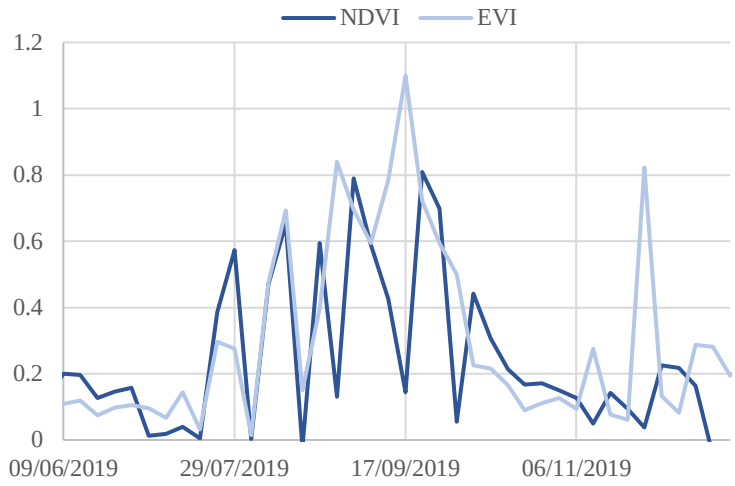


Figure 8: Example of an NDVI and EVI timeseries affected by clouds. The NDVI values is always reduced by clouds, while the EVI can be affected in both directions.

The Sentinel-2 level 2a data contains a band (QA60) with a pre-processed classification of clouds, which is intended to mask out the clouds from the images. However, this band is upon inspection unreliable in these study areas, as there are still cloud contaminated data that is not masked out. Lobell et al. (2019) encountered a similar challenge and as alternative to the cloud masking, they removed the effect of clouds by using an iterative smoothing that fitted to the upper envelope of the data points. Inspired by this, a similar approach was taken here, though with a different smoothing type.

6.2.2.1 Iterative MWLR

While Lobell et al. (2019) used a discrete Fourier transformation, we applied a moving window linear regression (MWLR), with a window size of 12 days i.e. 1 observation on either side, totalling 3 observations. The smoothing type and small size of the window were chosen to get the smoothing as close to the observations as possible. A goal of the analysis is to detect subtle differences between the fields in how the VI has developed at different periods. These might be overlooked if the smoothing has a generic form. Inside the window, a linear regression is made on the three observations and the middle observation is given the value of the trendline at the corresponding time. The window then moves one step and the process is repeated. The smoothing is comparable to a moving average and was satisfactory upon visual inspection.

In order to fit the smoothing to the upper envelope an iterative smoothing process was initiated: After the first smoothing a new timeseries identical to the original data was made but

if the original value was less than the smoothing, it was given the value of the smoothing instead. A new smoothing was then done on this newly created timeseries. This process was, similarly to Lobell et al. (2019) repeated 10 times. With every step the smoothing moves towards the upper envelope, though less and less the closer it gets (Figure 9).

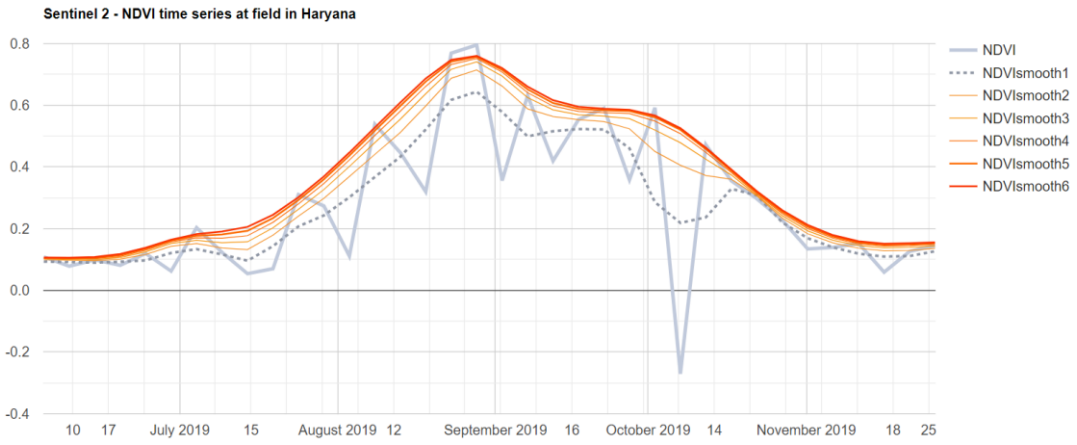


Figure 9: Example of the iterative process of smoothing an NDVI timeseries to the upper envelope of the data. On the graph is only showed 6 of the 10 smoothings that are applied in the study. The grey line represents the raw NDVI data, the dashed line the first MWLR smoothing and the orange and red lines represents the succeeding smoothings.

6.2.2.2 Removing EVI outliers

Fitting to the upper envelope works on NDVI timeseries as all cloud affected values are decreased. Because the EVI values can be influenced in both directions, it is necessary to first remove the positive outliers. This was done by first giving obvious outliers[4] a value equal to the mean of the four closest observations. The remaining outliers were more difficult to identify in the noisy dataset.

In the Timesat, a software specialised in extracting seasonality parameters, there are several ways of removing outliers. The one used for inspiration here, removes values that deviates a certain amount from the mean of the observations in a surrounding window (Eklundh & Jönsson, 2017). Because the timeseries of the AOIs are so heavily affected by clouds, the mean value alone is many places not a good measure to compare a potential outlier against (Figure 8). It was therefore deemed necessary to alter the method for detecting outliers to make it suitable for the specific conditions in the AOIs.

[4] Values below zero and values above one

The new measure was calculated as shown below (F3):

$$M = \frac{\sigma_w - \sigma_{w.o.}}{\sigma_w} * \left| \frac{\mu_w - \mu_{w.o.}}{\mu_w} \right|$$

(F3)

σ is the variation and μ is the mean. w and $w.o$ indicates whether the observation in question is included or not.

How the mean of the 4 surrounding observations compare to the value is still included, but it is multiplied by the fraction of variation that is added when the value in question is included in the window. This way, it is taken into account how likely it is that the observation should have a value close to the mean of the neighbours. If there is high variation in the surrounding observations, it is less likely to register as an outlier, reflecting a reduced certainty in it being an outlier. Calculating the difference in variance and mean in relative terms allows comparability between high and low VI values i.e. to detect outliers in all parts of the season. The cut-off value was determined by visually inspecting numerous timeseries and was set at a value of one. Observations with a measure value above one were then given the mean value of its four neighbours (Figure 10). This process was then repeated to reduce the influence from outliers being part of the four neighbouring observations.

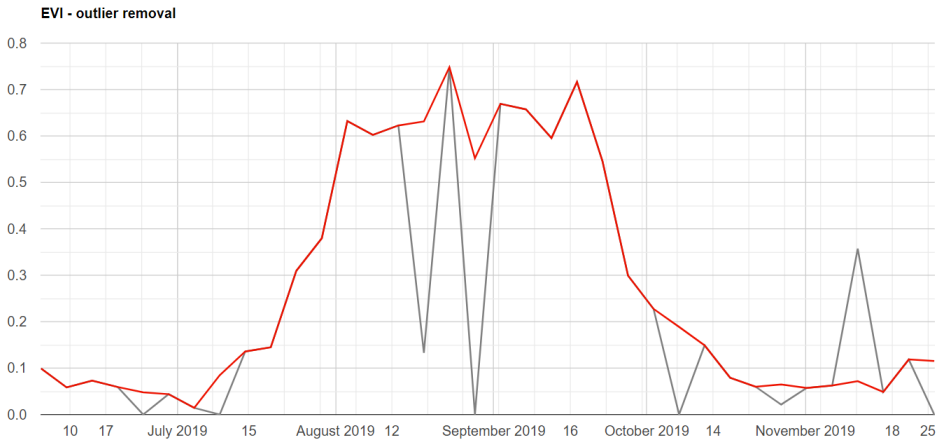


Figure 10: Example of EVI outlier removal. The grey line represents the raw EVI timeseries and the red line is the EVI timeseries after the outlier removal.

Due to the amount of cloudy observations, it was in some instances even challenging to distinguish outliers manually, but upon visual inspection the method appeared to identify the majority of the outliers. At rare occasions a seemingly correct observation was surrounded by

four observations, negatively affected by clouds, with similar values as each other and the value was therefore identified as an outlier. This is a source of uncertainty but in the rare occasions it happens the following iterative smoothing towards the upper envelope will likely correct some of the damage.

In GEE, four duplicate timeseries are made in order to calculate the measure that identify outliers in the EVI data. The four timeseries are shifted in time, by adding and subtracting five and ten days to and from the timestamp. These four timeseries are then joined so that each image in the timeseries also included the value five and ten days later and prior. The measure was then calculated (F3) and the value above the threshold given the value of the mean value of the bands containing the neighbouring values. The process was then repeated (See 11.11.1 For EVI *p. 126 for example of the script).*

6.2.2.3 Double logistic smoothing

Another type of smoothing was also applied to compare against the MWLR smoothing to get an indication of whether it improved the results and of how much uncertainty is created by the smoothing process.

For the second type of smoothing, a double logistic smoothing (DL) was selected. As opposed to the MWLR that closely follow the observations, the DL is less sensitive to the individual observations, turning all timeseries into a predefined format that is known to represent vegetative seasonality well. Upon visual inspection of several timeseries in Timesat the DL appeared to fit the timeseries well.

The DL smoothing was implemented in GEE, based on the template formulas presented in (Eklundh & Jönsson, 2017) and (Beck et al., 2006) though with slight alterations. The formula applied can be seen below (F4).

$$VI_{DL}(t) = w_1 VI + (mVI - w_1 VI) * \left(\frac{1}{1 + \exp(-mS * (t - S)}\right) + (mVI - w_2 VI) * \left(\frac{1}{1 + \exp(mA * (t - A)} - 1\right)$$

(F4)

It shows how the VI measure, VI_{DL} develops as a function of time, t. $w_1 VI$ represents the VI value at the beginning of the season before it increases and $w_2 VI$ the value at the end of the season after its decrease. mVI is the peak value of the timeseries. S is the inflection point at the

increasing part of the curve i.e. the point where the function goes from being convex to concave and A is the infliction point on the decreasing part. mS is the rate of increase at inflection point S and mA the rate of decrease at point A. Compared to the formula presented in (Beck et al., 2006) the part $(mVI - w_2VI)$ was added and multiplied to the second part of the formula to allow the smoothing to flatten out at a different value at the end of the season than before. An example of the two smoothing types can be seen on Figure 11.

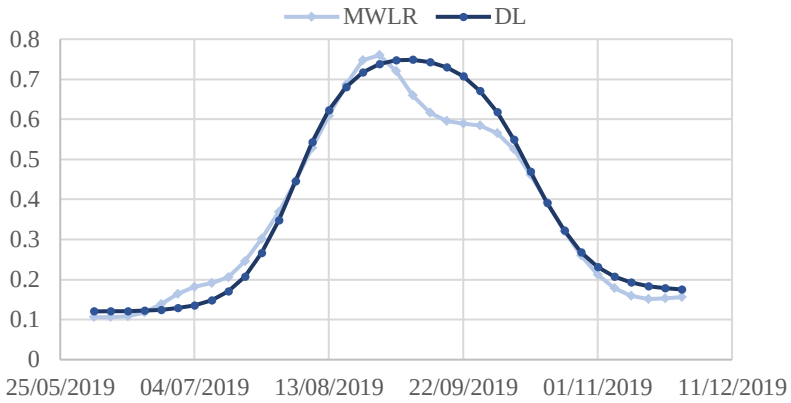


Figure 11: A comparison of the MWLR and DL smoothings (in this example the MWLR only has 6 of the 10 smoothings).

In GEE, first the peak value was found, which was then used to split the season in two, allowing us to find the minimum values in either part and then using these as w_1VI and w_2VI. The inflection points were set on the date that the increase and decrease had reached 50 % of the amplitude. The rates of changes were then found by taking the absolute slope of a linear regression of the observations in a 16-day window around the inflection points. The slopes were multiplied by eight to make the smoothing better fit the observations. The DL was applied on the 10th MWLR smoothing to reduce the chance of the variables in the formula being affected by cloudy observations. The DL was applied after adding observations for every day (see 6.2.3 Inserting a value for every day*), which reduces the potential error when selecting the inflection points (The GEE script can be found in 11.11.2* For the DL smoothings: *p. 131).*

6.2.3 Inserting a value for every day

A modification to the timeseries was needed before defining the season and extracting the VI variables.

There is only an observation every fifth day, so when using a VI value to determine the date of the season boundaries, the time point related to the value will most likely fall somewhere in the

five-day span between two observations. It will then set the boundary at the first date with a value above that VI threshold. This result in a potential leap from the time point that is requested to the time that is returned. The leap can in the worst case be almost five days. This practical hurdle would not be evident if analytical integration was possible, but that would require a function for the VI timeseries, which is not possible with our MWLR smoothing.

The implications of this is that a very small difference in the VI threshold value can determine whether an observation is included or not and thus lead to a large difference when aggregating the timeseries in the specific window to a VI variable. Two almost similar fields could for example end up with quite different areas under the curve. This difference is especially evident when extracting VI variables for the 3 phases, which are shorter and therefore include fewer observations. To address this challenge, values were added to those days without an observation, by assuming a simple linear development in the days between two original observations. This modification minimises the size of the potential leap when determining boundary dates and it is therefore expected to significantly improve the accuracy of the variables. This step is done prior to the double logistic smoothing, so that also the date of the inflection points can be more accurately determined. A graphical example of this can be found in the appendix (11.1 Inserting a value every day p. 104).

In GEE, this modification was made by first making two duplicate timeseries, adding five days to the time stamp for one and subtracting 5 days from the other. The three timeseries were then joined, so that each observation now also included the observation 5 days later and 5 days prior. Then four duplicate timeseries were made, adding one day to the first, two days to the second, and subtracting one and two days from the third and fourth. A new VI measure was then calculated for each, based on the two original observations it is located between. For the first, a fifth of the difference between the two original observations it lay between was subtracted and two fifths for the second, and so on and so forth. The five timeseries were then joined into one timeseries (see 11.11 GEE script: Data preparation and the temporal aggregation *p. 118).*

6.2.4 Defining the season

Before extracting VI variables it is necessary to first define the season. The season can either be defined by specific predetermined dates or based on the VI values. If it is based on dates, it is termed fixed seasonality as they will be identical for all fields. When it is based on the VI, the

dates can differ between fields and is therefore termed dynamic seasonality. The majority of the indices created in this study will be obtained through the dynamic method. Some indices will however also be created with the fixed method to be able to compare the difference.

6.2.4.1 Dynamic seasonality

The dynamic seasonality will both be used to determine the whole season and to determine the start and end of the three phases.

The start of season (SoS) will be determined as the date where the VI reaches 20% of the increasing amplitude and end of season (EoS) as the date where the VI has decreased 80% from the peak.

In GEE, the timeseries is first shortened with a fixed window, broad enough to contain the whole growing season for all fields within the study site. Then a new timeseries is made, within a window from the first day in the broad window to the day of VI peak. A similar window is made from the peak date to the last day of the broad window. In these two series the date with the minimum VI value is found. These are then used as the boundaries for the series used to find EoS and SoS (see 11.11 GEE script: Data preparation and the temporal aggregation *p. 118).*

6.2.4.1.1 Defining the crop phases

As described, paddy rice goes through different phenological stages through the growing season. A hypothesis that is investigated in this project is that a higher correlation between the VI variables and the yield can be achieved if the period covered by the variables are divided in phases according to the crop phenology. To be able to define the boundaries of each stage, a few steps were necessary.

The crop season was split into 3 phases according to the paddy rice phenology. The first was the VS phase, which include the early, mid and late vegetation stage combined into one. The second is FRS and the third is the RMS. Similar to the whole season, the phenological phases are here dynamic, determined by the VI´s instead of the date. This will allow for each field to have different timings of the phases.

To estimate at what VI value, the crop changes phases, the smart phone pictures were utilised. It has prior to our study been analysed manually (by Dvara) which stage is seen on each picture.

The date of the image is also available[5]. For each picture, the VI value of the satellite images was found, at the day the picture was taken, separated in whether it is before or after the peak date, i.e. whether the VI is on the ascending or descending part of the growth cycle. This resulted in 1263 pictures with a VI value and a crop stage. The pictures were then grouped according to their VI values. The groups were made by dividing the VI value range into equal intervals of 0.1 each. For each group of pictures, it was calculated how many percent of them that were of a field with crop stage VS, FRS and RMS[6]. This was also done using relative VI values instead of the absolute VI values, i.e. making the intervals based on how many percent of the season amplitude that is increased or decreased at the time the picture was taken (Figure 12).

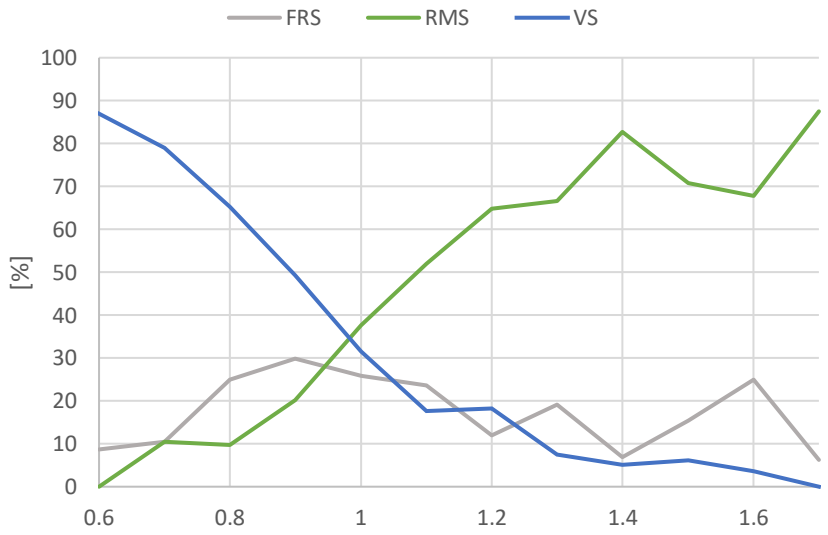


Figure 12: For every VI value (x-axis) the figure shows how many percent of the fields in the smartphone pictures that are in the different phases (y-axis). The x-axis is the VI values fraction of the amplitude divided into intervals. The fractions larger than one are for the descending part of the VI curve. At the x-axis value 1.4 the graphs show the percentage distribution for pictures taken the day, when the VI has declined 40% of the amplitude from the peak. As it is in intervals 40 covers VI values in the interval: [35%-45%]. Of the pictures taken of fields with a VI value in this interval, over 80% were in the RMS phase.

The VS phase is dominant on the increasing part of the season. Around the peak, a switch occurs, and the RMS phase becomes the most dominant. Less prominent is the FRS phase, which tops around the VI peak and again when the VI has decreased to around 60 % of the amplitude. The

[5] As many farmers have taken several consecutive pictures, the initial approach (inspired by Hufkens et al., 2019) was to look at which date the field on the pictures changes from one phase to the next and then see what the VI value was at that date. However, not enough farmers had taken enough pictures for this to give robust results. Another approach was therefore used.

[6] The output of this could for example be: Of all the pictures with a VI value in the interval]0.55:0.65], 90% were classified as being in the VS phase.

latter is disregarded as the FRS phase is known between the other two phases. Though the result is not unambiguous, it does give an estimate of which intervals of the VI values or VI percent of increase, that will most likely give a signal from the specific crop stages.

From the graphs, three windows were made with the aim of isolating the signal from that phase. Different temporal aggregations could then be applied to gain variables for the length of the crop phase, the average value and the sum (see 6.3.1 Temporal aggregation). A preliminary analysis comparing the obtained measures with the farm yield showed that intervals based on the fraction of NDVI had higher correlation with the farm level yield than absolute NDVI. Similarly, broader windows appear to give better results than more narrow ones. These were therefore used henceforth. A graphical representation of the identified windows and how they relate to the crop stage and VI value can be seen on Figure 13 and ***Table 2*** shows the exact interval boundaries.

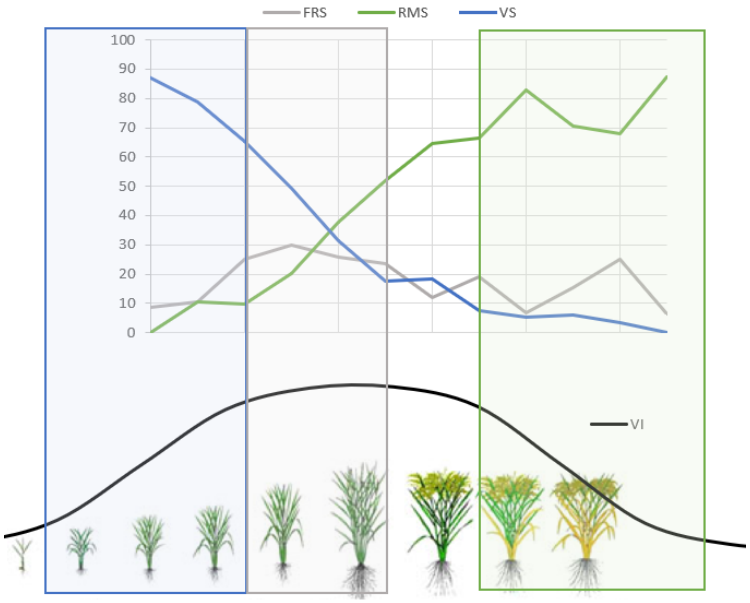


Figure 13: Graphical representation of the phase-windows. The graphs in the top are identical to the previous figure. The boxes represent the windows applied to isolate each phase. The black graph is a visual representation of the VI values and the plants represents the development stage of the rice plant. The blue box represents the window to isolate the VS phase, and the grey and green box the windows to isolate the FRS and RMS phase.

Table 2: *Overview of the used VI windows boundaries. They numbers are the fraction of the amplitude. Values above one are on the decending part.*

Phase	VS	FRS	RMS
VI interval	[0.4: 0.8]	[0.8: 1.1]	[1.3: 1.7]

6.2.4.2 Fixed seasonality

The fixed seasonality is determined by specific boundary dates for the selected windows. These window boundary dates were also found by utilising the smartphone pictures. By using the crop stage classification accompanying each picture, the pictures were separated into the three phases and the distribution of the picture dates for each phase was produced (Figure 14).

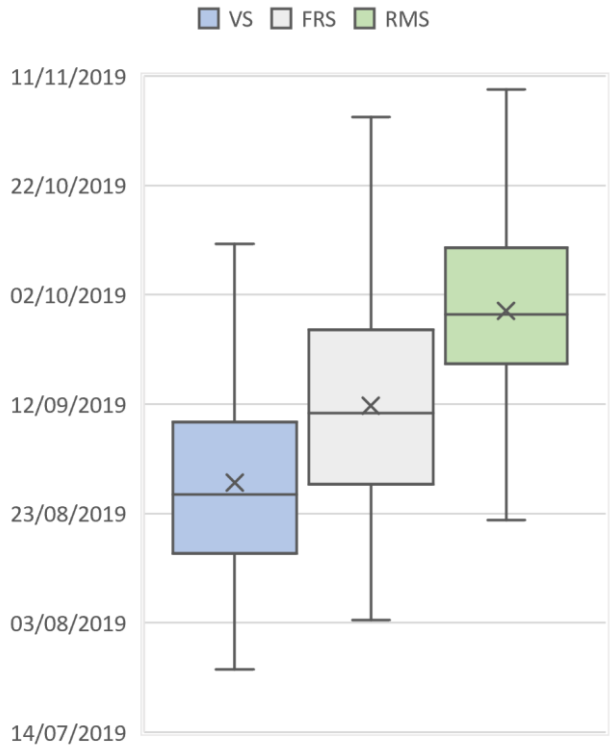


Figure 14: Haryana: A boxplot showing for each phase the distribution of dates on which the smartphone picture was taken.

The boxplots show that the observations for each phase are spread across a long time period, but that the middle 50 percent falls within much narrower windows, with the correct chronological order of the phases. Similar to when determining the phases by the VI measure, the FRS phase is more difficult to distinguish from the other two. The dates of the 25th and 75th percentile were used as the boundaries of the phases for the variables with fixed seasonality.

The 25th percentile of the VS phase and the 75th percentile of the RMS was used to create the fixed window for the whole season. The specific dates can be seen in ***Table 3***.

Table 3*: The fixed boundary dates of the phases and the whole season.*

	Boundary dates
VS	*15/08/2019 – 08/09/2019*
FRS	*28/08/2019 – 25/09/2019*
RMS	*19/09/2019 – 10/10/2019*
WS	*15/08/2019 – 10/10/2019*

6.3 Aggregations – Creating the VI variables

Here the temporal and spatial aggregations are described. The former is used to transform a timeseries into a single image, and the latter to transform that single image into one value for each field with yield data.

6.3.1 Temporal aggregation

The smoothed timeseries were transformed into 88 single images for Haryana and 64 for Odisha, see ***Table 6*** and ***Table 7*** in the appendix for an overview of the triggering measures used for the temporal aggregation.

The series of VI values for each pixel were aggregated to a single value in several different ways, both for the entire season and the three different crop stages. Each triggering measure used for the aggregation is included because of its potential to reveal information about the growing season and therefore the yield. The measures will be explained in the following and a graphical overview can be found on Figure 15.

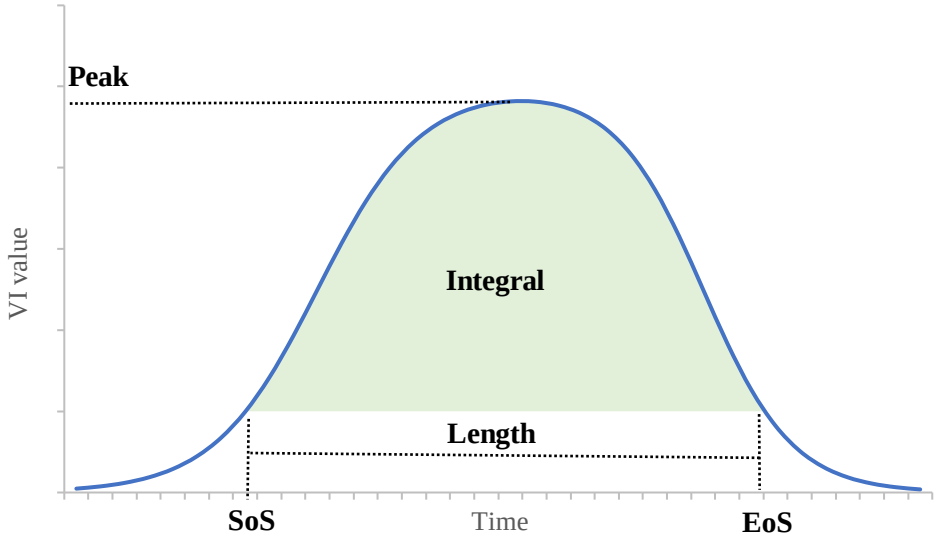


Figure 15: Graphical overview of the triggering measures. The example here is for the whole season.

Peak value: The peak value is the highest value during the season. It is one of the simplest seasonal measures to indicate how well the harvest has been. In studies similar to this one, this measure is one of the most commonly used when estimating yield from timeseries of VI´s (Lobell et al., 2019; IFAD; 2017; Lambert et al., 2018; Azzari et al., 2017; Guan et al., 2018; Lambert et al., 2017, IFAD, 2017).

Integral: The integral is found by numeric integration (i.e. the sum of the daily values) and is an estimate of the area under the curve. For our study, the area beneath 20 percent of the amplitude was subtracted as this gave higher correlations with yield in a preliminary analysis. It thus resembles the area between two curves where the second is representing the constant signal of the fields. The integral is used in several similar studies. It is expected to contain more information about the season than the peak as it is composed of all the values of the season (Flatnes et al., 2019; Lambert et al., 2017; Morel et al., 2014, IFAD, 2017).

Length: The length of the season and of the individual crop stages is the number of days in the period. These measures could reveal different aspects of the season. The length of the FRS phase could be related to how many seeds the plants will produce. The length of the RMS phase could indicate how long the grains are ripening and thus be related to size of the grains. These could therefore contribute to explaining the yield. The length of the season will add to the information from the integral measure, if they are combined. A large integral value could both be due to a

short season with high values or a long with lower values. This could be revealed if the measures are combined.

SoS and EoS: The SoS and EoS determines the temporal boundaries of the season. For the dynamic seasonality, the start of the season is the date that the VI value has increased 20 % of the amplitude from the minimum to the peak and the end of the season is when it has decreased 80% of the amplitude. These measures will reveal information about the timing of the growth season, which could be related to the yield.

SoFRS and EoFRS: The start and end date of the FRS phase. The FRS is an important phase for the yield as it is here the plant develops the flowers which determines the number of rice grains. If this period is timed poorly compared to the weather, it could have an influence on the yield. Only the FRS date boundaries are included as the VS and RMS boundary dates would be similar to FRS phase dates or SoS and EoS for the whole season. The phase date boundaries can therefore be viewed as more generally informing about the timing of the crop phenology.

Mean: The mean value is only found for the three phases. Even with the modification described earlier, inserting a value for each day, one day added or not when calculating the integral could be a noteworthy percentage of these shorter periods. The mean value is less sensitive to this and is therefore included[7].

6.3.2 Spatial aggregation

The timeseries have now been reduced to single images with one value per pixel. A spatial aggregation is then needed to get the VI measure for the individual fields (of which crop cut yield data is available). Using the coordinate of the CCE, the yield data was uploaded as points in GEE. Polygons were then manually drawn around the fields containing a yield data point, using the very high spatial resolution background-image in GEE as reference (Figure 16).

[7] The mean could however also be affected, especially of the phases on the slopes of the timeseries, as one more value in either the high or low end could change the mean in that direction.

Figure 16: An example of the polygons manually drawn to fit the fields containing a yield data point.

Though the CCE was instructed to be undertaken 5 meters from the field edge (see Figure 54), many of the points were quite close to the edge, which made it more challenging to identify the associated field. The GPS accuracy might also have led to some complications. The polygons were drawn for all fields but separated in to three categories based on the confidence in the point laying in a field. In Haryana in the North, 262 points were clearly inside a field, 68 points were more questionable laying either between fields or on parcels of land not visually resembling paddy rice and 4 points were inside cities. Only the 262 points were used later for the later analyses. In Odisha, 83 points were clearly inside a field, 33 were more questionable and 11 were clearly not on a field. Again, only the 83 point were included in the further analyses.

To get one value per field, the median VI value of the pixels within the field was found. The fields are relatively small compared to the size of the pixels. This increases the effect of edge pixels. Natural vegetation or a human made structures located right next to the field might bias the result if a pixel is not fully within the field. Taking the median value is expected to reduce the impact of extreme values within the field but to further reduce the effect of noisy edge pixels, a buffer of 5 meters was made inside the polygon and used as the field boundary instead (Figure 17).

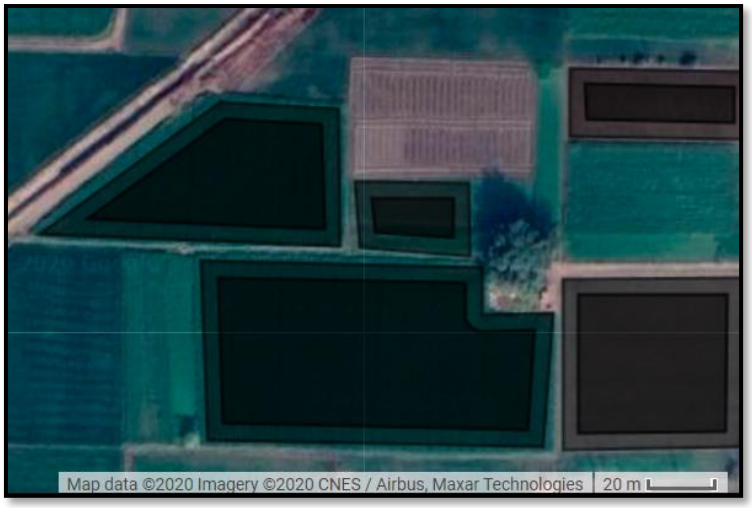


Figure 17: An example of the 5-meter internal buffers on the fields, created to reduce the effect of edge-pixels in the spatial aggregation.

6.4 Preparing the supplementary data

Aside from the VI variables, different sources of data were used to supplement the analyses. How these were prepared and why they were included will be explained in the following.

6.4.1 Bias correcting measures

The VI measures should be a result of the all the aspects influencing the crop. However, some measures were identified that could potentially have affected the yield data without being detected by the VIs. These will be referred to as bias creating variables and were included in the analysis to assess whether the results could be improved, when taking these potential biases into account.

6.4.1.1 Timing of the CCE

The timing of the CCE could bias the data. If assuming that the crops develop heterogeneously among the fields, the time when they are at their optimal harvest point would differ. The timing of the CCE date would thus have to be individualised to avoid potentially biasing the results. This would however be very impractical in practise. It is therefore hypothesised that some fields are harvested for the CCE at a suboptimal time which could lead to a lower yield. To take this into account, a new variable was calculated as the number of days from the EoFRS to the CCE.

6.4.1.2 Rice variety

The relationship between the yield and the VIs differ between different types of crops. At both study sites, several types of paddy rice varieties were farmed. The results from Guan et al. (2018) indicates that even between different varieties of the same crop, there can be differences in the relationship between VI and the yield.

In Haryana, 106 of the final 211[8] fields had information about the rice variety, while all 73 of the final fields in Odisha had the variety information. To be able to include this information in the multiple regression and RF classification analyses, the varieties were ranked by the average grain yield for that variety and numbered according to their rank, giving a value of one to the variety with the lowest average grain yield and two to the second lowest and so on (Figure 50).

6.4.1.3 Soil type

The soil type is also included as a bias creating variable. If the yield is negatively affected by the soil type it would be expected to affect the crop health and therefore be detectable by the VIs. It was however included, as it was hypothesized that nutrient composition of the soil could have an effect on the grain size, thereby affecting the yield without the being detected by the VIs.

For both study sites, there were several different soil types. Similar to the rice variety, they were ranked and numbered according to their average grain yield (Figure 51).

6.4.2 Input variables and ground data

Several other variables were also prepared. These are the input variables that will be used to understand what affects the yield. They include two climatic variables; precipitation and temperature and four ground variables from the yield dataset. All six variables are expected to have an influence on the yield.

6.4.2.1 Temperature

To estimate the temperature, Modis Land Surface Temperature was used. The spatial resolution of the data is 1 km and therefore limits the ability to see inter-field differences. The temperature is however expected to be spatially homogeneous and the data was therefore deemed satisfactory.

[8] The 262 fields with harvest data were reduced to 211, as the remaining were either not located in the same satellite image as the rest or removed because the value was an outlier.

The temperature data have week-to-week variation and is affected by seasonality (Figure 18). Like the Sentinel-2 data it is heavily affected by clouds, creating long periods without data. This only allowed us to extract the mean temperature, and not measures such as numbers of days with temperature above a certain threshold, as was initially intended. To find the mean temperature, the data was smoothed with a MWLR smoothing with a window of 40 days (Figure 18).

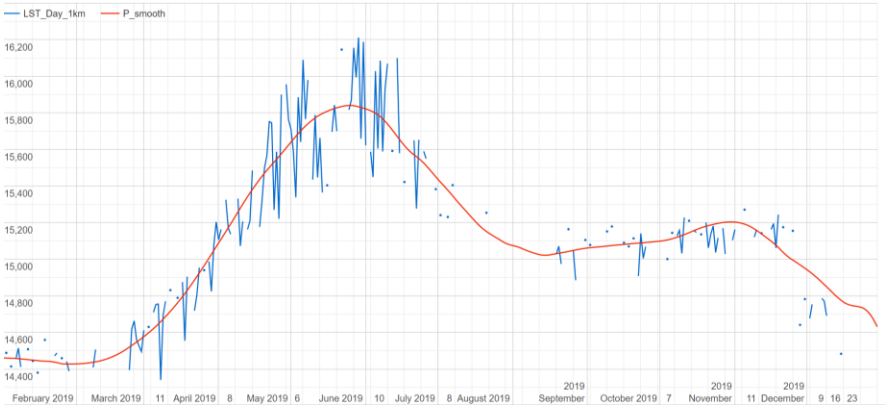


Figure 18: Haryana: An example of the cloud affected LST data (blue) and the smoothed timeseries (red).

The mean temperature was found both for the whole season and the three phases, as defined by the VI values. This was only done for Haryana case, and only with the phases defined by the MWLR smoothing. As the timing of the crop development is unique for each field, the mean temperature in the different phases will also be individualised.

6.4.2.2 Precipitation

The Chirps daily data is used to estimate the precipitation. Similar to the temperature, the spatial scale is large (0.05º ≈ 5 km at the study sites), so the inter-field variation primarily stems from individualised phases in which the mean and sum of the precipitation was calculated. Though the sum and the mean are expected to be quite similar, they could differ based on the length of the periods. This was also only calculated for Haryana, and only in phases created by VI variables with the MWLR smoothing.

6.4.2.3 Money spent

For each smartphone picture in the yield dataset, the farmer was also asked how much money that were spent on the field since last image (on fertiliser, labour etc.). When summing this for each field, it gives an indication of how much the farmer has spent during the season. The

uncertainty of this measure is expected to be high, primarily due to the difference in number of pictures taken.

6.4.2.4 Latitude and Longitude

Both the longitude and latitude of the fields were hypothesised to have an impact on the yield and were therefore included.

6.4.2.5 Sowing date

The sowing date could also be a decisive variable for the yield and was therefore included in the analyses.

6.5 Three statistical analyses

To test the created variables´ ability to estimate yield, three statistical analyses were made: A linear regression, testing the variables´ ability to individually explain the yield variation. A multiple regression, testing the ability of a combined group of variables, and lastly a RF Classification, testing the ability of groups of variables to correctly classify the samples according to their yield.

6.5.1 Linear Regression

Linear regression is used regularly in comparable studies to assess the ability of VI variables to explain the variation in yield. The relation found in the regression would then serve as the basis for estimating yield of a field with a known VI value but unknown yield (Lambert et al., 2017; Lobell et al., 2019; Burke & Lobell, 2017; Guan et al., 2018; Jain et al., 2016).

The prepared VI variables, bias correcting variables and the input-based and ground data variables were compared individually to the yield in a linear regression analysis to assess the strength of the correlation. As the primary aim of this was to compare the measures against each other, the linear regressions were made with the total yield as the dependent variable even though it is the grain yield that is the essential parameter. The rationale for this was that the strength of the correlations was expected to be higher than with the grain yield and that higher correlations are more suitable for comparison, as difference are less likely to be by chance.

In linear regression, a straight line is fitted to the data points by minimising the sum of the squared residuals. This line represents the best overall relationship between the dependent and independent variables. The strength of the relationship is indicated by the coefficient of

determination (R^2), which can be interpreted as what fraction of the variation in the dependent variable that can be explained by the independent variable (McGrew & Monroe, 2009).

The linear regression was first done for all variables in both Haryana and Odisha using all the available samples (fields with VI value and yield data). To test if the bias correcting variables could improve the strength of the relationship, the linear regression analysis was repeated multiple times but each time only including a certain subset of the samples.

In Haryana it was repeated five times for all the variables, each time including only the fields with: 1) The dominant rice variety "12". 2) Soil type "Loam". 3) Soil type "Sandy Loam". 4) Variety "12" & "Loam". 5) Variety "12" & "Sandy Loam".

In Odisha it was repeated five times, only including samples with: 1) The rice variety "Pusa Basmati 1509". 2) The variety "12". 3) Soil type "Loam". 4) Soil type "Sandy Loam". 5) The rice variety "Pusa Basmati 1509" & soil type "Sandy Loam".

Lastly the linear regressions were done again comparing with the grain yield instead of the total yield. This will be used later, when assessing how large a role the imperfect correlation between total yield and grain yield plays in the analyses.

6.5.1.1 t-test and p-value

To assess whether the correlations are statistically significant, a t-test was made. This is especially important when comparing the R^2 between two regression analysis with a different number of observations, as datasets with fewer observations tend to have higher R^2. The test statistic used can be seen in the formula below (F5) (McGrew & Monroe, 2009).

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-R^2}} \tag{F5}$$

t is the test statistic. n is the number of observations. r is the correlation coefficient and R^2 is the coefficient of determination.

From the test statistic and the number of observations a probability value or "p-value" can be calculated. This value indicates the probability to get that exact test statistic (t) if the hypothesis of no relation between the variables was true. If the p-value is very small, we can reject the

hypothesis that there is no relation as there is only a very small change that an error is made. The level of significance i.e. the p-value threshold used in this study is $\alpha = 0.05$. Linear regression with a p-value below this will be considered statistically significant (McGrew & Monroe, 2009).

6.5.2 Multiple Regression

The majority of the variables are not expected to explain the yield variation alone, but rather in combination with other variables. Several multiple regression analyses were therefore made. Multiple regression analysis is an extension of the linear regression analysis, where multiple independent variables are included to explain the variation of the dependent variable. The output coefficients of a multiple regression could then be used for yield estimation for fields were the independent variables are known.

Multiple regression analyses were made in Excel for several different groups of variables. The variables were divided into the following groups for both Haryana and Odisha[9].

1. WS_VI: A group with only VI variables from the entire season i.e. no variables that differentiated between crop stages.
2. WS_VI_Bias_Corrected: A group with the same WS_VI-variables and the three bias correcting variables (rice variety, soil type and days-to-CCE).
3. ALL_Fixed: A group with the WS variables and the five VI variables found with fixed season boundaries (integral of the three phases and WS, and the peak value). This one was only done for Haryana.
4. All_Fixed_Bias_Corrected: A group with the same variables as ALL_fixed, but also including the bias correcting variables.
5. Phases_VI: A group with VI variables from both the whole season, but also the three phases.
6. Phases_VI_Bias_Corrected: A group with the same variables as Phases_VI and the three bias correcting variables.
7. NO_VI: A group with the bias correcting variables, the input-based variables and the ground data. For Odisha, neither precipitation nor temperature were included.

[9] Group 3 and 4 are not made for Odisha.

For each of the groups 1, 2, 5 and 6, four different groups were made, separating the variables based on the vegetation index and smoothing type. Splitting the variables in groups will give indications of:

- Which vegetation index and smoothing type is to be preferred.
- How much the variables from the phases improves the explanatory ability compared to when only including variables of the entire season.
- Whether the dynamic seasonality is preferred over the fixed.
- Whether including the bias creating variables improves the correlation.
- Which ground and input variables that affect the yield.

Details of the groups can be found in the appendix (11.5 Overview of grouped variables p. 108).

In multiple regression analysis it is important that each variable contributes to the model. An iterative process of running the analysis and then removing the variable with the highest p-value was therefore done. The p-value for each variable in multiple regression indicates the probability of that variable not contributing to the model. Variables were removed until only variables with p-value below 0.05 remained. Another important aspect in multiple regression is to eliminate multicollinearity i.e. not have independent variables with strong correlations among them. Two of the triggering measures, the mean and the integral are very similar, and it was therefore decided to prior to the multiple regression, remove that variable of the two with the lowest individual correlation with the yield (McGrew & Monroe, 2009; ArcGIS Pro, 2020).

The output of the analyses is an R^2 value which has been adjusted to the number of variables used (adjusted R^2) and a significant F statistic indicating whether the model is statistically significant.

6.5.3 Random forest classification and variable importance

To supplement the results of the multiple regression analysis and to gain further information about the importance of each variable, several Random Forest Classifications (RF Classifications) were performed. Having discretised the yield data into five categories, the analysis will show to what extent the variables can be used to correctly categorise the yield data. Additionally, it will show which of the input variables that contributes most to this

categorisation. The RF Classification, trained on the samples, could then be used on an AOI to classify each pixel to a yield category. It would thus give a yield estimate for all fields.

The RF classifier is an ensemble model classification method consisting of multiple classification threes (Belgiu & Dragut, 2016).

A classification three can be interpreted as a ruleset or a set of binary questions through which observations are divided into homogeneous subgroups. It much resembles a decision tree but allocates qualitative data where a decision tree allocates quantitative data. Each binary question splits the data in two subgroups, and this continuing process grows the classification tree. The process is called recursively binary splitting, indicating that successive splits of subgroups are dependent on the previous splits. Which input variable and which threshold that is used to make each split is governed by what split of the observations that will create the most homogeneous groups i.e. minimize the number of observations not belonging in the dominant class of the subgroups. The homogeneity of the subgroup is referred to as the purity (James et al., 2013; Boehmke & Greewell, 2020).

For each tree in the RF Classification, only a randomly selected subset of the input variables is considered as candidates for each split. This is done to reduce the amount of correlation between the trees. Additionally, each classification three is trained on an individual subset of the samples, found with a "bagging approach". For each classification tree, the bagging approach randomly selects samples from the dataset equivalent to around two thirds of the data. A process that allows for samples to be used multiple times, thereby increasing the number of different training samples available (James et al., 2013; Boehmke & Greewell, 2020; Belgiu & Dragut, 2016).

In this study, 500 decision trees were used, a typical amount for a study like this. The classification trees each assign all the samples to a yield class. This is aggregated into one classification result for the entire random forest using a majority vote i.e. each sample will be assigned to the class where most classification trees allocated it (James et al., 2013; Boehmke & Greewell, 2020; Belgiu & Dragut, 2016).

The remaining third of the samples are used as validation samples (out-of-bag samples) to estimate the performance of the classification. The output variable of this cross-validation

method is the proportion of out-of-bag samples that are correctly classified by the random forest classification, measured by the Cross-Validation Score (CVS) (James et al., 2013; Boehmke & Greewell, 2020; Belgiu & Dragut, 2016).

Two essential measures for statistical learning methods, such as RF classification is the bias and the variance. The bias refers to the ability of the model to fit to the dataset while the variance refers to how much the accuracy changes, when using a different training dataset. These often poses a trade-off, as very flexible models will have low bias but high variance (overfitting) and vice versa. The advantages of using ensembles of decision trees and the bagging approach in the random forest classification is that each decision tree can minimize the bias while the high variance is reduced by averaging the results of the trees (James et al., 2013; Boehmke & Greewell, 2020; Belgiu & Dragut, 2016).

The random forest classifications also produce an importance-measure for each variable. This indicates how well the variable on average has been able to split the samples into pure subgroups in the RF classification. For this study, the measure will be used as an indicator of how suitable the VI variable is for use in index creation. This will be done by observing the difference in appearance among the top 10 most important VI variables for each site, when the RF classification is run with all VI variables.

For the multiple regression analysis, it is important to reduce multicollinearity, but as RF Classification is less sensitive to this, the RF Classification was first run using all VI variables and then on all VI variables and the bias correcting, the input-based and the supplementary ground data. This was done for both Haryana and Odisha (Belgiu & Dragut, 2016; McGrew & Monroe, 2009).

It was then run on the same groups of variables as presented above in the description of the multiple regression analysis and with the same objectives. The analyses were made in the software "Spyder" using 500 threes[10]. The used script can be found in the appendix (11.13 Spyder script – RF Classification p. 136). The grain yield was discretised into five equal intervals for both Haryana and Odisha though merging the highest and lowest to the adjacent interval to avoid intervals with only very few observations (***Table 4***).

[10] The default setting was chosen for the number of variables to select when randomly selecting a subset.

Table 4: *Grain yield intervals for the RF Classification.*

Haryana	***Odisha***
]10: 17.5]	]2.5: 7.5]
]17.5: 22.5]	]7.5: 12.5]
]22.5: 27.5]	]12.5: 17.5]
]27.5: 32.5]	]17.5: 22.5]
]32.5: 40]	]22.5: 27.5]

Lastly, a RF classification will be run on the total yield using all variables for later comparison of the effect of the mismatch between total yield and grain yield.

6.6 Uncertainty from choice of smoothing and VI

From the analyses above, it is already possible to compare the VIs, the smoothings, the effect of bias creating measures and the effect of the imperfectly correlation between total and grain yield. However, to elaborate on the uncertainty from the smoothings and choice of VI, an uncertainty analysis was done to assess which aspects of the index-creating method that creates the most uncertainty. This was done for the smoothing type by parring the previously obtained VI variables, so that the only difference is the smoothing type. For each pair, the coefficient of determination (R^2) was then calculated with linear regression. The higher the R^2 values obtained, the more similar the pairs are, and the less decisive is the choice of smoothing. This was then done similarly for the two VIs and the two design options were then compared. The results can be interpreted as an indication how much uncertainty is created from the choice of VI and smoothing type.

7 Results

In this section, the results of the analyses will be presented. The section will be structured according to the research questions. For each question the results of all three statistical analyses for both study sites will be included. Conclusions from the results will first be drawn and discussed in the discussion section that follows immediately after the result section.

7.1 Factors influencing the yield

Here it will be presented which input-variables influence the yield according to our results and how much of the yield variation can be explained by them (Research Question 1).

7.1.1 Linear Regression

In Haryana, the linear regression showed that of the ground- and bias correcting variables; longitude, sowing date, rice variety and two of the four variables with "days to CCE", had a significant correlation with the total yield. Of the eight precipitation variables 3 were significant and of the 8 temperature variables 2 were significant. Longitude had the highest R^2 (0.07), while the rest could explain less than 5% of the total yield variation. See appendix for a full overview of the results (11.9 Linear regression – No VI variables p. 116).

In Odisha, the date of the CCE, the rice variety, four of the four temperature variables and four of the four "days to CCE" had a significant correlation with total yield. Six of these could explain more than 15% of the yield variation (11.9 Linear regression – No VI variables p. 116).

7.1.2 Multiple Regression

When only using the climatic, the ground and the bias correcting variables (i.e. no VI variables) the multiple regression show that 12% of the grain yield variation could be explained in Haryana and 17% in Odisha.

7.1.3 RF Classification

When using no VI variables, the RF classification could correctly classify 46% of the samples in Haryana, which is the highest CVS score in Haryana.

In Odisha 40% were correctly classified by the RF classification, which is among the lowest CVS scores for Odisha.

7.2 VI variables´ ability to explain yield variation

In this section the results showing the overall ability of the VI variables to estimate the field level yield will be presented (Research Question 2).

7.2.1 Linear Regression

The key results of this comprehensive analysis will be presented here. A visual representation of the full results of the analysis can be found in the appendix (11.6 Results of the individual correlations p. 110).

In Haryana, 49% of the 70 produced VI variables had a statistically significant correlation with the total yield, when using all available observation pairs (Figure 19). Much fewer variables (16%) had a significant correlation with the grain yield. The mean R^2 of only the significant variables was below 0.05 for both total yield and grain yield. This indicates that the VI variables on average can explain less than five percent of the variation in the farm level yield in Haryana. The variable with the highest R^2, could explain 15% of the total yield variation and 5% of the grain yield variation.

In Odisha more variables had a significant correlation with the grain yield (46% of the 61 VI variables) than with the total yield (23%). Even though the total yield and grain yield were more similar in Odisha, this result was unexpected and might be an indication that caution should be taken when comparing results of correlations with low R^2. The significant VI variables in Odisha could on average explain 9% of the variation in total and grain yield, while the best could explain 15% and 20 % respectively (Figure 19).

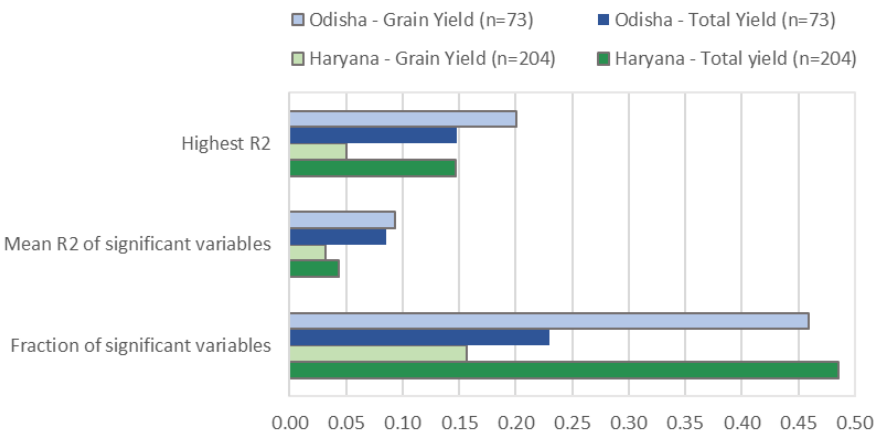


Figure 19: Overview of the linear regression results.

7.2.2 Multiple Regression

In Haryana, 19 multiple regressions were made on groups of variables and grain yield. 16 of them resulted in significant models. These significant models could on average explain 10 % of the grain yield variation and the best model could explain 24 % (Figure 20). In Odisha only 1 of the 17 multiple regression results were not significant, and the significant variables could on average explain 26 % of the variation in yield. The best model in Odisha could explain 53% of the variation in grain yield (Figure 20). A table with all the results of the multiple regression analysis can be found in the appendix (11.7 Full result of the multiple regression analyses p. 112).

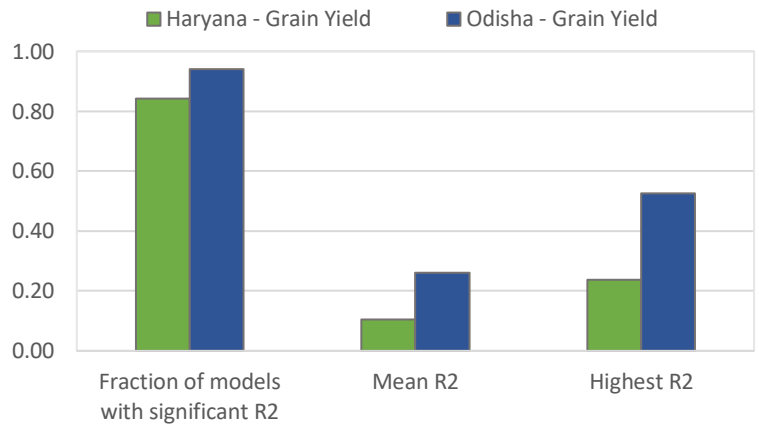


Figure 20: Overview of the multiple regression results.

7.2.3 RF Classification

The result of the RF Classification run with all VI variables showed that they could correctly classify around 38 % of the samples into the correct grain yield class in Haryana and around 50 % in Odisha (Figure 21). The output graphs also show that the additional gain in classification accuracy from including more 15-20 variables is limited and even worsening in the Haryana example. A result that justify the limited number of variables in the different groups on which the RF Classification was run subsequently.

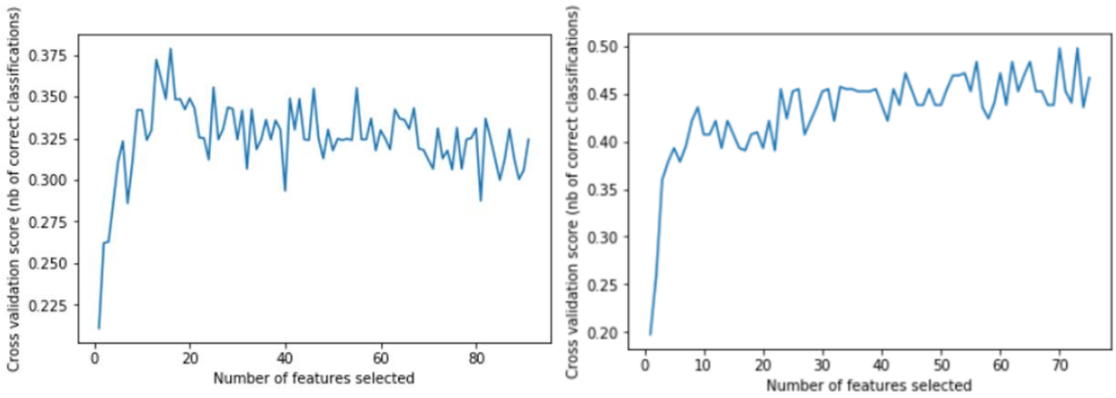


*Figure 21: CVS as a function of Number of features selected in the RF Classification with all VI variables for Haryana (**Left**) and Odisha (**Right**).*

Two groups of variables in Haryana and one in Odisha did not return any results. The results of the remaining runs using the different groups of variables show that on average 34% were correctly classified for Haryana and 41% for Odisha and that the best group could correctly classify 46% and 47% in Haryana and Odisha respectively (Figure 22). Tables with the full results of the RF classification analyses can be found in the appendix (11.8 Full results of the RF classification p. 114).

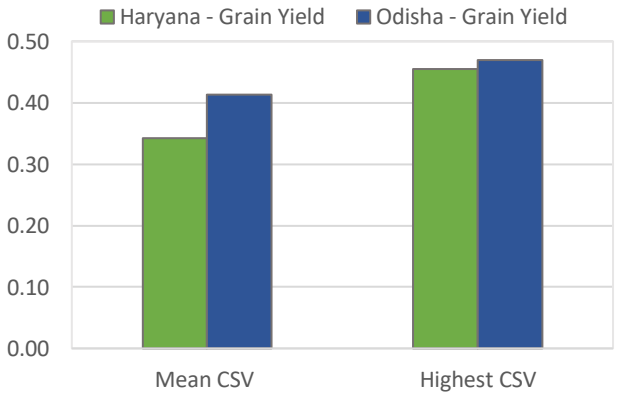


Figure 22: Overview of the RF Classification results (excluding the runs with all variables).

7.3 Suitability of the design options

The results of the analyses will here be extracted to compare the different design options (Research Question 3). The comparisons based on the linear regressions will be against the total yield to use the highest R^2 values. For the multiple regressions and RF classifications the

comparison will be based on the ability to estimate grain yield. The design options to be assessed are the VI (EVI or NDVI), the smoothing type (MWLR or DL), the triggering measurement (peak, integral, mean, length, SoS or EoS), the period (phenologically tailored phases or the whole season) and the seasonality (dynamic or fixed seasonality). It is important to note that there is a high degree of uncertainty when comparing very low R^2, which are evident in especially the linear regression results.

7.3.1 Vegetation index: EVI and NDVI

As described previously, the EVI should have less tendency to saturate and might therefore be better at distinguishing between the high VI values found in the study areas (Son et al., 2013). However, the EVI timeseries were more complicated to smooth due to its two-sided respond to cloud and cloud shadow disturbances. This might also have played a role in the following results.

7.3.1.1 Linear Regression

For Haryana, 21 of the 37 NDVI-based variables were significant, while only 15 of the 37 EVI-based were. Additionally, the significant NDVI-variables could explain more of the total yield variation (5% for NDVI and 3% for EVI) (Figure 23 - Left). In Odisha, both NDVI and EVI only had 9 significant variables of the 32, but they could on average explain 10% for NDVI and 9% for EVI (Figure 23 – Right).

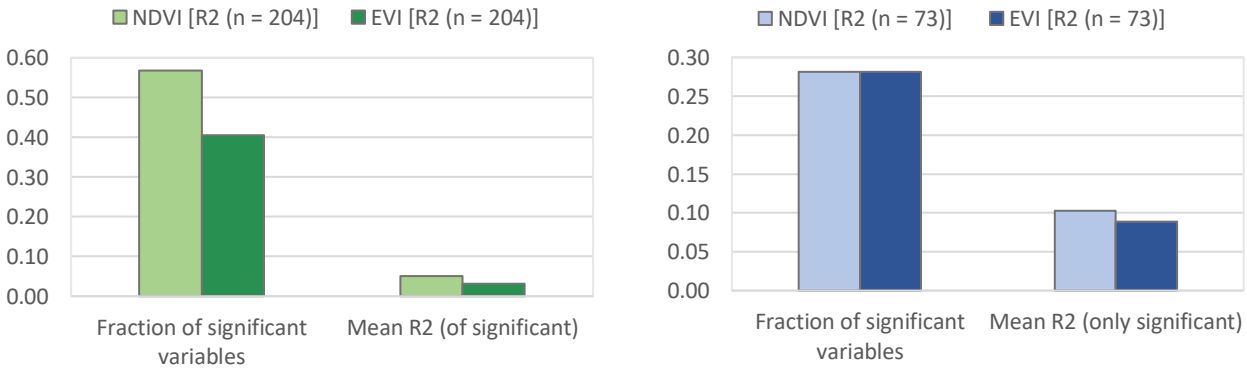


*Figure 23: Comparison of the linear regression results between EVI and NDVI for Haryana (**Left**) and Odisha (**Right**).*

For both study sites the difference between NDVI and EVI thus appear quite small. When looking more closely at the significant variables, the difference between EVI and NDVI is still negligible for Odisha, but it becomes clearer for Haryana as the NDVI variables are almost

consistently higher and have several variables with around double the R^2 compared to EVI (Figure 24).

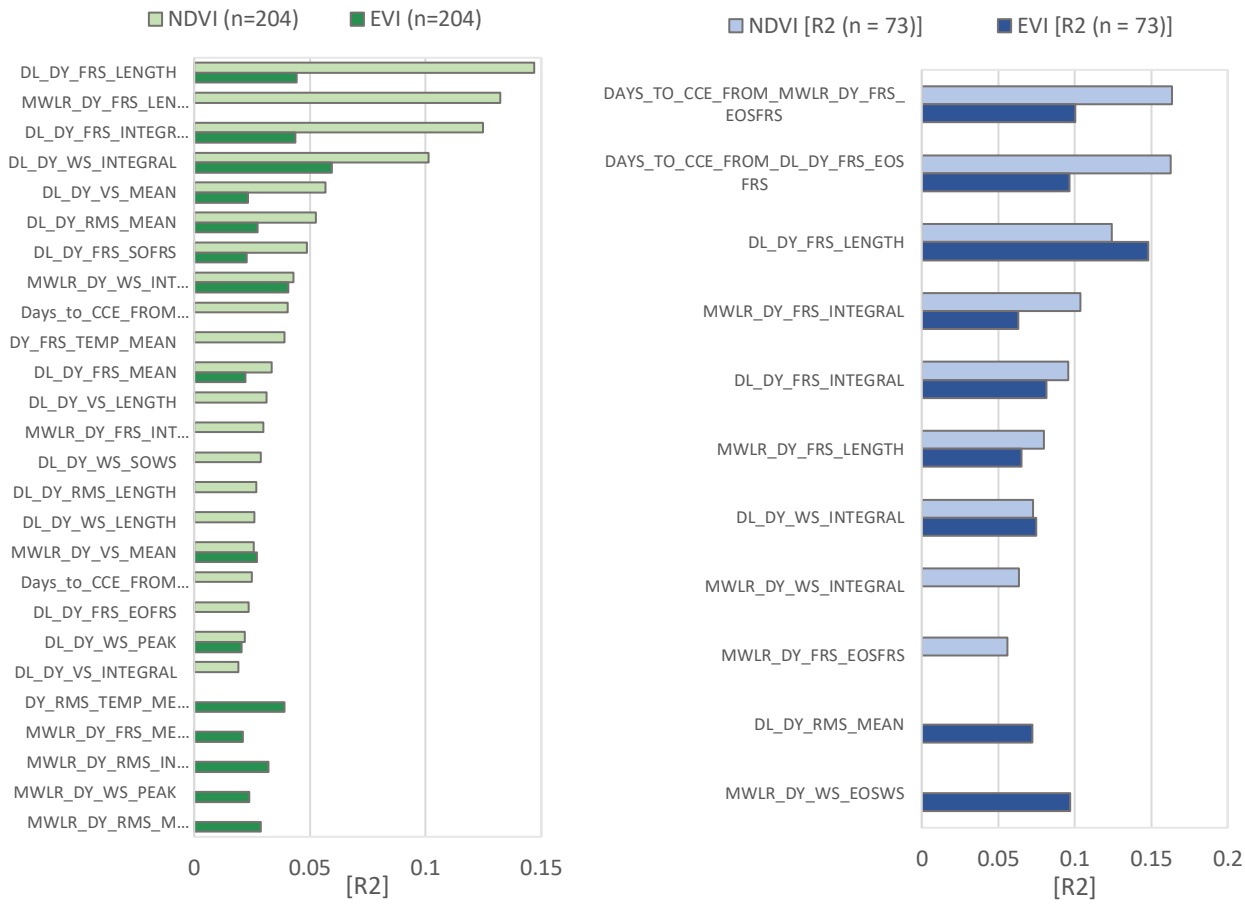


*Figure 24: Comparison of the linear regression results for each comparable variable between EVI and NDVI for Haryana (**Left**) and Odisha (**Right**).*

7.3.1.2 Multiple Regression

For Haryana, all three groups that did not yield a significant result were EVI groups, but the one insignificant group for Odisha was an NDVI group. When comparing the groups against each other with only the VI as difference, the results for Haryana is not very clear. On average the NDVI groups are slightly better, but three of the four categories with R^2 above 0.1 are based on EVI (Figure 25 - Left). The results for Odisha consistently show higher R^2 for the EVI groups and have a more than 50% higher mean value compared to the NDVI groups (Figure 25 - Right).

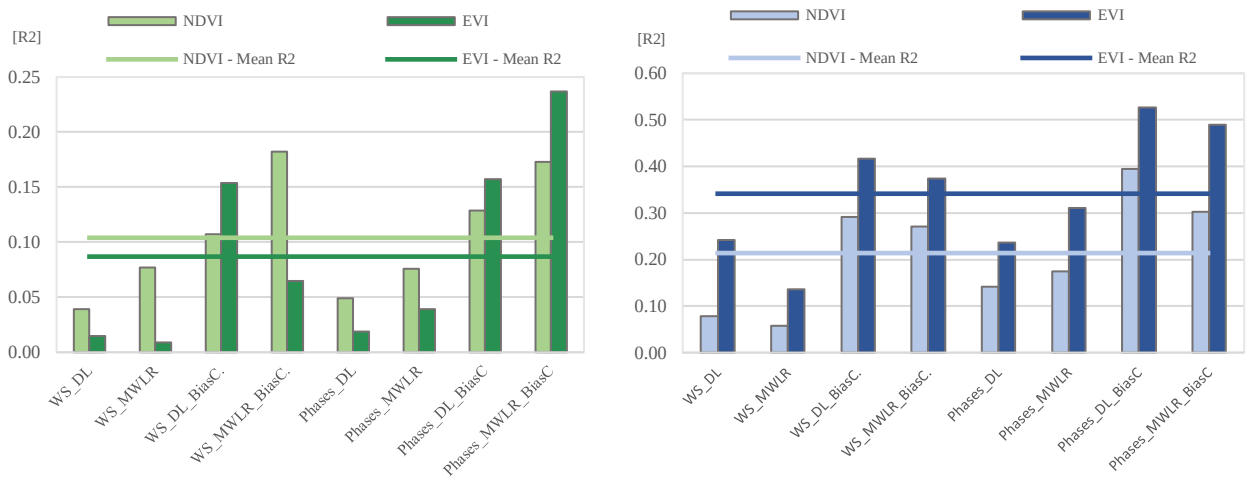


*Figure 25: Comparison of the multiple regression results between NDVI and EVI for Haryana (**Left**) and Odisha (**Right**).*

7.3.1.3 RF Classification

The RF classifications showed only very little difference in accuracy between the classifications from the EVI and NDVI groups, with slightly higher average CVS for EVI in Haryana and slightly lower CVS for EVI in Odisha (Figure 26).

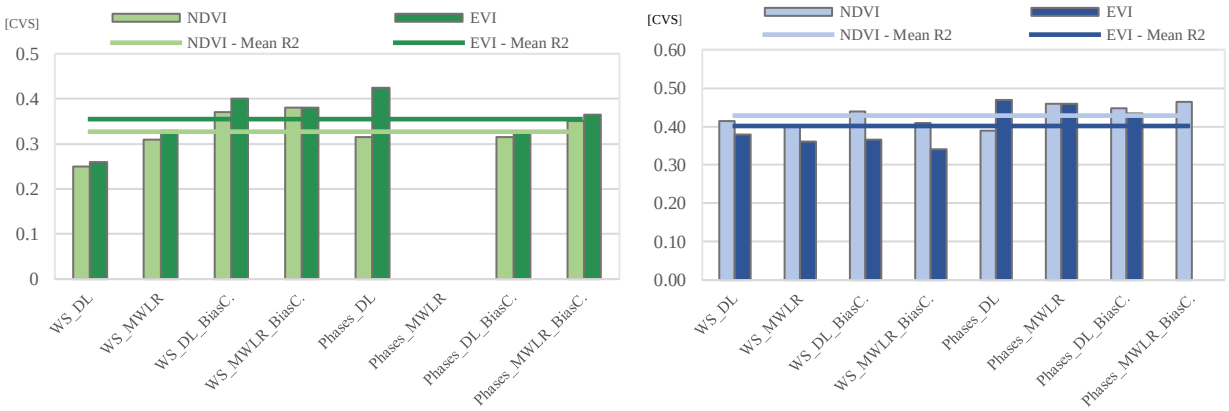


*Figure 26: Comparison of the RF classification results between NDVI and EVI for Haryana (**Left**) and Odisha (**Right**).*

The variable importance measure from the RF classification also can reveal information about which VI performs better. To assess this, the 10 most important variables in Haryana and 10 most important variables in Odisha for when the RF classification is run with all VI variables, is observed. Of these 20 variables, 8 were with NDVI and 12 with EVI (Figure 33).

7.3.2 Smoothing: MWLR and DL

Two different types of smoothing were used to reduce the effect of cloud and cloud shadows in the timeseries. The aim of the MWLR was to fit it as closely to the data as possible allowing for

subtle inter-field differences. In contrast, the DL smoothing was expected to be less sensitive to the individual datapoints but more secure, due to its predefined shape.

7.3.2.1 Linear Regression

More of the DL-based variables were significant compared to the MWLR for both Haryana and Odisha and the average R^2 of the significant variables were higher. The differences are however negligible (Figure 27).

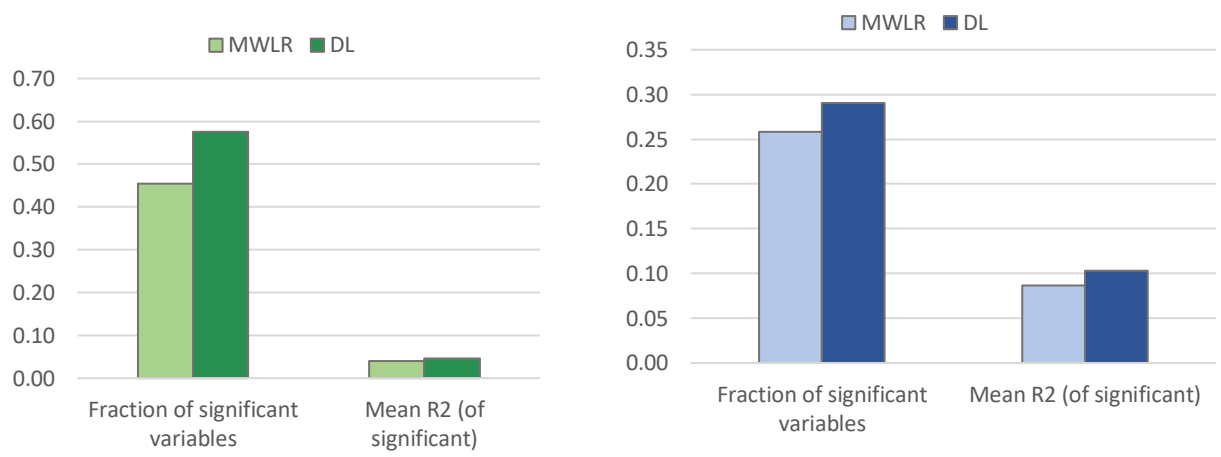


*Figure 27: Comparison of the linear regression results between MWLR and DL for Haryana (**Left**) and Odisha (**Right**).*

7.3.2.2 Multiple Regression

For Haryana, the groups based on the MWLR returned the highest R^2 in the multiple regression analysis. In Odisha it was opposite, as the DL-based groups consistently showed higher R^2 (Figure 28).

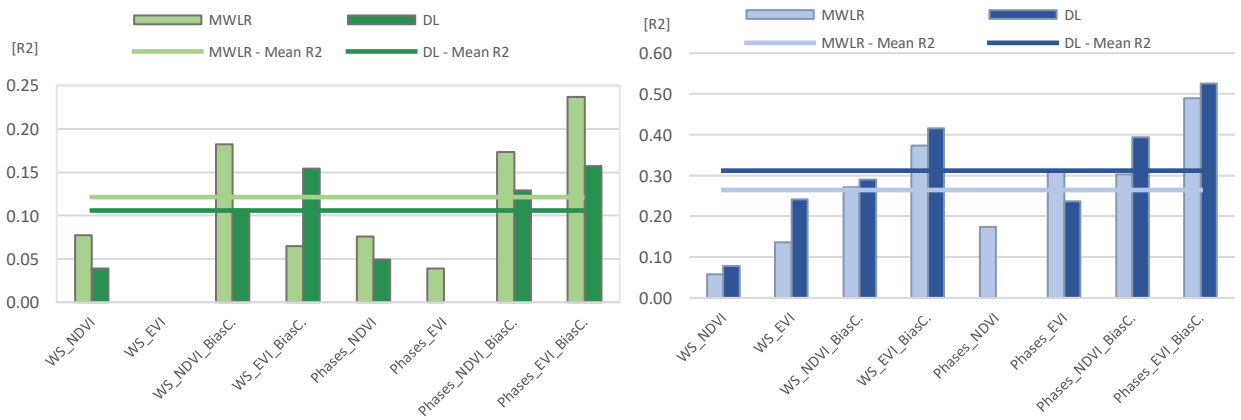


*Figure 28: Comparison of the multiple regression results between MWLR and DL for Haryana (**Left**) and Odisha (**Right**).*

7.3.2.3 RF Classification

The RF classification did not show any clear differences between the groups with MWLR-based variables and DL-based variables in neither Haryana nor Odisha (Figure 29).

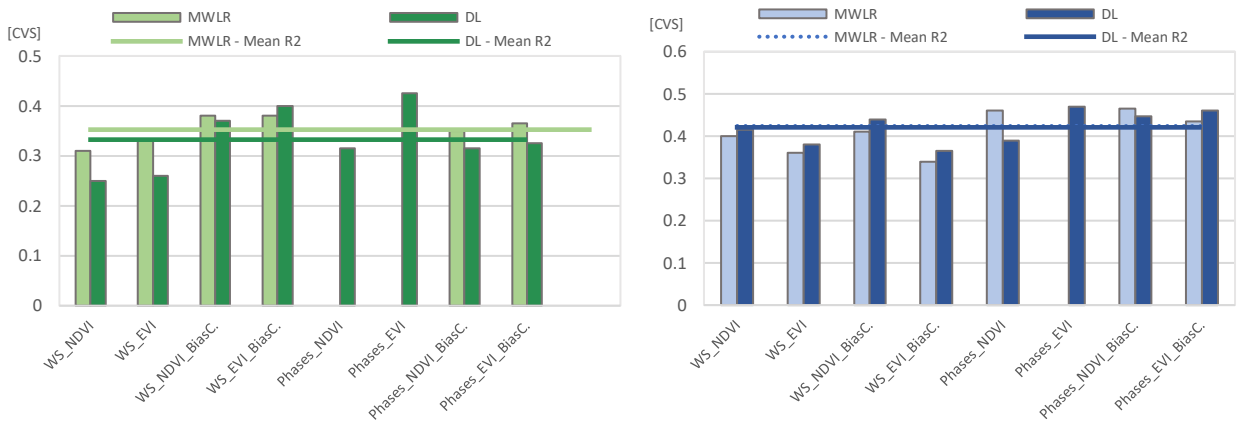


*Figure 29: Comparison of the RF classification results between MWLR and DL for Haryana (**Left**) and Odisha (**Right**).*

Of the 20 most important variables, 9 were with the MWLR smoothing and 11 were with the DL smoothing (Figure 33).

7.3.3 Triggering measure: peak, integral, mean, length, SoS or EoS

The triggering measures´ influence on the results will here be compared. The "peak" and the "integral" are the most commonly used in similar studies and are therefore expected to yield the highest results. The other measures could however also contribute with valuable information of the crop season and might therefore be more important in the multiple regression and RF classification. The triggering measures are more difficult to directly compare, as not all measures were used in all design combinations. There are for example no "peak" variables for the phases. The mean result is therefore also be affected by whether the triggering measure is included more frequently with other design options that performs well.

7.3.3.1 Linear Regression

In Haryana, the two highest correlations were with "length" as the trigger measurement and the two second highest were with "integral". The remaining variables were all considerably lower (Figure 30 - Left). Looking more closely at these (Figure 30 - Right), it becomes apparent that the "integral", "length" and "mean" on average have the highest values, though with a very high variance. The variables with the lowest R^2 are the EoS and SoS. The peak is on average in

the middle, but with much less variance. This might however also be due to it being used in fewer variables.

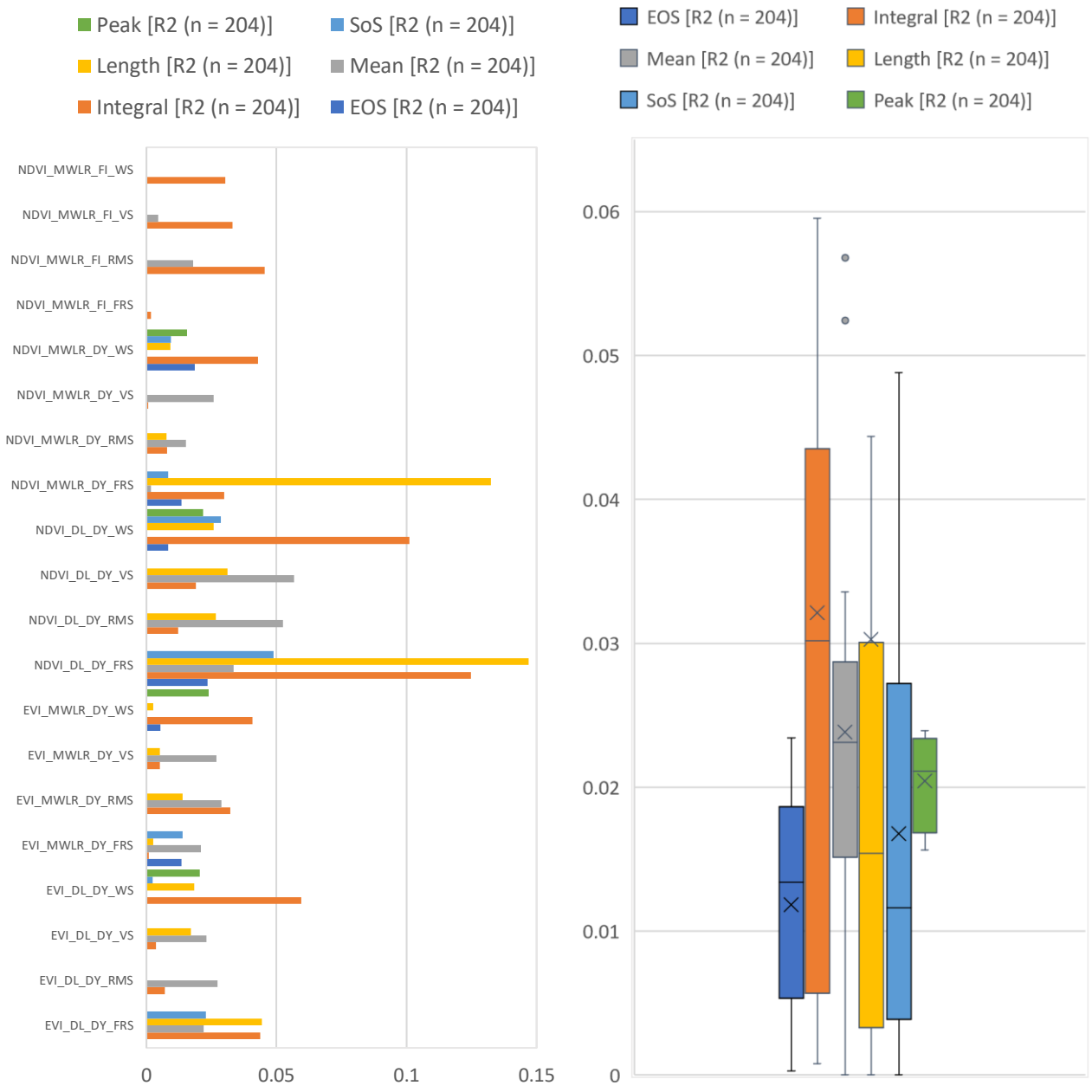


*Figure 30: For Haryana: Comparison of correlation with total yield between variables with different triggering measurements, first for the individual variables (**left**) and then summarised in boxplots (**Right**). The four highest R^2 from the left are not included in the boxplots to the right.*

In Odisha, the highest R^2 were from variables with "length" and the "integral", while the EoS variables had the highest mean R^2 and SoS the lowest. The peak variables were the second lowest on average and with a higher variance than in Haryana (Figure 31).

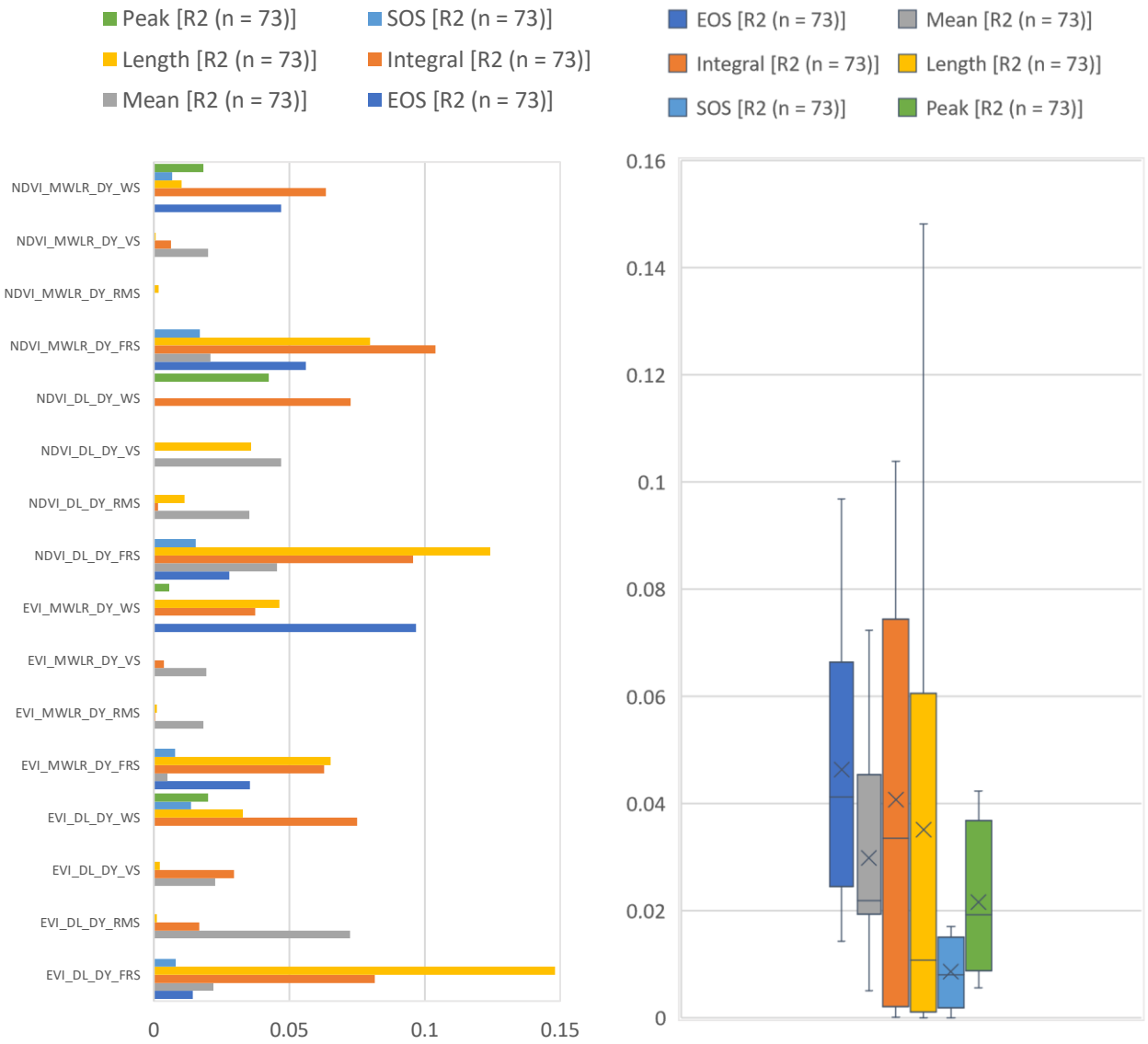


*Figure 31: For Odisha: Comparison of correlation with total yield between variables with different triggering measurements, first for the individual variables (**left**) and then summarised in boxplots (**Right**).*

7.3.3.2 Multiple Regression

The results of both the multiple regression and the RF classification does not directly allow for comparison of the different triggering measurements, as they are in groups with each other.

As an alternative, it was assessed how many times a variable with the specific triggering measurement was included as a significantly contributor in the multiple regression analysis. Of the 32 multiple regression analysis[11], all triggering measurements were included regularly. VI variables with "length" and "integral" were however used more often (Figure 32).

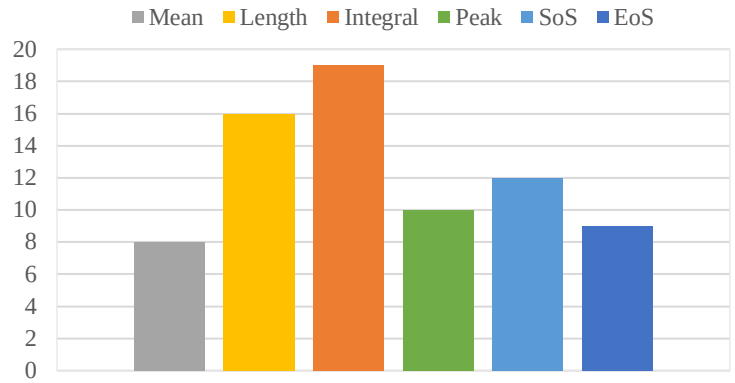


Figure 32: Number of times a variable with the specific triggering measure was used in the 32 multiple regression analyses for Haryana and Odisha

7.3.3.3 RF Classification

The variable importance output from the RF classification can also reveal information about the suitability of the different triggering measures. Of the 10 most important variables for Haryana and 10 most important variables for Odisha, six had "integral" as the triggering measure, six had "mean", another six had "length" and two had the "peak". Neither "SoS" nor "EoS" featured in the top 20 most important variables from the two RF classifications using all VI-variables (Figure 33).

[11] 8 for WS and 8 for Phases for both Haryana and Odisha.

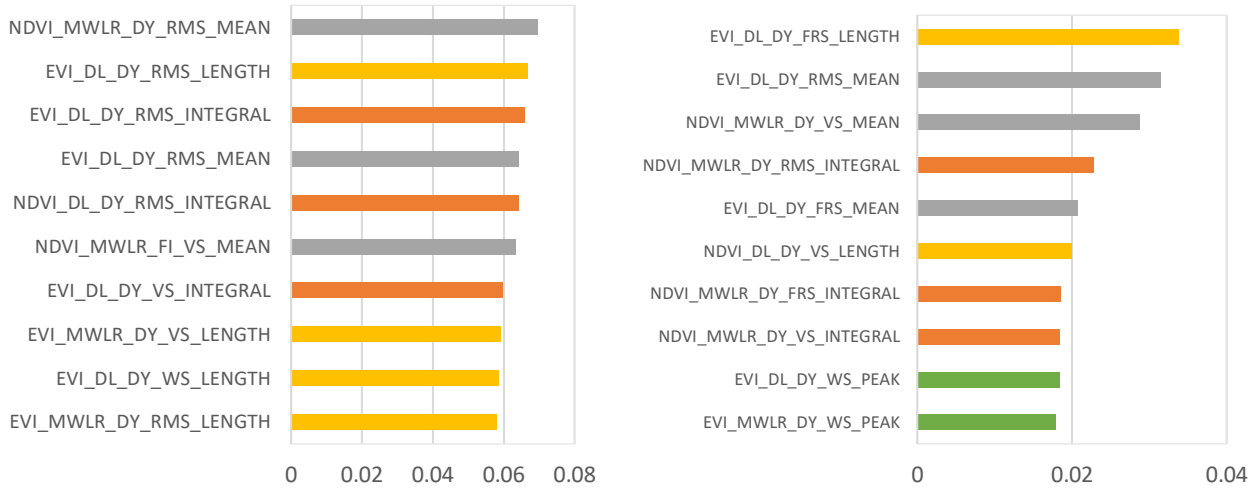


*Figure 33: Top 10 most important variables in the RF classification with all VI variables for Haryana (**Left**) and Odisha (**Right**). The bars are color coded according to the triggering measure.*

7.3.4 Period: Whole season and phases

Here it will be assessed whether including variables derived from phenologically tailored phases can increase the accuracy of the yield estimations compared to only including variables for the whole season.

7.3.4.1 Linear Regression

For the individual correlations with the total yield, the VI variables derived from the FRS phase have the highest R^2 for both Haryana and Odisha, followed by the variables from WS (Figure 34).

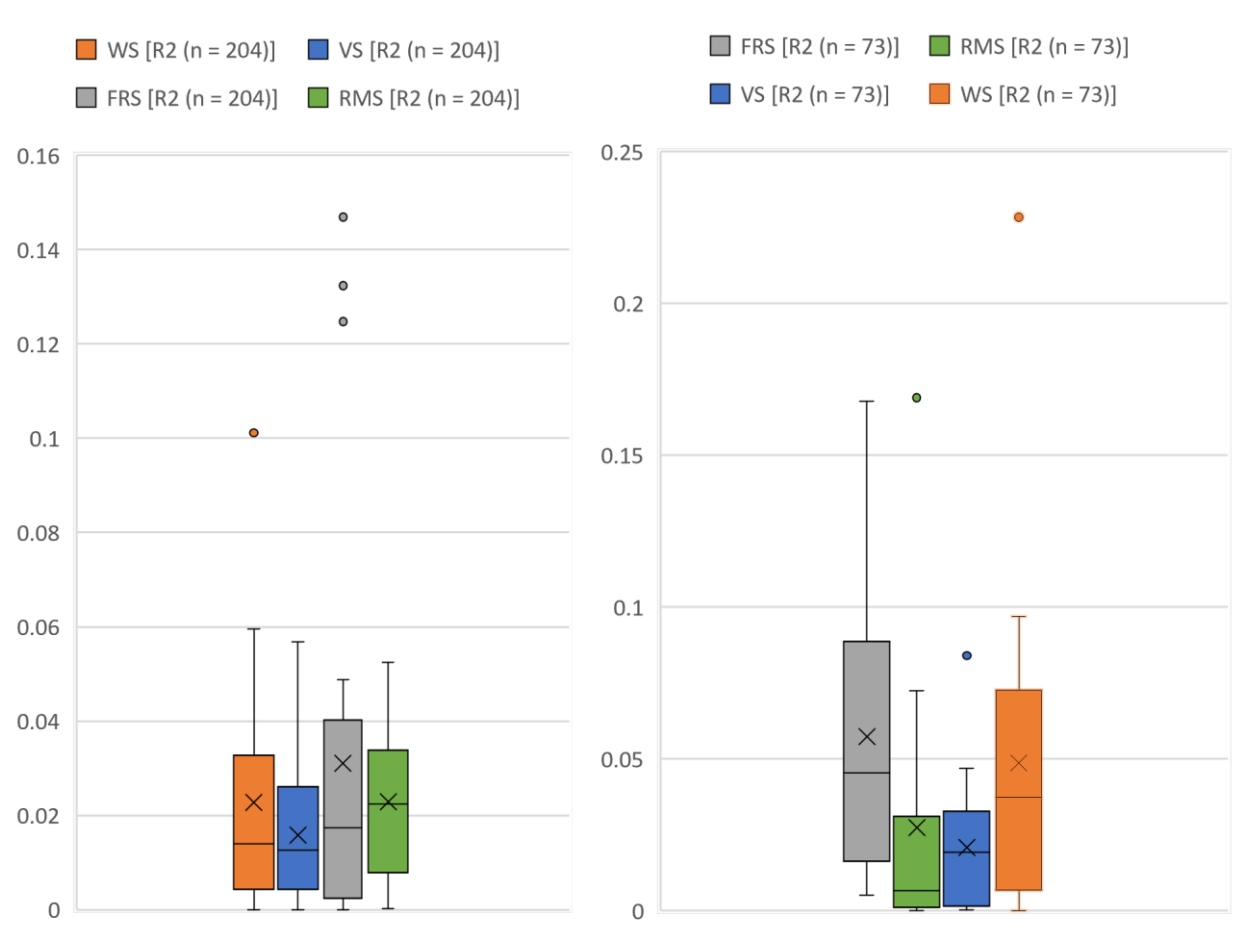


*Figure 34: Comparison between the whole season and the phases of the R^2 found in the linear regression with total yield for Haryana (**left**) and Odisha (**Right**).*

7.3.4.2 Multiple Regression

On average, the adjusted-R^2 of the multiple regression increased when including the variables from the three phases, especially in Odisha where it improved around 50%. The phase-group that could explain most of the grain yield variation, could explain 24% in Haryana and 53% in Odisha, while the best WS-groups could only explain 18% and 42% (Figure 35).

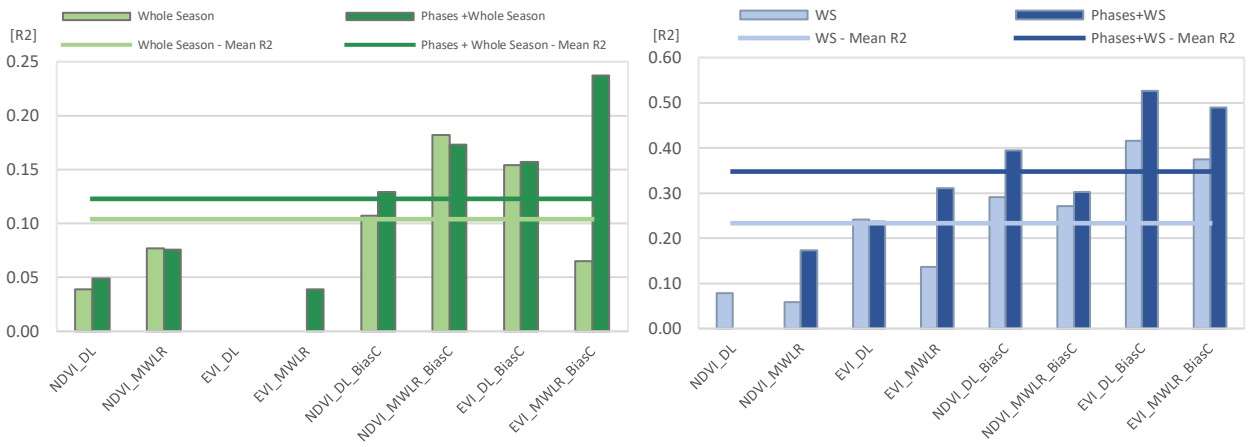


*Figure 35: Comparison between the whole season and the phases of the R^2 found in the multiple regressions with grain yield for Haryana (**left**) and Odisha (**Right**).*

7.3.4.3 RF Classification

In Haryana, the mean CVS increased when including the phase-variables, but only very little. For Odisha, including the phase-variables also improved the groups´ ability to correctly classify the samples into the correct grain yield category, with a mean CVS around 15% higher than when only including WS-variables (Figure 36).

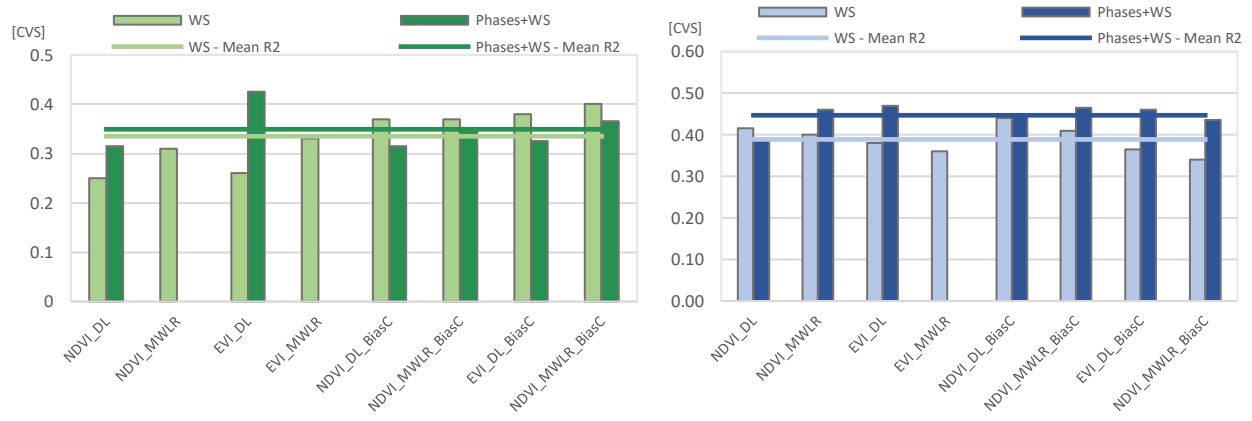


*Figure 36: Comparison between the whole season and the phases of the CVS found in the RF Classifications with grain yield for Haryana (**left**) and Odisha (**Right**).*

The variable importance measure from the RF classification with all VI variables also reveal the importance of the phases (Figure 33). Of the 10 most important variables for each study site, only three were for the whole season, placed 9th Haryana and 9th and 10th in Odisha.

7.3.5 Seasonality: Dynamic and fixed

The last design option to be assessed is whether a dynamic seasonality based on the VI value or a fixed seasonality based on specific dates will result in the variables best able to estimate grain yield. The dynamic seasonality is capable of individualising the timing for when the variables are extracted and is therefore expected to be better able to identify differences in farm level yield. The fixed variables were only done for Haryana with NDVI and the MWLR smoothing, resulting in eight variables (integral and mean for the three phases, and integral and peak for the whole season).

7.3.5.1 Linear Regression

The linear regression did not show any clear differences between the dynamic and fixed seasonality. Of the eight fixed-season variables, only three had a significant correlation with total yield. So had the dynamic seasonality for the corresponding variables. Neither could explain more than 5% of the yield variation (Figure 37).

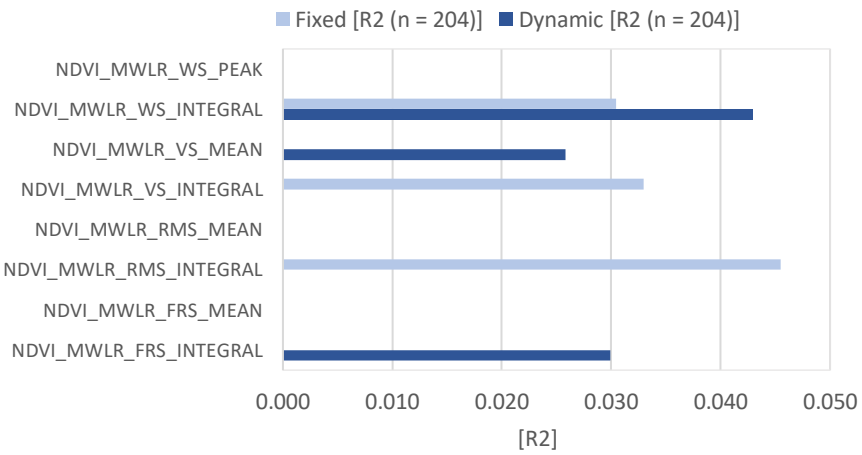


Figure 37: Haryana: Comparison between the dynamic and fixed seasonality of the R^2 found in the linear regression with total yield.

7.3.5.2 Multiple Regression

The results of the multiple regression analysis showed that for the two groups without bias correcting variables, the R^2 was around 70% higher for the group with dynamic seasonality. For the two groups with bias correcting variables, the group with dynamic variables had a 16% higher R^2 (Figure 38).

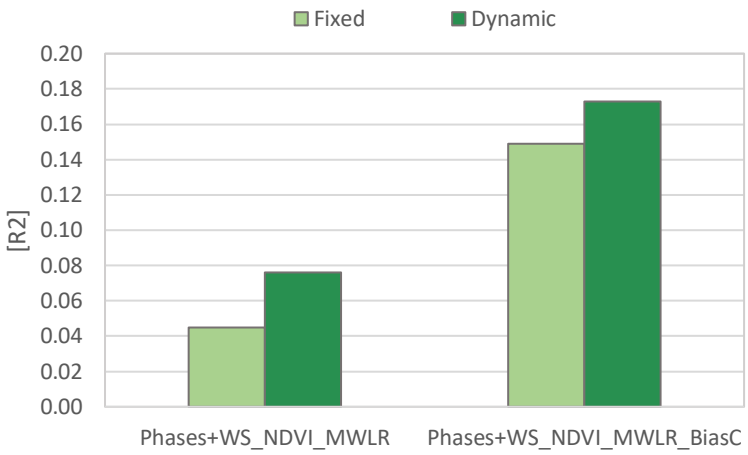


Figure 38: Haryana: Comparison between the dynamic and fixed seasonality of the R^2 found in the multiple regression with grain yield.

7.3.5.3 Rf Classification

The group with dynamic seasonality was only slightly better at correctly categorising the samples in the RF classification (Figure 39). Had the design option been compared over more examples, the results would likely have been more clear.

Figure 39: Haryana: Comparison between the dynamic and fixed seasonality of the CVS found in the RF classification of the grain yield. One of the dynamic runs did not return a result.

7.4 Sources of uncertainty in the index creation process

In this section, results will be presented to assess the magnitude of uncertainty from different sources. The aim of this is to give an indication of what should be focussed on in succeeding studies (Research Question 4).

7.4.1 Uncertainty from VI and smoothing

For both VI type and choice of smoothing, the correlation between matching variables was found. These R^2 values for the different variables were presented as boxplots in the graphs below for both Haryana and Odisha (Figure 40). For both design options and both study sites, there was a variance in the R^2 values. The correlations between MWLR and DL variables were on average around 0.53 for Haryana and 0.58 for Odisha. Significantly lower were the correlations between NDVI and EVI variables, with an average of 0.29 in Haryana and 0.33 in Odisha. This indicates that the choice of VI is the most decisive of the two.

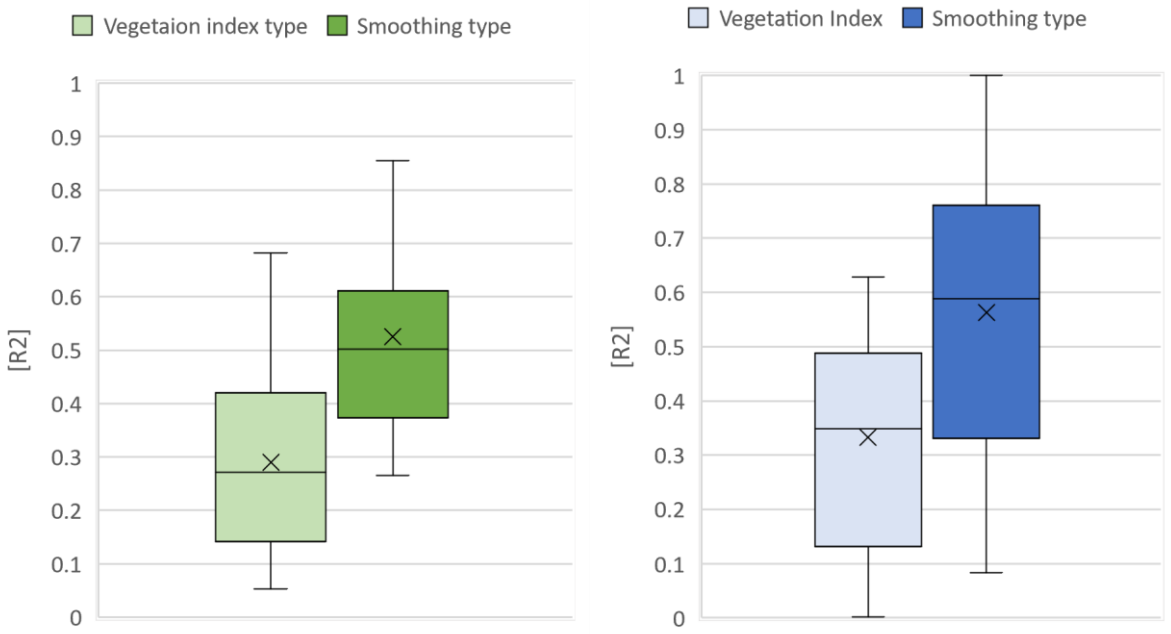


*Figure 40: Comparison between the R^2 distribution of vegetation index and smoothing type for a linear regression of matching variables, for Haryana (**Left**) and Odisha (**Right**).*

7.4.2 Effect of bias correcting variables

Several bias correcting variables were identified. It will here be assessed whether including these could improve the results.

7.4.2.1 Linear Regression

The linear regression analysis was also done on certain subsets of the samples, split according to rice variety and soil type. See appendix for the full result (11.6 Results of the individual correlations p. 110). An important thing to note is that the number of samples in each group varies and is significantly lower than when including all the samples.

For Haryana, more VI variables had a significant correlation with the total yield, when including all samples. For grain yield it did however increase the fraction of significant VI variables when running the regression with only samples with rice variety "12" and when running with samples with the soil type "Loam" (Figure 41 - Left). In Odisha, isolating the samples with rice variety "12" increased the fraction of VI variables with a significant correlation with the total yield and only slightly decreases it for grain. The other isolated sample groups reduced the fraction considerably for both total yield and grain yield (Figure 41 - Right).

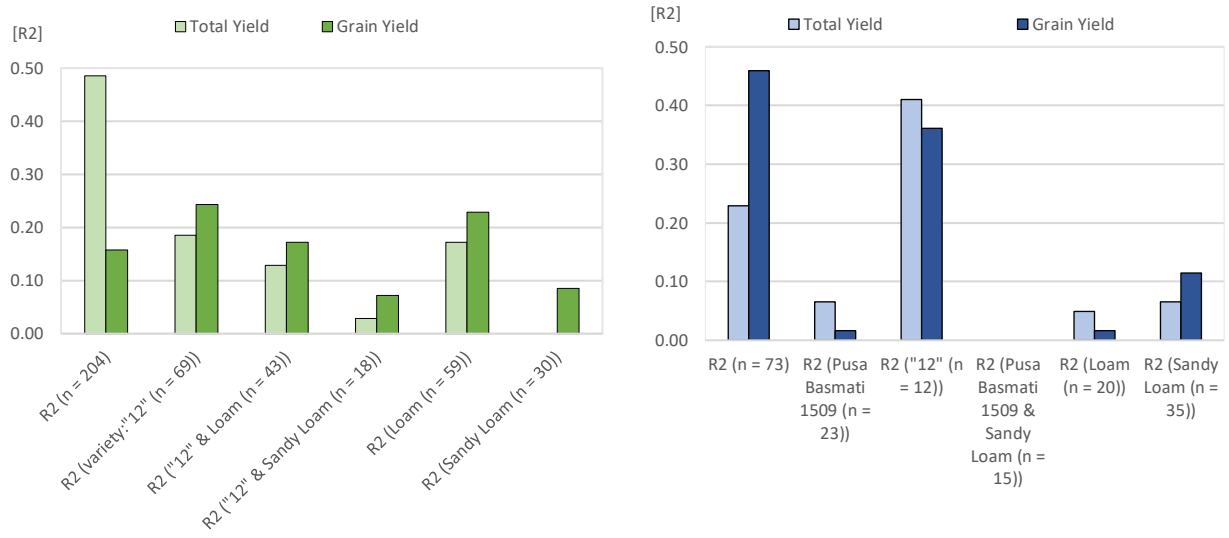


*Figure 41: Fraction of variables with a significant correlation with yield for the different sample groups for Haryana (**Left**) and Odisha (**Right**).*

The mean R^2 of these isolated groups of the samples can be seen below on Figure 42. In both Haryana and Odisha, the mean R^2 of the groups are higher for all the groups compared to including the full sample. For both study sites the highest R^2 are found on the groups with the fewest samples and there appears to be a consistently lower R^2 the more samples are included (Figure 42 & Figure 43).

In Haryana, the significant VI variables can on average explain more than 25% of the variation in total yield and grain yield of the field samples with "Sandy Loam" & "12" (Figure 42). As seen on Figure 41 it is however only a very little fraction of the variables that are significant, which challenge the robustness of the result.

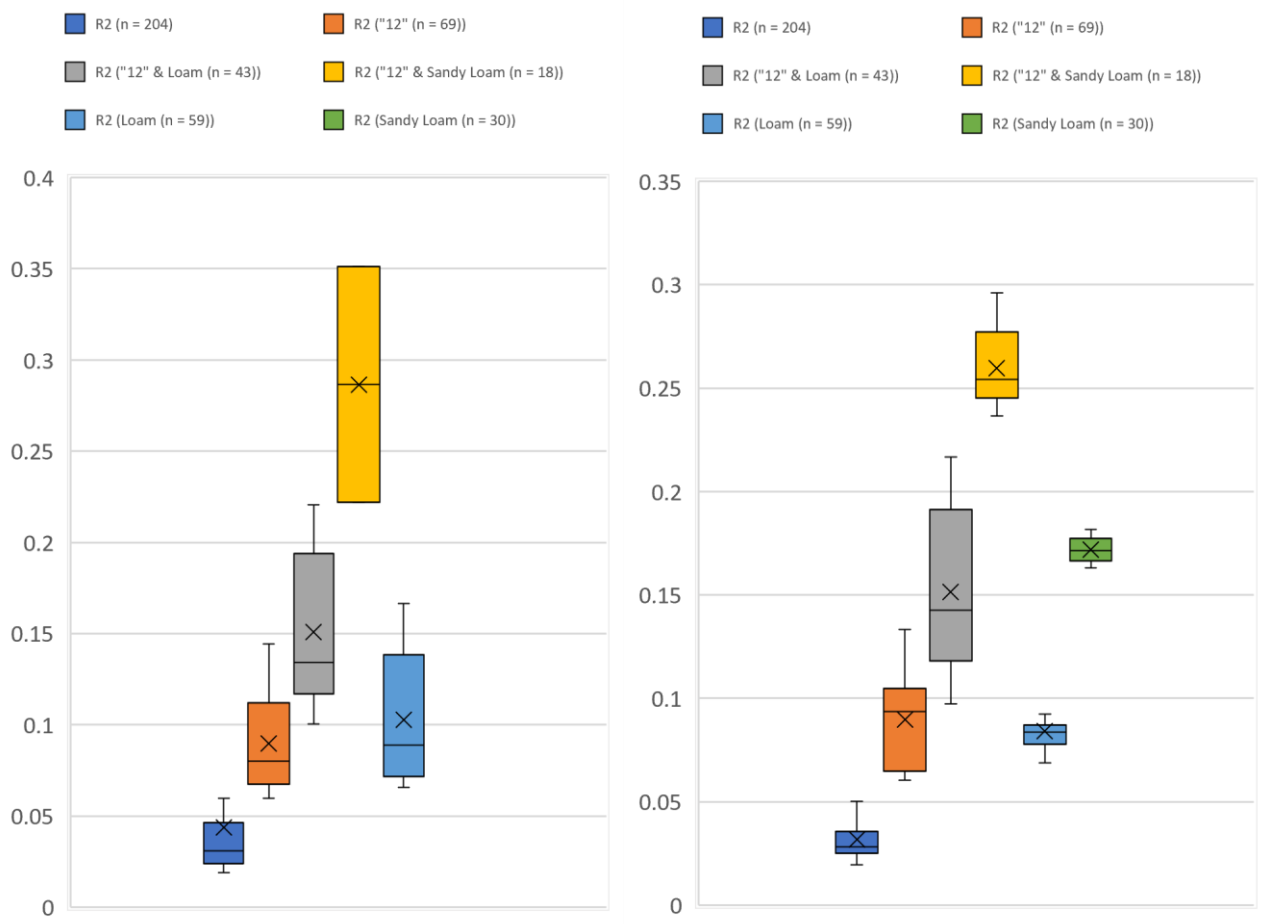


*Figure 42: Haryana: Comparison between the different groups of samples of the R^2 from a linear regression with total yield (**Left**) and grain yield (**Right**).*

In Odisha, the highest mean R^2 was found when isolating the samples with rice variety "12" (Figure 43). The VI variables could on average explain 45% of the total yield variation and 65% of the grain yield variation, when isolating this specific variety. The best VI variables could with this sample subset explain around 86% (Figure 43 & Figure 44) of the grain yield variation and 15 VI variables could explain more than 60%. The consistently high R^2 across multiple VI variables, also seen on the large fraction of significant variables (Figure 41), increase the robustness of the results and suggest that the R^2 can be improved if analysing the rice varieties separately. Had the high R^2 only been due to the small sample size, it would also have been expected to find some significant correlations when isolating the 15 samples with variety "Pusa Basmati 1509" & soil type "Sandy Loam", but none were found.

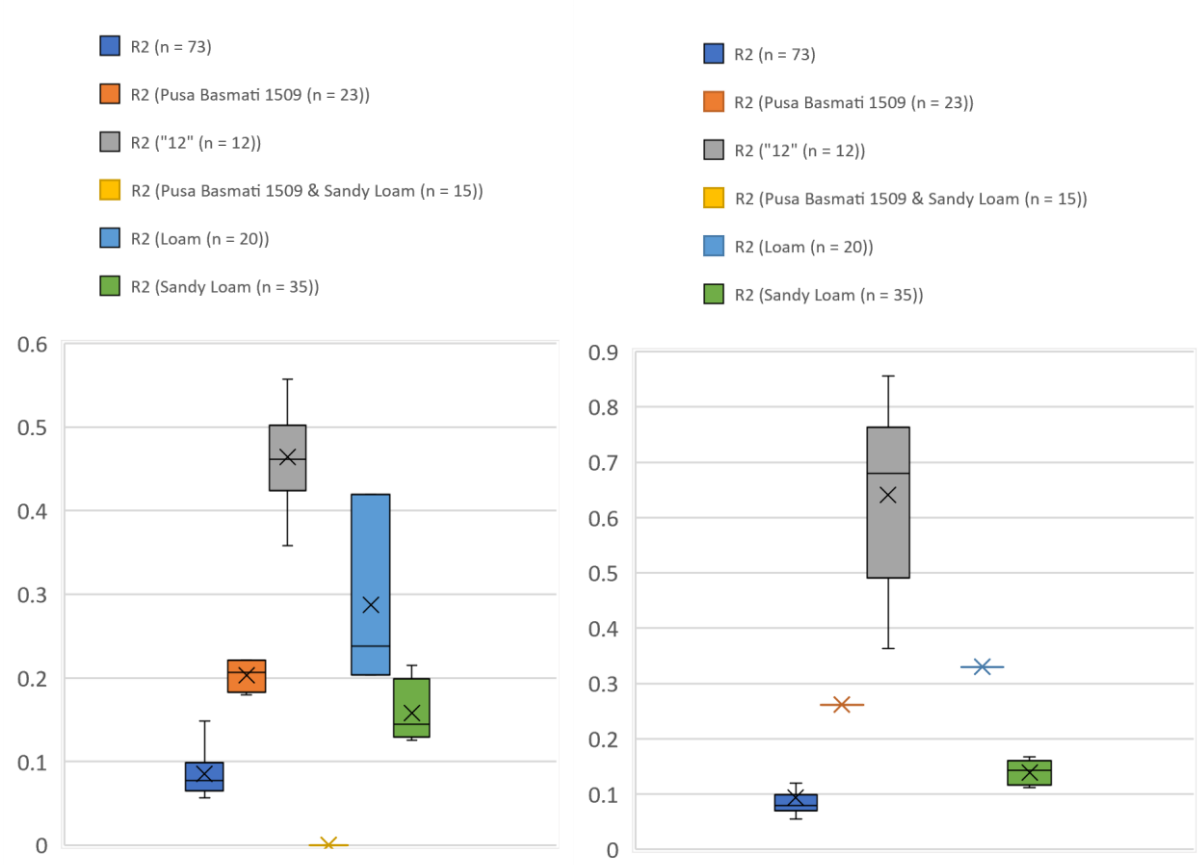


*Figure 43: Odisha: Comparison between the different groups of samples of the R^2 from a linear regression with total yield (**Left**) and grain yield (**Right**).*

A scatterplot of the best correlation for Odisha found with a subset of the samples (variety "12") can be seen on Figure 44.

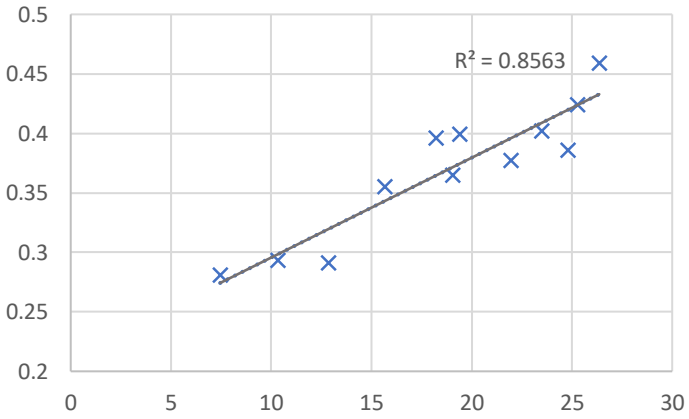


Figure 44: Odisha: A scatterplot of the best correlation with grain yield. The variable is: EVI_MWLR_DY_RMS_MEAN and it is with the sample subset: rice variety "12".

7.4.2.2 Multiple Regression

For the multiple regressions, the potential bias in the yield data was accommodated by including three variables. One with the rice variety ranked and numbered according to the mean yield for that variety. One with soil type, similarly ranked and numbered by the mean yield, and lastly one with the number of days to the CCE from either EoFRS (for phase variables) or SoS (for WS variables).

For both Haryana and Odisha, more groups of variables became significant when including the bias correcting variables, and the mean adjusted-R^2 of the significant groups increased with 169% and 117% respectively. Including the bias correcting variables thus dramatically improved how much of the variation in grain yield that could be explained by the different groups of VI variables (Figure 45).

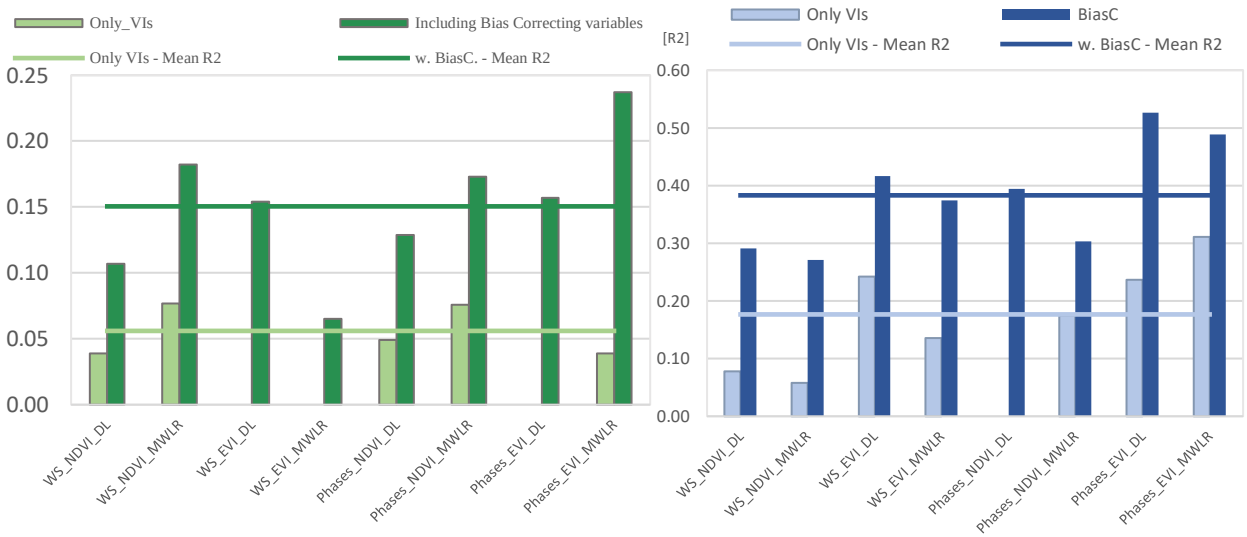


*Figure 45: Comparison between groups with and without bias correcting variables of R^2 from multiple regressions with grain yield for Haryana (**Left**) and Odisha (**Right**).*

7.4.2.3 RF Classification

The effect of including the bias correcting variables in the groups for the RF classification was less clear. In Haryana, the mean CVS increased 14% when included, while there was almost no difference in Odisha (2%) when including the bias correcting variables (Figure 46).

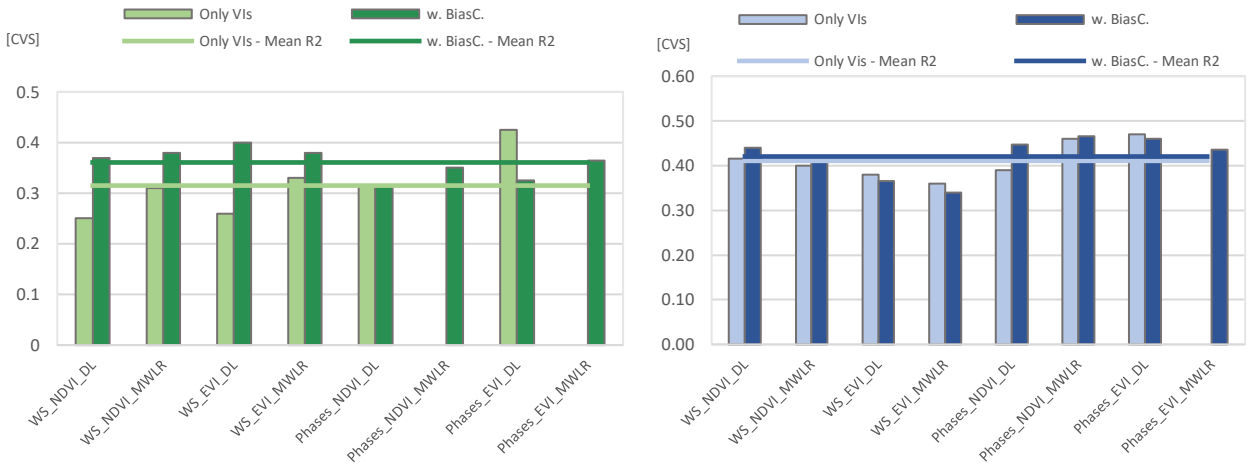


*Figure 46: Comparison between groups with and without bias correcting variables of CVS from RF classifications of grain yield for Haryana (**Left**) and Odisha (**Right**).*

7.4.3 Effect of the imperfect relation between total and grain yield

As established earlier (in 6.1 Yield data preparation p. 27), the grain yield does not perfectly correlate with the total yield. The total yield is the closest to what the satellite observes and the imperfect correlation with grain yield therefore poses a challenge for the efforts to estimate grain yield.

7.4.3.1 Linear Regression

The correlation with the individual variables found in the linear regression showed that a considerable higher fraction of the VI variables had a statistically significant relation with total yield compared to grain yield. This was however opposite for Odisha (Figure 19).

7.4.3.2 Multiple Regression

When running the multiple regressions for the four bias corrected phase-groups using total yield as the dependent variable instead of grain yield, they could on average explain 38% more of the yield variation in Haryana, while it had no effect in Odisha (Figure 47).

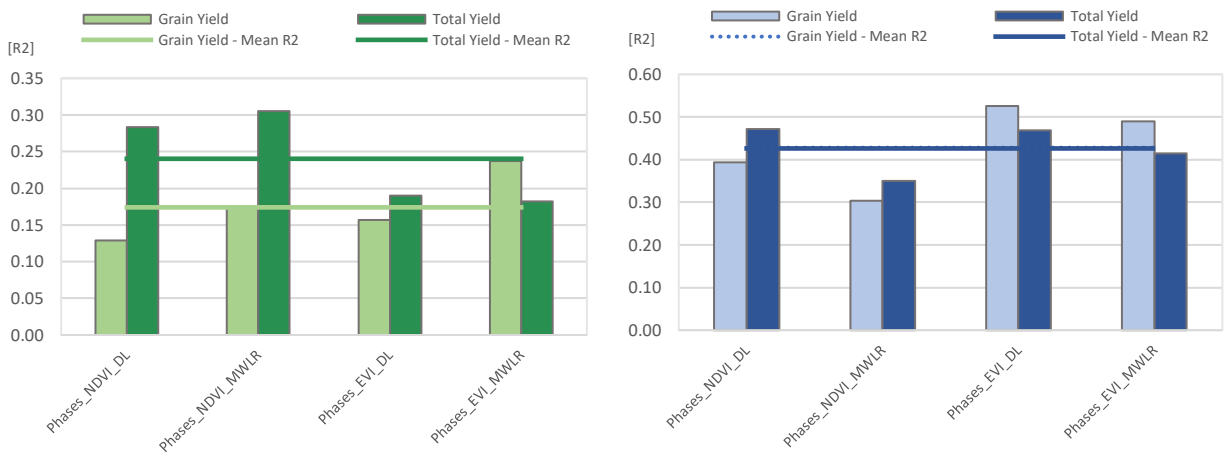


*Figure 47: Comparison of R^2 between having grain yield or total yield as the dependent variable in multiple regressions for Haryana (**Left**) and Odisha (**Right**).*

7.4.3.3 RF Classification

An RF Classification was run again with all VI variables (no ground variables etc.) and with discretized total yield as the categories (same number of categories as for grain yield). For Haryana the ability of the groups to correctly classify the samples was 0.41, an 8% increase, while it was 0.45 in Odisha, a decrease of 10% compared to the CVS for the grain yield classes (Figure 48).

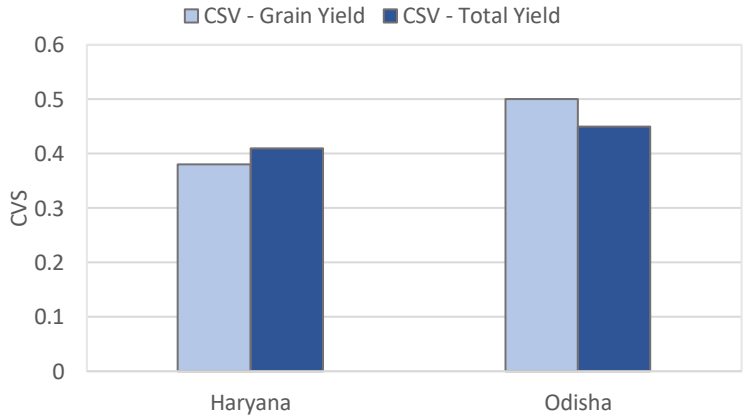


Figure 48: Comparison of CVS between RF classifications of grain yield or total yield.

8 Discussion

In this section, the methods and results of the study will be discussed. The discussion will be structures as follows: First the uncertainties in the method of creating the variables will be discussed, followed by a discussion of the three statistical tests. Then it will be discussed how the results answer the research questions. This will be followed by a discussion of what could have improved the results. It will then be discussed how our findings contribute to the research field, which will be followed by a discussion of recommended further research.

8.1 Discussion of uncertainties in the methods

Before assessing how the results of the study answered the research questions, the uncertainties in the method of creating the variables and assessing them in the three statistical tests will first be discussed to create a basis for the following discussions.

8.1.1 Creating the variables

In the processes of creating the variables used in the three statistical analyses, several sources of uncertainty occurred.

8.1.1.1 The VIs

NDVI is among the most used VIs for vegetation assessment and is considered reliable in its ability to estimate the biomass accumulation. The EVI is a modified version of the NDVI and is also widely used. The EVI is less affected by the soil background and less prone to saturation, which the low inter-field NDVI variation indicated might be a problem for the NDVI (Lillesand et al., 2015; Son et al., 2013; Lambert et al., 2017; Lobell et al., 2019; Burke & Lobell, 2017; Guan et al., 2018). Many other indices are included in similar studies for example GCVI, MTCI, LAI, NDVI705, NDVI740 (Lobell et al., 2019; Guan et al., 2018; Lambert et al., 2017; Lobell et al., 2018; Burke & Lobell, 2017). These could also have been included in this study. This would however have dramatically increased the number of variables and thus not given us the possibility to also differ between the other design options. The NDVI and EVI were generally considered reliable and supplemented each other well and was therefore deemed a good choice for this study.

One challenge of the chosen VIs is that they react differently to clouds and cloud shadows. An extra processing step was therefore needed for EVI to adjust for this. This however reduces the

ability to directly compare the two indices, as it cannot be assessed to what degree the differences between NDVI and EVI is because one VI is better at estimating the yield and what is because the efforts to remove the effect of positive cloud outliers in the EVI data was insufficient.

The process of removing positive outliers in the EVI timeseries was complicated due to the heavy influence by clouds. Though developing a method that took the variation of the neighbouring observations into account, this processing step still contributed with an amount of uncertainty, as correct observations sometimes appeared as positive outliers because the previous and succeeding observations all was similarly reduced by clouds. It is therefore recommended that future research develop this processing step further.

8.1.1.2 The smoothings

The timeseries were heavily affected by clouds and cloud shadow, which complicated the process of creating smoothed timeseries. This is problematic as the objective of the study is to be able to differentiate between individual fields which might only have subtle differences in the shape of the VI timeseries. The iterative MWLR smoothing appeared through visual inspection to be a good way to remove the effect of the clouds without cancelling out the inter-field differences. Smoothing to the upper envelope have prior to this study been observed to lead to overestimations of the VI in the start and end of the season as smoothed values can be dragged up by the high values of the previous and succeeding seasons (Kong et al., 2019). The short window is however expected to limit the effect of this in this study.

The number of iterative smoothings was set at 10, as in Lobell et al. (2019). To assess whether 10 was enough to approximate the upper envelope a test was done, comparing the correlation between a VI with 6 smoothings and 10 smoothings. The high R^2 revealed that more than six smoothings only added very limited to the upper-envelope fitting. The test did however not directly indicate if the 10^{th} smoothing is in fact at the upper envelope. Additionally, the test could have been performed more systematically to find the optimal number of smoothings

The DL smoothing is widely used and appeared to have a good fit with the visually inspected examples (Eklundh & Jönsson, 2017). It was much less dependent on the individual observations and its generic form was a good supplement to the MWLR as a smoothing type in the other end of the spectra. The implementation of the DL smoothing in GEE also created some

uncertainty as parameters, such as the inflection points and rate of increase and decrease are influenced by individual observations. This is however limited by the inserted daily values and the preceding MWLR smoothing.

8.1.1.3 Creating the phase-windows

The process of determining the phase windows based on the VIs is also associated with a degree of uncertainty. The smartphone pictures, taken a date where the VI is on the increasing part, were dominantly of crops in the VS stage and similarly for of the RMS stage in the decreasing part. Less dominant was the FRS around the peak of the VI timeseries. At best, only around 30% of the pictures, taken a date where the field had reached the peak of the VI timeseries, were classified as being in the FRS stage. This result also entails that if inspecting the crops around the time that the VI peaks, only 30% will be in the FRS phase. Whether it is possible to extract the variables with information for that phase alone is therefore questionable. However, there was expected to be a lot of noise in the picture data and the manual classification hereof and the ability to isolate the FRS phase will therefore likely be higher than what the 30% indicates. It is for example unlikely that many fields should be in the FRS phase after the VI timeseries has decreased 60%. The clear signal of the VS and RMS phase alone increases the confidence that the FRS phase is located in between them, around the peak of the VI timeseries. Efforts could be done in further research to find the optimal VI boundaries for each phase though it is not expected to have a large impact on the results.

8.1.1.4 The supplementary data

There are large sources of uncertainties related to the supplementary data.

Both temperature and precipitation had a low spatial resolution thus only allowing to differentiate between fields by the differing time windows created by the dynamic seasonality. In addition, the temperature data was heavily influenced by clouds.

The “money-spent” variable was only an estimate as described previously and the information was not available for all fields.

The rice variety and soil type were also not available for all fields and could only be included in the linear regression by isolating samples with certain varieties or soil types, often leading to only few included samples, and thus less reliable results. To include them in the multiple

regression analyses and RF classification they were ranked and numbered according to their average yield. This does however only reflect which variety or soil type that on average have a higher yield but not how much higher it was.

8.1.1.5 The temporal aggregation

In the temporal aggregations were also several potential sources of uncertainty, especially related to the temporal resolution of the satellite data. A timeseries with an observation every fifth day naturally lacks information on the days in between. The direct effect of this in this study is however expected to be limited due to the relatively slow development of the crops. A potential indirect bias from the five-day span between observations was however identified. This related to the potential jump when going from a VI boundary value to a date boundary. Inserting a value for each day by assuming a linear development of the VI between the observations is expected to have eliminated most of this uncertainty. It is however not able to remove the uncertainty from the numerical integration as an approximation of the area beneath the VI curve. If comparing two smoothings with known functions, analytical integration could have been applied and the uncertainty thereby reduced.

8.1.1.6 The spatial aggregation

The fields were on average 45m x 45m in Haryana and 22m x 22m in Odisha. This is small relative to the size of the pixels. Some fields would only be covered by a few pixels and have a high risk of being affected by edge-pixels.

Several measures were taken to reduce the uncertainty from the spatial aggregation. All the fields were drawn in manually with very high-resolution imagery as reference and the datapoints not connected to an obvious field was excluded. The effect of the edge pixels was reduced by creating a five-meter inner buffer and by taking the median value of the pixels. An inner buffer of 10 meters, as used in Lambert et al. (2017), would have had a greater effect, but was rejected as too many fields would have been erased due to the very small field.

8.1.2 The three statistical analyses

Three statistical analyses were made to assess the ability of the created variables to explain the inter-field yield variation. Each analysis does however contain certain limitations in doing so, which will be discussed in the following.

The methodological triangulation has led to more ambiguous results and therefore more conservative conclusions. This is however also considered a strength as the results are expected to be more reliable when assessed in multiple ways.

8.1.2.1 Linear Regression

All variables could individually only explain a minor part of the yield variation. Such low R^2 values does not create a good foundation for comparing the different design options, as the differences in correlations can be heavily influenced by randomness. As the comparisons of the design options are done across multiple variables the uncertainty is expected to be reduced. It will however still be acknowledged in the following discussion of the results.

An assumption of the linear regression is that the relationship between the variables is linear, which is often not the case. Upon visual inspection of several scatterplots of yields and VI variables, the low R^2 did not appear to be due to non-linear relationship, but merely because of weak association between the variables i.e. there did not appeared to be systematic biases in the residuals. The relation between yield and the VI variable could still be non-linear, but the uncertainty from this appears to be cancelled out by the low covariance (McGrew & Monroe, 2009).

8.1.2.2 Multiple Regression

Similar to the linear regression, multiple regression also assumes a linear relationship between the variables, which again might not be the case in reality.

To improve the models, efforts were done to only include variables that contributed to the model, by iteratively removing the variable with the highest p-value and running the analysis again until only significant variables remained. Additionally, multicollinearity was reduced by not including both “mean” and “integral” variables. There are however more steps that could have been taken to improve the reliability of the multiple regression models. The remaining variables could have been tested for redundancy and the residuals could have been tested for biases, including spatial biases (ArcGIS Pro, 2020).

8.1.2.3 RF Classification

Random forest is a powerful tool with high predictive accuracy. Compared to the other two tests, it allows for non-linear relations between the variables and is less sensitive to

multicollinearity. In this study it was used on discretised yield data. With a slight methodological alteration the RF Classification can also function on continuous data which might have been a better fit for this study, as the aim is to estimate the exact yield, which the five categories does not fully allow[12] (James et al., 2013; Boehmke & Greewell, 2020; Belgiu & Dragut, 2016).

Though the rationale behind the outcome of a single classification tree can be easily understood, the reasoning behind the outcome of the RF Classification cannot be intuitively interpreted. For both the linear and multiple regression analysis the relationship between the dependent variable and each independent variable can be assessed. This grants the possibility to check if the relations are scientifically sound i.e. whether the dependent and independent variables have the expected relationship. A check that is not possible with the RF Classification (Boehmke & Greewell, 2020).

A method that has been proposed to increase the accuracy of the RF classification, is to iteratively remove the least contributing features, much similar to the process in the multiple regression analysis. This might have increased the accuracy of the classifications, especially of those groups with many variables (Belgiu & Dragut, 2016).

An important point to note, is that the CVS does not consider, that in a random classification, some samples would also be correctly classified. This is referred to as the expected accuracy. With the relatively few categories used in this study, the correctly classified samples in a random classification would be a significant amount of the CVS. The Cohen´s Kappa Coefficient is a measure similar to the accuracy but adjusted for the expected accuracy and would therefore have been a beneficial measure to use in this study. In absence of this, it will in the following discussion be assumed that the actual accuracy is somewhat lower than the CVS (Lee et al., 2019).

8.2 Discussion of the results

In this part, it will be discussed how the results of the analyses answer the research questions. The question regarding the overall ability of the VI variables to explain the yield will be saved

[12] Though it's a slight methodological alteration to go from RF classification to RF regression, the practical implications can be large, which is why the RF Classification was used in this study.

for last and will include a comparison to similar studies and a discussion of the implications of the results for the research field and in relation to the aim of the study.

8.2.1 Input variables influencing the yield

The aim of this analysis was to get an understanding of what factors affected the yield, as this could be useful when afterwards estimating the yield (Research Question 1). From the results it however became apparent that a more thorough analysis would be necessary to explain the yield variation with input variables.

In the linear regression many of the chosen variables did not have a significant correlation with the yield and each variable could only explain very little of the two study sites´ yield variation. Similarly, the multiple regression could only explain 12% of the variation in Haryana and 17% in Odisha. The RF Classification was able to correctly classify 46% in Haryana and 40% in Odisha. This should however also be considered an optimistic score as the actual accuracy will be lower.

The generally low results are likely due to several aspects: Some input variables essential to the yield have likely not been included in this analysis. The uncertainties of the used data might have reduced the variables´ ability to explain the yield variation. And lastly, the variables should likely have been more specifically tailored to better estimate the yield. It could for example be the case, that it is not the mean temperature but rather the number of very warm days that is decisive, or that the very heavy rainfalls should have been omitted from the precipitation data etc.

8.2.2 Design options

An important aim of the study was to compare different design options to see which are preferable and to get an indication of the importance of choosing the best suited design options for the specific case (Research Question 3). The comparison was done across multiple variables and with three different tests. Though this might lead to less unambiguous results, it will increase the overall robustness of the results.

8.2.2.1 The VIs

In Haryana, the linear regression showed that more NDVI-based variables had a significant R^2 and that the significant variables on average had a higher R^2. The low R^2 did however not

provide a good basis for comparison. A more indicative result was that four NDVI variables could explain more than 10% while there was only one EVI variable that could explain more than 5%. In the multiple regressions, the mean R^2 of the significant groups was higher for NDVI, but the group with the highest R^2 had EVI-based variables. The RF classification also only had small differences, with slightly higher CVS for the EVI-based groups.

In Odisha, the linear regression showed almost no difference between NDVI and EVI variables. The multiple regression did however clearly show that the groups with EVI variables performed better. Less clear was the results of the RF classification, which showed a minor advantage to the NDVI.

When assessing the variable importance from the RF classification with all VIs for both study sites, the EVI-based variables had a slightly higher representation in the top 20 of most important variables.

Overall, no clear conclusions can be drawn on whether the NDVI or EVI is to be preferred. The clearest results were from the multiple regression in Odisha, were EVI was considerable higher. While it cannot be concluded from this result alone that EVI is to be preferred, the result does show that the choice of VI can lead to large differences and it can thus be concluded that it is important to assess several different VIs when estimating yield.

8.2.2.2 The choice of smoothing

In both Haryana and Odisha, the differences between the two smoothings were small for all three statistical tests and not consistently in favour of either. In Haryana the three groups with the highest R^2 in the multiple regression were all with MWLR smoothed variables, but the two groups with the highest CVS in the RF classification were DL smoothed groups. The remaining results for Haryana and the results of all three statistical test in Odisha were all less indicative. This either indicates that whether the smoothing closely follows the datapoints as the MWLR or has a more generic form as the DL does not have a decisive impact on the results, or that general results were not strong enough to detect the differences from the smoothings. As larger differences could be observed comparing other design options, it indicates the former.

8.2.2.3 The triggering measures

In both Haryana and Odisha, the variables with the highest R^2 in the linear regression were "length" and "integral". In Haryana the variables with these triggering measures on average also had the highest R^2, while they were surpassed by the EOS-variables in Odisha.

In the multiple regressions the "length" and "integral" were also contributing more often to the models, almost twice as often as the other variables. In the RF classifications using all VI variables, the 6 of the 20 most important variables had "integral" as triggering measure and 6 had "length". Another 6 of the most important variables had "mean" as triggering measure.

Overall, the results clearly showed differences between the triggering measures, with "integral" and "length" as the most suitable options. They did however also show that all triggering measures were used in a considerable part of the multiple regressions, indicating that they do provide useful, supplementary information of the crop season.

The "peak", which is most commonly used in similar studies, was generally less suitable to estimate yield according to results of this study, though not the worst performing triggering measure. This implies that if only using the "peak" as measure, some explanatory power might be lost (Lambert et al., 2017; Lobell et al., 2019; Guan et al., 2018).

8.2.2.4 The time period

The variables with the highest R^2 in the linear correlations were from the FRS phase for Haryana and the WS for Odisha. For both study sites the average R^2 of significant variables was higher for the variables from the FRS phase. This is however not a very robust result, due to the low R^2. The advantages of focussing on the phases were also primarily expected to be evident when used in combination with each other. The multiple regression reinforced this hypothesis, as the groups including the phase variables on average had 18% and 42% higher adjusted R^2 for Haryana and Odisha respectively. The RF classification also supported this, with a 15% higher CVS in Odisha though only slightly higher in Haryana. Additionally, the 16 most important variables in the RF classification with all VI variables were from one of the phases, while the WS had only 3 in top 20. The result is considered robust due to the consistency in the results and it clearly indicates that valuable information of the crop season can be obtained when taking the phenology of the crop into consideration when creating the VI variables. The

results also indicate that separating the phases using the relative a VI value did to at least some extent enable isolation of the specific crop phases.

8.2.2.5 The seasonality

Only a limited number of variables were calculated with the fixed seasonality, and only in Haryana. This decreases the reliability of the results compared to the other design options that were compared across more variables.

The linear regression did not result in any clear differences between the dynamic and fixed variables. The multiple regression showed 70% (non-biased corrected group) and 16% (bias corrected group) higher R^2 for the dynamic seasonality. Similarly, the RF classification showed slightly higher CVS for the group with variables determined with dynamic seasonality. The result indicates that the dynamic seasonality is a better design option when estimating inter-field yield variation. The differences were however not as clear as expected. Had the design option been compared over more VI variables, the differences might have been clearer.

8.2.3 Uncertainties and biases

Efforts were also done to assess the sources of uncertainty and bias in the study (Research Question 4). The results of these efforts will be discussed in the following.

8.2.3.1 Smoothing type or VI type

It was assessed how much agreement there were between variables which only differed from the type of smoothing, and afterwards for the choice of VI. For both Haryana and Odisha there was a markedly higher agreement between variables with differing smoothing type than between variables with variables with differing VI. This is even though the smoothings were chosen purposely to be on either end of the spectrum regarding how closely they follow the initial observations. The results thus indicate that the choice of VI is more decisive for the output and including more VIs in succeeding research would therefore be recommendable over including more different smoothings types. This is also supported by the results described previously, where only minor differences were found when comparing the three statistical test results of MWLR and DL variables. If the EVI variables consistently performed poorly in the three statistical tests, the results of the bias test could have been due to an inadequate removal of positive outliers in the extra processing step of the EVI. Had that been the case, it could not have been concluded that the VI generally is more decisive.

There is however the reservation that only two smoothing types and two VIs were used in this analysis. Had more been included the results would have been more robust.

8.2.3.2 Bias creating variables

The linear regression of isolated groups of samples revealed that considerable higher R^2 of significant variables could be achieved when differentiating between different rice varieties and soil types. The results for the majority of these groups were however not very robust, as only few of the VI variables returned significant correlations. In Odisha, isolating the samples with rice variety "12" drastically increased the R^2 of many of the VI variables, achieving R^2 up to 0.65 for total yield and 0.86 for grain yield. The consistent increase in the R^2 across multiple variables strongly indicates that the rice variety have a large influence on the results.

The multiple regressions where the rice variety, soil type and days to CCE had been included, consistently resulted in adjusted R^2 much higher than when not included. On average the R^2 increased 169% in Haryana and 117% in Odisha. For the RF classification the results were less clear but pointed in the same direction. It should be noted that these results are from including the ranked and numbered rice varieties and soil type, which are not perfect representations of the variables. Even better results might be achieved using more representative variables.

Overall, the results clearly indicate that correcting for biases in the yield data and differing between rice variety and soil type in the analyses can drastically improve the results.

From the result it can also be deducted that if other types of biases exist in the yield data it could be limiting the ability of the variables to explain the yield variation.

8.2.3.3 Mismatch between grain yield and total yield

The last source of uncertainty assessed in this study was a result of the mismatch between the total yield observed by the satellite and the grain yield that was to be estimated. This was especially evident in Haryana where the R^2 for the correlation between total yield and grain yield was only 0.22. In Odisha there was a much higher agreement, with an R^2 of 0.74.

On average across the three statistical tests, the VI variables in Haryana were considerably more able to explain the variation in total yield than grain yield. This was opposite for Odisha but with less consistent across the three statistical tests. Between the two study sites, the analyses from Odisha consistently showed higher results.

The results thus support that the mismatch can have a significant influence on the results and that the size of this challenge appears to be directly related to the agreement between the total yield and grain yield. It is therefore recommended for future studies that the correlation is tested prior to the analyses and that effort are taken to understand and reduce the mismatch.

8.2.4 How well the VI variables explain the yield variation

Here the overall results of how well the VI variables are able to explain yield variation will be discussed (Research Question 2). The results will be compared to similar studies, taking difference in approaches into account. Lastly, the results will be discussed in relation to the chosen school of research.

8.2.4.1 The results compared to similar studies

The linear regression of all the field samples showed that less than half the created VI variables had a significant correlation with the yield and that these significant VI variables could on average only explain a minor part of the yield variation (<5% for Haryana and <10% for Odisha). The best VI variables could explain 15% of the variation in total yield in both Haryana and Odisha, while the highest correlations with grain yield could explain 5% in Haryana and 20% in Odisha.

Compared to similar studies that estimate yield on field level, the results of this study are generally somewhat lower (***Table 5***). Though they all use linear regression to assess the variables, a direct comparison is difficult as there are multiple differences in analysis design, including both different crop types, satellite data, continents and VIs. The comparison should however give some indication of quality of this study´s results.

Table 5: *Overview of comparable studies.*

~R^2	Crop	Notes	VI	Reference
0.6-0.8	Cotton & Millet	Only able to obtain the high values when using a subset of the most homogeneous fields as samples (n<10)	NDVI & LAI	Lambert et al. (2017)
0.2-0.6	Maize & Sorghum			
0.25	Sorghum		NDVI & GCVI	Lobell et al. (2019)
0.3-0.4 (2014)	Maize		NDVI & GCVI	Burke & Lobell (2017)
0.15-0.2 (2015)				
0.27-0.33	Wheat		LAI & GCVI	Jain et al. (2016)
0.4	Rice	Only use 71 of the 255 available fields as they removed fields that might be influenced by nearby landcovers types.	NDVI, EVI & GCVI	Guan et al. (2018)
0.06		When including all samples		
0.69		When isolating rice varieties		

When isolating the samples with a certain rice variety or soil type in this study, the mean R^2 of the significant variables considerably improved for almost all sample groups. The isolated group of samples which gave the highest correlation with grain yield, returned an average R^2 of over 0.25 for the significant variables in Haryana and around 0.65 in Odisha, with the best being 0.35 in Haryana and 0.86 in Odisha. These are in the high end compared to similar studies. It should however be noted that the sample size of these is quite small and the result therefore less reliable. This was however also the case for several of the studies, especially Lambert et al. (2017) and Guan et al. (2018) when isolating the rice varieties. The consistently high correlations with grain yield across different types of VI variables when isolating the dominant rice variety in Odisha (variety "12") does however indicate that it is a more reliable result.

The multiple regression and RF classification allows for the variables to supplement each other. The majority of the groups of variables returned a significant result for both tests. Through the multiple regressions, the significant groups could on average explain 20% of the grain yield variation in Haryana and 26% in Odisha, while the best groups could explain 24% in Haryana and 53% in Odisha. The RF classification, which also allow for non-linear relationships between dependent and independent variables could for the best groups of variables correctly classify just less than half the samples, though without adjusting for the expected accuracy.

These results are less comparable to the similar studies but implies that the variables can supplement each other and thereby better estimate the yield.

8.2.4.2 The results and the school of research

The results of this study indicate that there is still some work to be done before the methods can be used for index insurance. If indices were created on the basis of the correlations found in this study the mean error would be too high and the insurance thus be too unreliable to benefit the farmers. The results do however also indicate that there are still substantial improvements to be gained and that the objective might therefore not be unachievable.

The alternative source of farm level yield data will in many cases be retrospective farm yield surveys which have been shown to very inaccurate. Considering this, the accuracy of the yield estimations found in this study suggests that the approach used here can be an effective and scalable way to identify yield gaps and assess the impacts of policy interventions. Measures that are both relevant for accelerating rural development (Lambert et al., 2017; Burke & Lobell, 2017; Lambert et al., 2018; Lobell et al. 2018).

8.2.5 Limiting factors

In this section, it will be discussed which aspects, aside from the already discussed uncertainties in the methods, that are limiting the ability of the VI variables to explain yield variation and what measures could have been done to improve them.

8.2.5.1 The data used

One of the clearest results of the study is that adjusting the yield data for biases considerably improves the accuracy of the yield estimates. It is likely that there are other biases, which were not corrected for. This has likely been a limiting factor for the yield estimates. Given the improvement from adjusting to the identified biases, there might be potential to improve the results considerably, if the yield data can be further corrected for biases.

The other main source of data used in the study, the Sentinel-2 satellite data, might also have been a limiting factor. Though the spatial resolution of 10 meters is considered high, the very small field sizes in especially Odisha, meant that only few pixels could have covered each field and that the effect of edge pixels potentially could have been large. Though measures were taken to reduce the effect of this, it could have influenced the VI signal from the fields. It would

however not necessarily be beneficial for the analysis to trade off the temporal, radiometric or spectral resolution for a higher spatial resolution. A lower temporal resolution could amplify the challenges of removing the effect of clouds from the data, and as the VI values between fields were observed to be relatively close, a reduction in the sensitivity of the satellite could also degrade the results. It could however be worth analysing if the very high spatial resolution Planet data could achieve better results.

8.2.5.2 Field size

Beside the challenge of the spatial resolution of the satellite images, the small field size might also be limiting in other ways. The GPS accuracy of the yield data points, and the geolocation uncertainty of the satellite data do create some uncertainty in the analyses. The risk of a mismatch between which field the CCE is done for and where the satellite measures, is increased the smaller the size of the fields (Guan et al., 2018; ESA, 2020). The effect of the field size was not assessed in this study. The fields were generally smaller in Odisha were the results were better, but this is more likely to be attributed to stronger correlation between total yield and grain yield.

8.2.5.3 The clouds

The satellite data in the study areas was heavily influenced by clouds and cloud shadows, which created challenges for the data processing. The implications of this on the results is difficult to assess. Active sensors are not affected by clouds in the same way and might therefore seem a viable alternative. Guan et al. (2018) did however find that radar data did not improve their yield estimations. The similarities between the two smoothings used in this study also indicate that the effect of the clouds was not the most decisive factor.

8.2.5.4 Fertilizer

Differences in rates of fertilizer application might also complicate the yield estimations, as sufficient access to essential nutrients might increase the grain yield without proportionally increasing the VI signal observed by the satellite (Lambert et al., 2017). For a subset of the fields in Haryana, the yield dataset included information on whether fertilizer was applied. This showed that fertilizer had been applied to 81% of the fields. The limited amount of data did however not allow for the inclusion of this aspect in analyses. The fertilizer data was unfortunately not available for Odisha. The lower per capita income in Odisha might result in

lower fertilizer application rates and could thus be an explanation for why the total yield and grain yield had a stronger correlation in Odisha compared to Haryana. In future studies the effect of fertiliser application would be an interesting aspect to include, as the superior results from the study site in Odisha indicate that considerably gains in index accuracy can be achieved if the mismatch between grain yield and total yield can be reduced.

8.2.5.5 The causes of crop damage

A subset of the yield data also had information about the cause of crop damage, either self-reported by the farmer or classified based on the smartphone-images. The data reveals a diverse range of damage causes, including rain, wind, heat, fire and pest as the most frequently occurring causes. Several of these causes might be difficult to detect with the VI variables and might thus have been a limiting factor in this study. If the plants have been overturned by strong winds, it could destroy the grains, while the satellite still detects very green vegetation. It is therefore recommended that efforts are made to understand the effect of these damages on the VI values and to develop ways that can specifically detect damages from these sources.

8.2.5.6 Including more years

The analyses in this study were only done for 2019. As seen in the differing results between years in Burke & Lobell (2017) there can be interannual differences. Analyses with more years are thus another way to increase the robustness of the results.

8.2.5.7 Localised estimations

The differences between the two study sites indicate that though the crop might be the same, it is important that yield estimations are done on the basis of localised relations between yield and VI variables. The results also indicate that the most suitable design options might differ between locations.

8.3 Further research

The main recommendations for further research are summarised here.

- The very high spatial resolution and temporal coverage of the Planet data might improve the yield estimations on very small fields. A similar assessment as this, but with the use of Planet data is therefore recommended.

- The results indicated that analysing different rice varieties individually could improve the accuracy of the yield estimations considerably. It is recommended to consider this in further research and to verify this result on a larger dataset to get a more robust result.

- Based on the results, it is more important to include more VIs than to try more different smoothings, which is therefore recommended for succeeding studies.

- It is recommended that efforts to assess and understand the mismatch between grain yield and total yield are done prior to the analyses, as these most likely have a substantial impact on the results. Assessing the effect of fertilizer application rates is recommended as a starting point for these efforts.

- Isolating the season in phases according to the crop phenology does improve the results, and it is therefore recommended that further studies incorporate this and try to find methods to more accurately isolate the crop phases.

- It is also recommended that more effort is done to estimate the response of the VI values to specific types of damages, so that the yield estimates can be better tailored to the diverse causes of crop damage.

- Lastly it is recommended that yield estimates are done on localised conditions and that the analyses are done across more years to increase the robustness of the results.

9 Conclusions

The aim of the study has been to assess the ability of Sentinel-2 derived vegetation indices to explain inter-field yield variation in paddy rice and to systematically compare the suitability of selected design options. This has been done by creating VI variables for almost all the different combinations of the selected design options and assessing them through linear regression, multiple regression and RF Classification.

The preliminary analyses of input variables show that the selected variables can only explain a minor part of the yield variation. This can either be due to uncertainty in the data sources and applied methods, a need of more specifically tailored variables or because some important variables were not included.

The VI variables can generally only explain a small part of the variation in farm level yield, less than comparable studies. If only including a subset of the samples according to rice variety or soil type, the results significantly improve and are in the high end compared to similar studies. The few samples included and the inconsistency across the different variables does however reduce the robustness of the result. Isolating the dominant rice variety ("12") in Odisha does however result in consistently higher R^2, with the highest being 0.86.

The multiple regression and RF Classification have revealed the benefits of combining multiple variables and the best groups of variables can explain 24% of the grain yield variation in Haryana and 53% in Odisha and they can correctly classify 46% of the samples in Haryana and 47% in Odisha.

The assessment of the design options shows only small differences between the two smoothing types and not consistently in favour of either. The choice of VI creates larger differences, but the results are inconclusive on whether NDVI or EVI was more suitable. The assessment of the triggering measures suggests that the "integral" and "length" are better able to capture the inter-field yield variation, but that all triggering measures can contribute with information in the multiple regression analyses and RF Classifications. The variables from the phenologically tailored phases did contribute significantly to the explanatory ability of the multiple regressions and RF Classifications. Though less robust, the assessment also indicates that the dynamic seasonality is to be preferred over the fixed.

When assessing the sources of uncertainty in the results, the choice of VI is considerably more decisive than the smoothing type. The results also suggest that including the bias correcting variables significantly improves the results and that the mismatch between total yield and grain yield are decisive for the results.

The study generally finds that the VI variables obtained through the used methods cannot sufficiently capture the inter-field yield variation to use the approach for index insurance of individual paddy rice fields. The study does however identify and recommend several ways to potentially achieve significant gains in accuracy and therefore concludes that the objective might not be unachievable. Lastly, the result of this analyses indicate that the used methods can pose an effective and scalable way to identify yield gaps and specifically target and evaluate rural development efforts.

10 References

ArcGIS Pro, 2020. What they don't tell you about regression analysis. Available here: https://pro.arcgis.com/en/pro-app/tool-reference%20/spatial-statistics/what-they-don-t-tell-you-about-regression-analysis.htm. Last accessed 22/06/2020.

Azzari, G.; Jain, M.; Lobell, D. B. 2017. Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries. Remote Sensing of Environment 202 (2017) 129–141.

BD 2020, Business Dictionary - -Definition Morale Hazard. Available here: http://www.businessdictionary.com/definition/morale-hazard.html Last accessed 26/06/2020.

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 (2006) 321–334.

Bekhet, A. K. & Zauszniewski, J. A. 2012. Methodological Triangulation: An Approach to Understanding Data. Nurse Res. 2012;20(2):40-43.

Belgiu, M. & Dragut, L. 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing 114 (2016) 24–31.

Boehmke, B. & Greewell, B. 2020. Hands-On Machine Learning with R. Available here: https://bradleyboehmke.github.io/HOML/. Last accessed 17/06/2020.

Burke, M. & Lobell, D. B. 2017. "Satellite-based assessment of yield variation and its determinants in smallholder African systems. PNAS j February 28, 2017 j vol. 114 j no. 9 j 2189–2194.

Carter, M.; Janvry, A.; Sadoulet, E.; Sarris, A. 2017. Index Insurance for Developing Country Agriculture: A Reassessment. Annual Review of Resource Economics.

CGIAR 2013. Fourth Edition Rice Almanac, Source Book for One of the Most Important Economic Activities on Earth. Research program on Rice, Global Rice Science Partnership.

Chantarat, S.; Mude, A. G.; Barret, C. B.; Carter, M. R. 2013. Designing Index-Based Livestock Insurance for Managing Asset Risk in Northern Kenya. The Journal of Risk and Insurance, 2013, Vol. 80, No. 1, 205-237.

Cheston S. 2018, ACCION - Inclusive Insurance: Closing the Protection Gap for Emerging Customers. A joint report from the Center for Financial Inclusion at Accion and the Institute of International Finance. Available here: https://content.centerforfinancialinclusion.org/wpcontent/uploads/sites/2/2018/08/Inclusive-Insurance-Final-2018.06.13.pdf Last accessed 26/06/2020.

CLIMATE-DATA.ORGa, 2020. Panchkula Climate(India) – Data 1982-2012- Panchkula weather by month // Weather averages. Available here: https://en.climate-data.org/asia/india/haryana/panchkula-56586/ Last accessed 26/06/2020.

CLIMATE-DATA.ORGb. 2020. Bhubaneswar Climate(India) – Data 1982-2012 -Bhubane weather by month // Weather averages https://en.climate-data.org/asia/india/odisha/bhubaneswar-5756/ Last accessed 26/06/2020.

Dong J. and Xiao X. 2016. Evolution of regional to global paddy rice mapping methods: A review. Department of Microbiology and Plant Biology, Center for Spatial Analysis, University of Oklahoma.

Eklundh, L. & Jönsson, P. 2017. TIMESAT 3.3 with seasonal trend decomposition and parallel processing Software Manual. Available at http://www.nateko.lu.se/TIMESAT/. Last accessed 01/06/2020.

Enenkel, M.; Osgood, M.; Powell, B. 2018. The Added Value of Satellite Soil Moisture for Agricultural Index Insurance. Remote Sensing of Hydrometeorological Hazards, chp. 4.

ESA, 2020. Performance. Sentinel-2 MSI – Calibration and Validation. The European Space Agency. Available here: https://earth.esa.int/web/sentinel/technical-guides/sentinel-2-msi/performance Last accessed 25/06/2020.

ESA, 2020a. Sentinel Overview- Missions. Available here: https://sentinel.esa.int/web/sentinel/missions Last accessed 26/06/2020.

ESA. 2015. Sentinel-2 User Handbook- European Commission. Available here: https://earth.esa.int/documents/247904/685211/Sentinel-2_User_Handbook Last accessed 26/06/2020.

ESA. 2020b. Level-2A Algorithm Overview- Technical Guides. Available here: https://earth.esa.int/web/sentinel/technical-guides/sentinel-2-msi/level-2a/algorithm Last accessed 26/06/2020.

Flatnes, J. E.; Carter, M. R.; Mercovich, R. 2019. Improving the quality of index insurance with a satellite-based conditional audit contract.

Funk C.; Peterso P.; Landsfeld M.; Pedreros D.; Verdin N.; Shukla S.; Husak G.; Rowland.; Harrison L.; Hoell A. & Michaelsen J. 2015. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. Available here: https://www.nature.com/articles/sdata201566.pdf Last accessed 26/06/2020.

GIFF, 2020. What is index insurance. World bank group. The Global Index Insurance Facility. Available here: https://www.indexinsuranceforum.org/faq/what-index-insurance Last accessed 26/06/2020.

GIIF, 2019. Frequently Asked Questions. Global Index Insurance Facility. Available here: https://www.indexinsuranceforum.org/faq-page. Last accessed 05/06/20.

GIZ, 2016. Innovations and Emerging Trends in Agricultural Insurance. How can we transfer natural risks out of rural livelihoods to empower and protect people? Deutsche Gesellschaft für Internationale Zusammenarbeit.

GIZ, 2017. Inclusive Insurance and the Sustainable Development Goals. How insurance contributes to the 2030 Agenda for Sustainable Development. Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ).

GOD, 2020. Population – Odisha. Government of Odisha. Available here: https://www.odisha.gov.in/ - Last assessed 28/06/2020.

GOH, 2020. Demography - Government of Haryana. Available here: https://haryana.gov.in/demography/ - Last assessed 28/06/2020.

Greatrex, H.; Hansen, J.; Garvin, S.; Diro, R.; Blakeley, S.; Guen, M.; Rao, K.; Osgood, D. 2015. Scaling up index insurance for smallholder farmers: Recent evidence and insights. CCAFS Report No. 14. CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS).

Guan K.; Li Z.; Rao L. N.; Gao F.; Xie D.i.; Hien N. T., and Zeng Z. 2018. Mapping Paddy Rice Area and Yields Over Thai Binh Province in Viet Nam From MODIS, Landsat, and ALOS-2/PALSAR-2

Guan, Z. L. K.; Rao, F. G. L. N.; Xie, N. T. H. D.; Zeng, Z. 2018. Mapping Paddy Rice Area and Yields Over Thai Binh Province in Viet Nam From MODIS, Landsat, and ALOS-2/PALSAR-2. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Vol. 11, No. 7, July 2018.

Hansen, J.; Hellin, J.; Rosenstock, T.; Fisher, E.; Cairns, J.; Stirling, C.; Lamanna, C. Van Etten, J.; Rose, A.; Campell, B. 2017. Climate risk management and rural poverty reduction. Agricultural Systems 172 (2019) 28–46.

Hufkens, K.; Melaas, E. K.; Mann, M. L.; Foster, T.; Ceballos, F.; Robles, M.; Kramer. B. 2019. Monitoring crop phenology using a smartphone based near-surface remote sensing approach. Agricultural and Forest Meteorology 265 (2019) 327–337

IFAD, 2017. Remote sensing for index insurance Findings and lessons learned for smallholder agriculture. International Fund of Agricultural Development (IFAD).

Investopedia 2020. Economy – Economics – Adverse Selection. Available here: https://www.investopedia.com/terms/a/adverseselection.asp Last accessed 26/06/2020.

IPCC, 2018. Intergovernmental Panel on Climate Change - Global warming of 1.5C.

IRMI 2020. Glossary - Morale Hazard. International Risk Management Institute Available here: https://www.irmi.com/term/insurance-definitions/morale-hazard Last accessed 26/06/2020.

Ivanic, M. & Martin, W. 2018. Sectoral Productivity Growth and Poverty Reduction: National and Global Impacts. World Development 109 (2018) 429–439.

Jain, M.; Srivastava, A. K.; Balwinder-Singh, Joon, R. K.; McDonald, A.; Royal, K.; Lisaius, M. C.; Lobell, D. B. Mapping Smallholder Wheat Yields and Sowing Dates Using Micro-Satellite Data. Remote Sens. 2016, 8, 860.

James, G.; Witten, D.; Hastie, T.; Tibshirani, R. 2013. An Introduction to Statistical Learning: with Applications in R. Springer Texts in Statistics.

Klish, A. & Atzberger, C. 2016. Operational Drought Monitoring in Kenya Using MODIS NDVI Time Series. Remote Sens. 2016, 8, 267.

Kong, D.; Zhang, Y.; Gu, X.; Wang, D. 2019. A robust method for reconstructing global MODIS EVI time series on the Google Earth Engine. ISPRS Journal of Photogrammetry and Remote Sensing 155 (2019) 13–24.

Lambert, MJ.; Blaes, X.; Traoré, P. S.; Defourny, P. 2017. Estimate yield at parcel level from S2 time serie in sub-Saharan smallholder farming systems.

Lambert, MJ.; Traoré, P. C. S.; Blaes, X.; Baret, P.; Defourny, P. 2018 Estimating smallholder crops production at village level from Sentinel-2 time series in Mali's cotton belt. Remote Sensing of Environment 216 (2018) 647–657.

Lee, M. R.; Sankar, V.; Hammer, A.; Kennedy, W. G.; Barb, J. J.; McQueen, P. G.; Leggio, L. 2019. Using Machine Learning to Classify Individuals with Alcohol Use Disorder Based on Treatment Seeking Status. EClinicalMedicine 12 (2019) 70–78.

Lillesand, T. M.; Kiefer, R. W.; Cipman, J. W. 2015. Remote Sensing and Image Interpretation. 7th edition. Wiley.

Liu, Y. & Myers, R. J. 2016. The Dynamics of Microinsurance Demand In Developing Countries Under Liquidity Constraints And Insurer Default Risk. The Journal of Risk and Insurance. 83, No. 1, 121–138.

Lobell, D. B.; Azzari, G.; Burke, M.; Gourlay, S.; Jin, Z.; Kilic, T.; Murray, S. 2018. Eyes in the Sky, Boots on the Ground Assessing Satellite- and Ground-Based Approaches to Crop Yield Measurement and Analysis in Uganda.

Lobell, D. B.; Di Tommaso, S.; You, C.; Djima, I. Y.; Burke, M.; Kilic, T. 2019. Sight for Sorghums: Comparisons of Satellite- and Ground-Based Sorghum Yield Estimates in Mali. Remote Sens. 2020, 12, 100.

Lobell, D. B.; Thau, D.; Seifert, C.; Engle, E.; Little, B. 2015. A scalable satellite-based crop yield mapper. Remote Sensing of Environment 164 (2015) 324–333.

Makaudze, E. M. & Miranda, M. J. 2010. Catastrophic drought insurance based on the remotely sensed normalised difference vegetation index for smallholder farmers in Zimbabwe. Agrekon, 49:4, 418-432.

McGrew, J. C. & Monroe, C. B. 2009. Statistical Problem Solving in Geography. Second edition.

Miranda, J. M. & Farrin, K. 2012. Index Insurance for Developing Countries. Applied Economic Perspectives and Policy, Vol. 34, No. 3 (Autumn 2012), pp. 391-427.

Morel, J.; Todoroff, P.; Bégué, A.; Bury, A.; Martiné, JF.; 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 Sens. 2014, 6, 6620-6635.

NASA, 2020. TERRA The EOS Flagship – About terra. Available here: https://terra.nasa.gov/about Last accessed 26/06/2020.

Osgood, D.; Powell, B.; Diro, R.; Farah, C.; Enenkel, M.; Brown, M. E.; Husak, G.; Blakeley, S. L.; Hoffman, L.; McCarty, J. L. 2018. "Farmer Perception, Recollection, and Remote Sensing in Weather Index Insurance: An Ethiopia Case Study". Remote Sens. 2018, 10, 1887.

Pasimen M. Ri.; *Valente D.*; *Semeraro T.*; *Petrosillo I. and Zurlin G, 2019. Anthropogenic Landscapes: A Scientific and Social Challenge. University of Salento, Lecce, Italy.* Available here: https://www.sciencedirect.com/science/article/pii/B9780124095489106025

Platteau, JP.; Bock, O.; Gelade, W. 2017. The Demand for Microinsurance: A Literature Review. World development Vol. 94, pp. 139–156.

Rosema, A.; Huystee, J van; Foppes S.; Woerd J. van der; Klaassen, E.; Barendse, J.; Asseldonk, M. van; Dubreuil, M.; Régent, S.; Weber, S.; Karaa, A.; Reusche, G.; Goslinga, R.; Mbaka, M.; Gosselink, F.; Leftley, R.; Kyokunda, J.; Kakweza, J.; Lynch, R.; Stigter, K. 2014 "FESA Micro-

insurance: Crop insurance reaching every farmer in Africa", Scientific Final Report of Millennium Agreements Project no. 38. Prepared and published by EARS Earth Environment Monitoring BV, Delft, The Netherlands.

Son, N. T.; Chen, C. F.; Chen, C. R.; Chang, L. Y.; Duc, H. N.; Nguyen, L. D. 2013. Prediction of rice crop yield using MODIS EVI−LAI data in the Mekong Delta, Vietnam. International Journal of Remote Sensing, 34:20, 7275-7292.

Statista, 2020a - Per capita income across Haryana in India from financial year 2012 to 2017, with estimates until 2019. Available here: https://www.statista.com/statistics/1116827/india-per-capita-income-haryana/ - Last accessed 26/06/2020.

Statista, 2020b - Per capita income across Odisha in India from financial year 2012 to 2017, with estimates until 2019. Available here: https://www.statista.com/statistics/1117632/india-per-capita-income-odisha/ Last accessed 26/06/2020.

The World Bank Group, 2018. How Technology Can Make Insurance More Inclusive. Fintech Note | No. 2.

UN, 2019. The Sustainable Development Goals Report. United Nations New York, 2019.

USAID 2020. CHIRPS: Rainfall Estimates from Rain Gauge and Satellite Observations. Available here: https://www.chc.ucsb.edu/data/chirps Last accessed 26/06/2020.

USGS 2020. Data Smoothing - Reducing the "Noise" in NDVI - United States Geological Survey. Available here: https://www.usgs.gov/land-resources/eros/phenology/science/data-smoothing-reducing-noise-ndvi?qt-science_center_objects=0#qt-science_center_objects Last accessed 26/06/2020.

USGS, 2020. MOD11A2 v006 - MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1 km SIN Grid. Available here: https://lpdaac.usgs.gov/products/mod11a2v006/ Last accessed 26/06/2020.

11 Appendix

11.1 Inserting a value every day

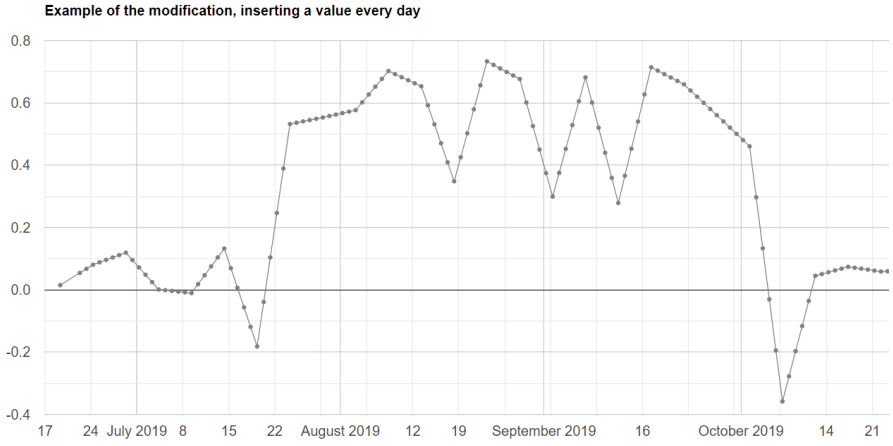


Figure 49: Example of an unsmoothed NDVI timeseries were a value has been inserted for every day.

11.2 Rice variety

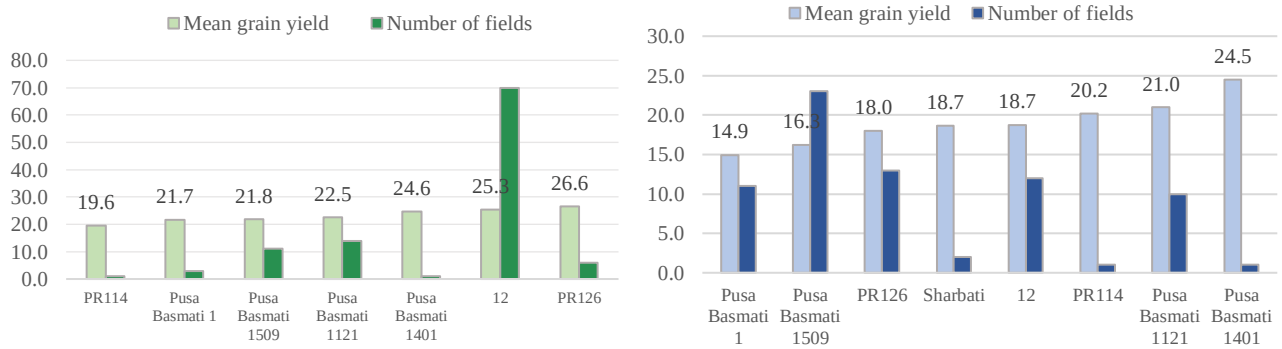


Figure 50: The bar charts shows the number of fields with the different rice varieties (left) and the average grain yield for each variety (right), for both Haryana (top) and Odisha (bottom).

11.3 Soil type

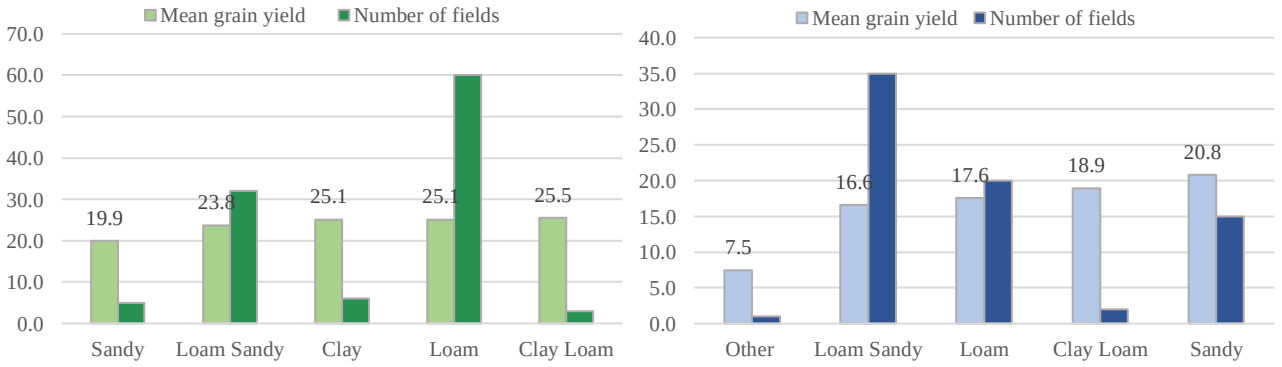


Figure 51: The bar charts shows the number of fields with the different soil types (left) and the average grain yield for soil type (right), for both Haryana (top) and Odisha (bottom).

11.4 Overview of the triggering measures

Table 6: *Overview of created variables in Haryana.*

Haryana							
WS		VS		FRS		RMS	
NDVI							
MWLR	DL	MWLR	DL	MWLR	DL	MWLR	DL
Peak value	*Peak*	*Sum*	*Sum*	*Sum*	*Sum*	*Sum*	*Sum*
Sum of values	*Sum*	*Preci-mean*	*Length*	*Preci-mean*	*Length*	*Preci-mean*	*Length*
Length	~~*Eos*~~	*Preci_sum*	*Mean*	*Preci_sum*	*Mean*	*Preci_sum*	*Mean*
SoS	*Sos*	*Temp_mean*		*Temp_mean*	*Eos*	*Temp_mean*	
EoS	*Length*	*Length*		*Length*	*Sos*	*Length*	
Precip mean		*Mean*		*Mean*		*Mean*	
Precip sum		*Mean_fix*		*Eos*		*Mean_fix*	
Temp mean		*Sum_fix*		*Sos*		*Sum_fix*	
Peak_fix				*Mean_fix*			
Sum_fix				*Sum_fix*			
EVI							
MWLR	DL	MWLR	DL	MWLR	DL	MWLR	DL
Peak value	*Peak*	*Sum*	*Sum*	*Sum*	*Sum*	*Sum*	*Sum*
Sum of values	*Sum*	*Temp_mean*	*Length*	*Temp_mean*	*Length*	*Temp_mean*	*Length*
Length	~~*Eos*~~	*Length*	*Mean*	*Length*	*Mean*	*Length*	*Mean*
SoS	*Sos*	*Mean*		*Mean*	*Eos*	*Mean*	
EoS	*Length*			*Eos*	*Sos*		
Temp mean				*Sos*			

Table 7: *Overview of created variables in Odisha*

Odisha							
WS		VS		FRS		RMS	
NDVI							
MWLR	DL	MWLR	DL	MWLR	DL	MWLR	DL
Peak value	*Peak*	*Sum*	*Sum*	*Sum*	*Sum*	*Sum*	*Sum*
Sum of values	*Sum*	*Length*	*Length*	*Length*	*Length*	*Length*	*Length*
Length	~~*Eos*~~	*Mean*	*Mean*	*Mean*	*Mean*	~~*Mean*~~	*Mean*
SoS	*Sos*			*Eos*	*Eos*		
EoS	*Length*			*Sos*	*Sos*		
EVI							
MWLR	DL	MWLR	DL	MWLR	DL	MWLR	DL
Peak value	*Peak*	*Sum*	*Sum*	*Sum*	*Sum*	*Sum*	*Sum*
Sum of values	*Sum*	*Length*	*Length*	*Length*	*Length*	*Length*	*Length*
Length	~~*Eos*~~	*Mean*	*Mean*	*Mean*	*Mean*	*Mean*	*Mean*
SoS	*Sos*			*Eos*	*Eos*		
EoS	*Length*			*Sos*	*Sos*		

11.5 Overview of grouped variables

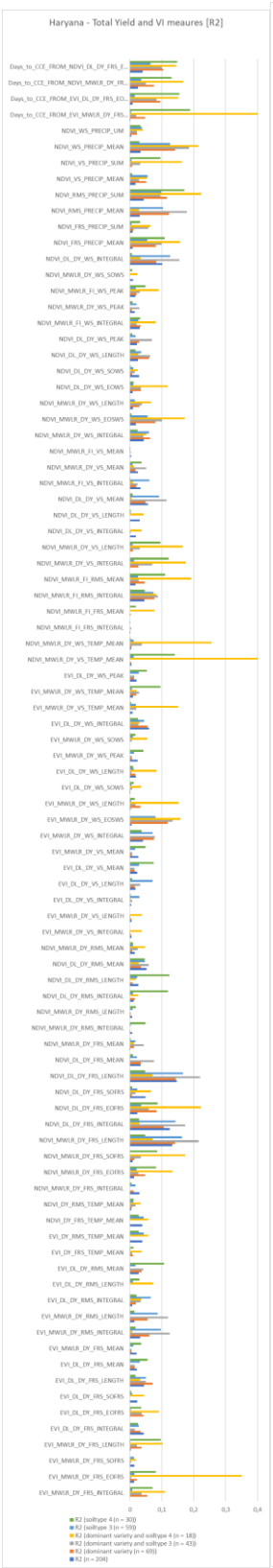
Table 8: *Overview of the grouped variables in Haryana and Odisha.*

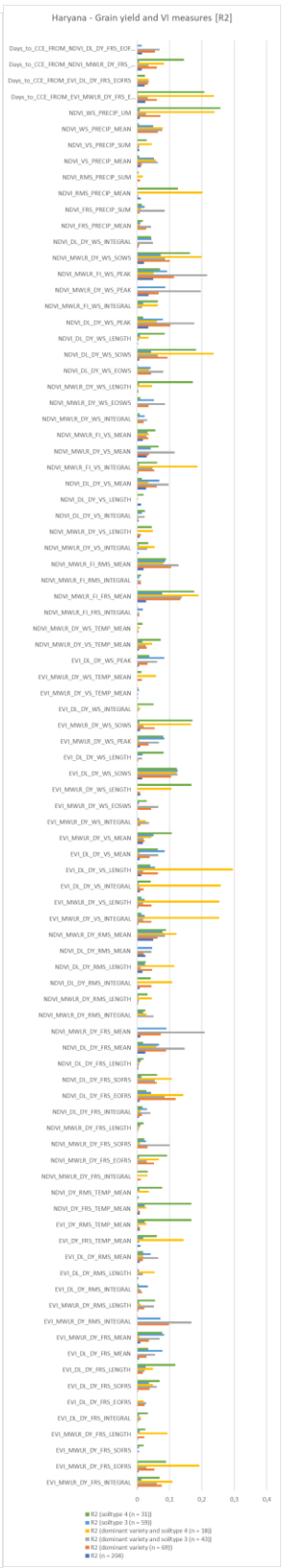
HARYANA	*NDVI_DL*	*NDVI_MWLR*	*EVI_DL*	*EVI_MWLR*
WS_VI	NDVI_DL_DY_WS_INTEGRAL NDVI_DL_DY_WS_LENGTH NDVI_DL_DY_WS_PEAK NDVI_DL_DY_WS_SOWS	NDVI_MWLR_DY_WS_EOSWS NDVI_MWLR_DY_WS_INTEGRAL NDVI_MWLR_DY_WS_LENGTH NDVI_MWLR_DY_WS_PEAK NDVI_MWLR_DY_WS_SOWS	EVI_DL_DY_WS_INTEGRAL EVI_DL_DY_WS_LENGTH EVI_DL_DY_WS_PEAK EVI_DL_DY_WS_SOWS	EVI_MWLR_DY_WS_EOSWS EVI_MWLR_DY_WS_INTEGRAL EVI_MWLR_DY_WS_LENGTH EVI_MWLR_DY_WS_PEAK EVI_MWLR_DY_WS_SOWS
WS_VI_BiasC	Variety_number SoilType_number NDVI_DL_DY_WS_EOWS NDVI_DL_DY_WS_INTEGRAL NDVI_DL_DY_WS_LENGTH NDVI_DL_DY_WS_PEAK NDVI_DL_DY_WS_SOWS Days_to CCE_fromSoWS	Variety_number SoilType_number Days_to_CCE_from_SoWS NDVI_MWLR_DY_WS_EOSWS NDVI_MWLR_DY_WS_INTEGRAL NDVI_MWLR_DY_WS_LENGTH NDVI_MWLR_DY_WS_PEAK NDVI_MWLR_DY_WS_SOWS	Variety_number SoilType_number Days_to_CCE_from_SoWS EVI_DL_DY_WS_INTEGRAL EVI_DL_DY_WS_LENGTH EVI_DL_DY_WS_PEAK EVI_DL_DY_WS_SOWS	Variety_number SoilType_number Days_to_CCE_from_SoWS EVI_MWLR_DY_WS_EOSWS EVI_MWLR_DY_WS_INTEGRAL EVI_MWLR_DY_WS_LENGTH EVI_MWLR_DY_WS_PEAK EVI_MWLR_DY_WS_SOWS
Fixed		NDVI_MWLR_FI_FRS_INTEGRAL NDVI_MWLR_FI_RMS_INTEGRAL NDVI_MWLR_FI_VS_INTEGRAL NDVI_MWLR_FI_WS_INTEGRAL NDVI_MWLR_FI_WS_PEAK		
Fixed_BiasC		Variety_number, SoilType_number DOY of CCE NDVI_MWLR_FI_FRS_INTEGRAL NDVI_MWLR_FI_RMS_INTEGRAL NDVI_MWLR_FI_VS_INTEGRAL NDVI_MWLR_FI_WS_INTEGRAL NDVI_MWLR_FI_WS_PEAK		
Phases_VI	NDVI_DL_DY_FRS_EOFRS NDVI_DL_DY_FRS_INTEGRAL NDVI_DL_DY_FRS_LENGTH NDVI_DL_DY_FRS_SOFRS NDVI_DL_DY_RMS_LENGTH NDVI_DL_DY_RMS_MEAN NDVI_DL_DY_VS_LENGTH NDVI_DL_DY_VS_MEAN NDVI_DL_DY_WS_INTEGRAL NDVI_DL_DY_WS_LENGTH NDVI_DL_DY_WS_PEAK NDVI_DL_DY_WS_SOWS	NDVI_MWLR_DY_FRS_EOFRS NDVI_MWLR_DY_FRS_INTEGRAL NDVI_MWLR_DY_FRS_LENGTH NDVI_MWLR_DY_FRS_SOFRS NDVI_MWLR_DY_RMS_LENGTH NDVI_MWLR_DY_RMS_MEAN NDVI_MWLR_DY_VS_LENGTH NDVI_MWLR_DY_VS_MEAN NDVI_MWLR_DY_WS_EOSWS NDVI_MWLR_DY_WS_INTEGRAL NDVI_MWLR_DY_WS_LENGTH NDVI_MWLR_DY_WS_PEAK NDVI_MWLR_DY_WS_SOWS	EVI_DL_DY_FRS_EOFRS EVI_DL_DY_FRS_INTEGRAL EVI_DL_DY_FRS_LENGTH EVI_DL_DY_FRS_SOFRS EVI_DL_DY_RMS_LENGTH EVI_DL_DY_RMS_MEAN EVI_DL_DY_VS_LENGTH EVI_DL_DY_VS_MEAN EVI_DL_DY_WS_INTEGRAL EVI_DL_DY_WS_LENGTH EVI_DL_DY_WS_PEAK EVI_DL_DY_WS_SOWS	EVI_MWLR_DY_FRS_EOFRS EVI_MWLR_DY_FRS_LENGTH EVI_MWLR_DY_FRS_MEAN EVI_MWLR_DY_FRS_SOFRS EVI_MWLR_DY_RMS_INTEGRAL EVI_MWLR_DY_RMS_LENGTH EVI_MWLR_DY_VS_LENGTH EVI_MWLR_DY_VS_MEAN EVI_MWLR_DY_WS_EOSWS EVI_MWLR_DY_WS_INTEGRAL EVI_MWLR_DY_WS_LENGTH EVI_MWLR_DY_WS_PEAK EVI_MWLR_DY_WS_SOWS
Phases_VI_BiasC.	Variety_number SoilType_number Days_to_CCE_FROM_NDVI_DL_DY_FRS_EOFRS NDVI_DL_DY_FRS_EOFRS NDVI_DL_DY_FRS_INTEGRAL NDVI_DL_DY_FRS_LENGTH NDVI_DL_DY_FRS_SOFRS NDVI_DL_DY_RMS_LENGTH NDVI_DL_DY_RMS_MEAN NDVI_DL_DY_VS_LENGTH NDVI_DL_DY_VS_MEAN NDVI_DL_DY_WS_INTEGRAL NDVI_DL_DY_WS_LENGTH NDVI_DL_DY_WS_PEAK NDVI_DL_DY_WS_SOWS	Variety_number SoilType_number Days_to_CCE_FROM_NDVI_MWLR_DY_FRS_EOFRS NDVI_MWLR_DY_FRS_EOFRS NDVI_MWLR_DY_FRS_INTEGRAL NDVI_MWLR_DY_FRS_LENGTH NDVI_MWLR_DY_FRS_SOFRS NDVI_MWLR_DY_RMS_LENGTH NDVI_MWLR_DY_RMS_MEAN NDVI_MWLR_DY_VS_LENGTH NDVI_MWLR_DY_VS_MEAN NDVI_MWLR_DY_WS_EOSWS NDVI_MWLR_DY_WS_INTEGRAL NDVI_MWLR_DY_WS_LENGTH NDVI_MWLR_DY_WS_PEAK NDVI_MWLR_DY_WS_SOWS	Variety_number SoilType_number Days_to_CCE_FROM_EVI_DL_DY_FRS_EOFRS EVI_DL_DY_FRS_EOFRS EVI_DL_DY_FRS_INTEGRAL EVI_DL_DY_FRS_LENGTH EVI_DL_DY_FRS_SOFRS EVI_DL_DY_RMS_LENGTH EVI_DL_DY_RMS_MEAN EVI_DL_DY_VS_LENGTH EVI_DL_DY_VS_MEAN EVI_DL_DY_WS_INTEGRAL EVI_DL_DY_WS_LENGTH EVI_DL_DY_WS_PEAK EVI_DL_DY_WS_SOWS	Variety_number SoilType_number Days_to_CCE_FROM_EVI_MWLR_DY_FRS_EOFRS EVI_MWLR_DY_FRS_EOFRS EVI_MWLR_DY_FRS_LENGTH EVI_MWLR_DY_FRS_MEAN EVI_MWLR_DY_FRS_SOFRS EVI_MWLR_DY_RMS_INTEGRAL EVI_MWLR_DY_RMS_LENGTH EVI_MWLR_DY_VS_LENGTH EVI_MWLR_DY_VS_MEAN EVI_MWLR_DY_WS_EOSWS EVI_MWLR_DY_WS_INTEGRAL EVI_MWLR_DY_WS_LENGTH EVI_MWLR_DY_WS_PEAK EVI_MWLR_DY_WS_SOWS
No_VI	-	Latitude CCE, Longitude CCE Date of the CCE, DOY of CCE Amount spent (in Rs), SowingDate Variety_number, SoilType_number NDVI_MWLR_DY_FRS_TEMP_MEAN NDVI_MWLR_DY_RMS_TEMP_MEAN NDVI_MWLR_DY_VS_TEMP_MEAN NDVI_MWLR_DY_WS_TEMP_MEAN NDVI_MWLR_DY_FRS_PRECIP_SUM NDVI_MWLR_DY_RMS_PRECIP_SUM NDVI_MWLR_DY_VS_PRECIP_MEAN NDVI_MWLR_DY_WS_PRECIP_MEAN	-	-

Odisha	*NDVI_DL*	*NDVI_MWLR*	*EVI_DL*	*EVI_MWLR*
WS_VI	NDVI_DL_DY_WS_SOWS NDVI_DL_DY_WS_LENGTH NDVI_DL_DY_WS_PEAK NDVI_DL_DY_WS_INTEGRAL	NDVI_MWLR_DY_WS_INTEGRAL NDVI_MWLR_DY_WS_EOWS NDVI_MWLR_DY_WS_LENGTH NDVI_MWLR_DY_WS_PEAK NDVI_MWLR_DY_WS_SOWS	EVI_DL_DY_WS_SOSWS EVI_DL_DY_WS_LENGTH EVI_DL_DY_WS_PEAK EVI_DL_DY_WS_INTEGRAL	EVI_MWLR_DY_WS_INTEGRAL EVI_MWLR_DY_WS_EOSWS EVI_MWLR_DY_WS_LENGTH EVI_MWLR_DY_WS_PEAK EVI_MWLR_DY_WS_SOSWS
WS_VI_BiasC.	Variety_number_ny Soiltype_NY NDVI_DL_DY_WS_SOWS NDVI_DL_DY_WS_LENGTH NDVI_DL_DY_WS_PEAK NDVI_DL_DY_WS_INTEGRAL DAYS_TO_CCE_FROM_NDV_DL_SOWS	Variety_number_ny Soiltype_NY NDVI_MWLR_DY_WS_INTEGRAL NDVI_MWLR_DY_WS_EOWS NDVI_MWLR_DY_WS_LENGTH NDVI_MWLR_DY_WS_PEAK NDVI_MWLR_DY_WS_SOWS DAYS_TO_CCE_FROM_NDVI_MWLR_SOWS	Variety_number_ny Soiltype_NY EVI_DL_DY_WS_SOSWS EVI_DL_DY_WS_LENGTH EVI_DL_DY_WS_PEAK EVI_DL_DY_WS_INTEGRAL Days_to_CCE_from_EVI_DL_SOWS	Variety_number_ny Soiltype_NY EVI_MWLR_DY_WS_INTEGRAL EVI_MWLR_DY_WS_EOSWS EVI_MWLR_DY_WS_LENGTH EVI_MWLR_DY_WS_PEAK EVI_MWLR_DY_WS_SOSWS Days_to_CCE_from_EVI_MWLR_SOWS6.19109641552169E-07
Phases_VI	NDVI_DL_DY_FRS_EOFRS NDVI_DL_DY_FRS_SOFRS NDVI_DL_DY_FRS_LENGTH NDVI_DL_DY_FRS_INTEGRAL NDVI_DL_DY_RMS_LENGTH NDVI_DL_DY_RMS_MEAN NDVI_DL_DY_VS_LENGTH NDVI_DL_DY_VS_MEAN NDVI_DL_DY_WS_SOWS NDVI_DL_DY_WS_LENGTH NDVI_DL_DY_WS_PEAK NDVI_DL_DY_WS_INTEGRAL	NDVI_MWLR_DY_FRS_INTEGRAL NDVI_MWLR_DY_FRS_EOFRS NDVI_MWLR_DY_FRS_SOFRS NDVI_MWLR_DY_FRS_LENGTH NDVI_MWLR_DY_RMS_INTEGRAL NDVI_MWLR_DY_RMS_LENGTH NDVI_MWLR_DY_VS_LENGTH NDVI_MWLR_DY_VS_MEAN NDVI_MWLR_DY_WS_INTEGRAL NDVI_MWLR_DY_WS_EOWS NDVI_MWLR_DY_WS_LENGTH NDVI_MWLR_DY_WS_PEAK NDVI_MWLR_DY_WS_SOWS	EVI_DL_DY_FRS_INTEGRAL EVI_DL_DY_FRS_EOSFRS EVI_DL_DY_FRS_SOSFRS EVI_DL_DY_FRS_LENGTH EVI_DL_DY_RMS_LENGTH EVI_DL_DY_RMS_MEAN EVI_DL_DY_VS_INTEGRAL EVI_DL_DY_VS_LENGTH EVI_DL_DY_WS_SOSWS EVI_DL_DY_WS_LENGTH EVI_DL_DY_WS_PEAK EVI_DL_DY_WS_INTEGRAL	EVI_MWLR_DY_FRS_INTEGRAL EVI_MWLR_DY_FRS_EOSFRS EVI_MWLR_DY_FRS_SOSFRS EVI_MWLR_DY_FRS_LENGTH EVI_MWLR_DY_RMS_LENGTH EVI_MWLR_DY_RMS_MEAN EVI_MWLR_DY_VS_LENGTH EVI_MWLR_DY_VS_MEAN EVI_MWLR_DY_WS_INTEGRAL EVI_MWLR_DY_WS_EOSWS EVI_MWLR_DY_WS_LENGTH EVI_MWLR_DY_WS_PEAK EVI_MWLR_DY_WS_SOSWS
Phases_VI_BiasC	Variety_number_ny Soiltype_NY DAYS_TO_CCE_FROM_NDVI_DL_DY_FRS_EOFRS NDVI_DL_DY_FRS_EOFRS NDVI_DL_DY_FRS_SOFRS NDVI_DL_DY_FRS_LENGTH NDVI_DL_DY_FRS_INTEGRAL NDVI_DL_DY_RMS_LENGTH NDVI_DL_DY_RMS_MEAN NDVI_DL_DY_VS_LENGTH NDVI_DL_DY_VS_MEAN NDVI_DL_DY_WS_SOWS NDVI_DL_DY_WS_LENGTH NDVI_DL_DY_WS_PEAK NDVI_DL_DY_WS_INTEGRAL	Variety_number_ny Soiltype_NY DAYS_TO_CCE_FROM_NDVI_MWLR_DY_FRS_EOFRS NDVI_MWLR_DY_FRS_INTEGRAL NDVI_MWLR_DY_FRS_EOFRS NDVI_MWLR_DY_FRS_SOFRS NDVI_MWLR_DY_FRS_LENGTH NDVI_MWLR_DY_RMS_INTEGRAL NDVI_MWLR_DY_RMS_LENGTH NDVI_MWLR_DY_VS_LENGTH NDVI_MWLR_DY_VS_MEAN NDVI_MWLR_DY_WS_INTEGRAL NDVI_MWLR_DY_WS_EOWS NDVI_MWLR_DY_WS_LENGTH NDVI_MWLR_DY_WS_PEAK NDVI_MWLR_DY_WS_SOWS	Variety_number_ny Soiltype_NY DAYS_TO_CCE_FROM_EVI_DL_DY_FRS_EOSFRS EVI_DL_DY_FRS_INTEGRAL EVI_DL_DY_FRS_EOSFRS EVI_DL_DY_FRS_SOSFRS EVI_DL_DY_FRS_LENGTH EVI_DL_DY_RMS_LENGTH EVI_DL_DY_RMS_MEAN EVI_DL_DY_VS_INTEGRAL EVI_DL_DY_VS_LENGTH EVI_DL_DY_WS_SOSWS EVI_DL_DY_WS_LENGTH EVI_DL_DY_WS_PEAK EVI_DL_DY_WS_INTEGRAL	Variety_number_ny Soiltype_NY DAYS_TO_CCE_FROM_EVI_MWLR_DY_FRS_EOSFRS EVI_MWLR_DY_FRS_INTEGRAL EVI_MWLR_DY_FRS_EOSFRS EVI_MWLR_DY_FRS_SOSFRS EVI_MWLR_DY_FRS_LENGTH EVI_MWLR_DY_RMS_LENGTH EVI_MWLR_DY_RMS_MEAN EVI_MWLR_DY_VS_LENGTH EVI_MWLR_DY_VS_MEAN EVI_MWLR_DY_WS_INTEGRAL EVI_MWLR_DY_WS_EOSWS EVI_MWLR_DY_WS_LENGTH EVI_MWLR_DY_WS_PEAK EVI_MWLR_DY_WS_SOSWS
No_VI	Latitude CCE Longitude CCE Date of the CCE DOY of CCE SowingDate Variety_number_ny Soiltype_NY			

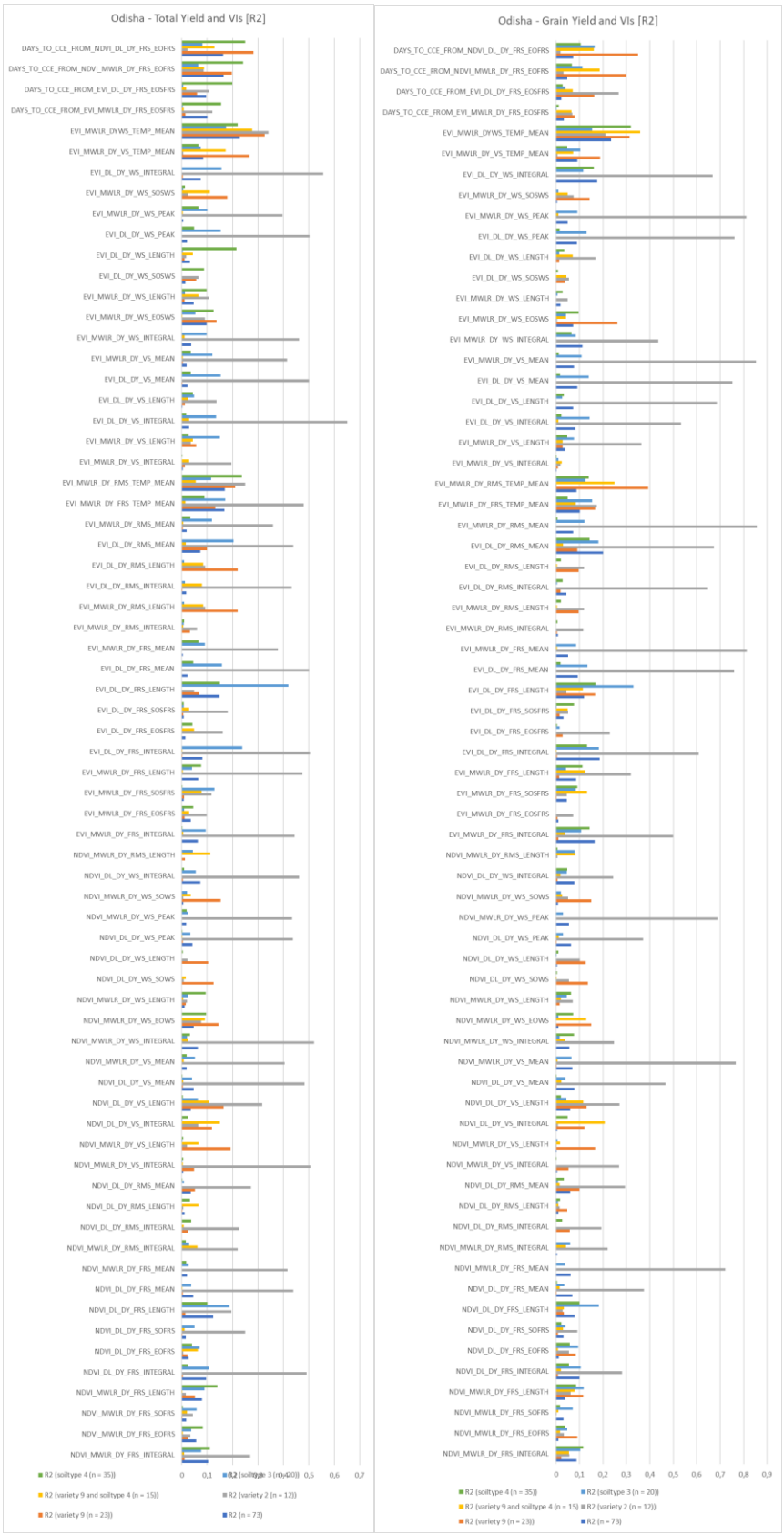
11.6 Results of the individual correlations

11.6.1 Haryana





11.6.2 Odisha


Odisha - Total Yield and VIs [R2]
DAYS_TO_CCE_FROM_NDVI_DL_DY_FRS_EOFRS
DAYS_TO_CCE_FROM_NDVI_MWLR_DY_FRS_EOFRS
DAYS_TO_CCE_FROM_EVI_DL_DY_FRS_EOSFRS
DAYS_TO_CCE_FROM_EVI_MWLR_DY_FRS_EOSFRS
EVI_MWLR_DYWS_TEMP_MEAN
EVI_MWLR_DY_VS_TEMP_MEAN
EVI_DL_DY_WS_INTEGRAL
EVI_MWLR_DY_WS_SOSWS
EVI_MWLR_DY_WS_PEAK
EVI_DL_DY_WS_PEAK
EVI_DL_DY_WS_LENGTH
EVI_DL_DY_WS_SOSWS
EVI_MWLR_DY_WS_LENGTH
EVI_MWLR_DY_WS_EOSWS
EVI_MWLR_DY_WS_INTEGRAL
EVI_MWLR_DY_VS_MEAN
EVI_DL_DY_VS_MEAN
EVI_DL_DY_VS_LENGTH
EVI_DL_DY_VS_INTEGRAL
EVI_MWLR_DY_VS_LENGTH
EVI_MWLR_DY_VS_INTEGRAL
EVI_MWLR_DY_RMS_TEMP_MEAN
EVI_MWLR_DY_FRS_TEMP_MEAN
EVI_MWLR_DY_RMS_MEAN
EVI_DL_DY_RMS_MEAN
EVI_DL_DY_RMS_LENGTH
EVI_DL_DY_RMS_INTEGRAL
EVI_MWLR_DY_RMS_LENGTH
EVI_MWLR_DY_RMS_INTEGRAL
EVI_MWLR_DY_FRS_MEAN
EVI_DL_DY_FRS_MEAN
EVI_DL_DY_FRS_LENGTH
EVI_DL_DY_FRS_SOSFRS
EVI_DL_DY_FRS_EOSFRS
EVI_DL_DY_FRS_INTEGRAL
EVI_MWLR_DY_FRS_LENGTH
EVI_MWLR_DY_FRS_SOSFRS
EVI_MWLR_DY_FRS_EOSFRS
EVI_MWLR_DY_FRS_INTEGRAL
NDVI_MWLR_DY_RMS_LENGTH
NDVI_DL_DY_WS_INTEGRAL
NDVI_MWLR_DY_WS_SOWS
NDVI_MWLR_DY_WS_PEAK
NDVI_DL_DY_WS_PEAK
NDVI_DL_DY_WS_LENGTH
NDVI_DL_DY_WS_SOWS
NDVI_MWLR_DY_WS_LENGTH
NDVI_MWLR_DY_WS_EOWS
NDVI_MWLR_DY_WS_INTEGRAL
NDVI_MWLR_DY_VS_MEAN
NDVI_DL_DY_VS_MEAN
NDVI_DL_DY_VS_LENGTH
NDVI_DL_DY_VS_INTEGRAL
NDVI_MWLR_DY_VS_LENGTH
NDVI_MWLR_DY_VS_INTEGRAL
NDVI_DL_DY_RMS_MEAN
NDVI_DL_DY_RMS_LENGTH
NDVI_DL_DY_RMS_INTEGRAL
NDVI_MWLR_DY_RMS_INTEGRAL
NDVI_MWLR_DY_FRS_MEAN
NDVI_DL_DY_FRS_MEAN
NDVI_DL_DY_FRS_LENGTH
NDVI_DL_DY_FRS_SOFRS
NDVI_DL_DY_FRS_EOFRS
NDVI_DL_DY_FRS_INTEGRAL
NDVI_MWLR_DY_FRS_LENGTH
NDVI_MWLR_DY_FRS_SOFRS
NDVI_MWLR_DY_FRS_EOFRS
NDVI_MWLR_DY_FRS_INTEGRAL
0 0,1 0,2 0,3 0,4 0,5 0,6 0,7
R2 (soiltype 4 (n = 35))
R2 (soiltype 3 (n=20))
R2 (variety 9 and soiltype 4 (n = 15))
R2 (variety 2 (n = 12))
R2 (variety 9 (n = 23))
R2 (n = 73)
Odisha - Grain Yield and VIs [R2]
DAYS_TO_CCE_FROM_NDVI_DL_DY_FRS_EOFRS
DAYS_TO_CCE_FROM_NDVI_MWLR_DY_FRS_EOFRS
DAYS_TO_CCE_FROM_EVI_DL_DY_FRS_EOSFRS
DAYS_TO_CCE_FROM_EVI_MWLR_DY_FRS_EOSFRS
EVI_MWLR_DYWS_TEMP_MEAN
EVI_MWLR_DY_VS_TEMP_MEAN
EVI_DL_DY_WS_INTEGRAL
EVI_MWLR_DY_WS_SOSWS
EVI_MWLR_DY_WS_PEAK
EVI_DL_DY_WS_PEAK
EVI_DL_DY_WS_LENGTH
EVI_DL_DY_WS_SOSWS
EVI_MWLR_DY_WS_LENGTH
EVI_MWLR_DY_WS_EOSWS
EVI_MWLR_DY_WS_INTEGRAL
EVI_MWLR_DY_VS_MEAN
EVI_DL_DY_VS_MEAN
EVI_DL_DY_VS_LENGTH
EVI_DL_DY_VS_INTEGRAL
EVI_MWLR_DY_VS_LENGTH
EVI_MWLR_DY_VS_INTEGRAL
EVI_MWLR_DY_RMS_TEMP_MEAN
EVI_MWLR_DY_FRS_TEMP_MEAN
EVI_MWLR_DY_RMS_MEAN
EVI_DL_DY_RMS_MEAN
EVI_DL_DY_RMS_LENGTH
EVI_DL_DY_RMS_INTEGRAL
EVI_MWLR_DY_RMS_LENGTH
EVI_MWLR_DY_RMS_INTEGRAL
EVI_MWLR_DY_FRS_MEAN
EVI_DL_DY_FRS_MEAN
EVI_DL_DY_FRS_LENGTH
EVI_DL_DY_FRS_SOSFRS
EVI_DL_DY_FRS_EOSFRS
EVI_DL_DY_FRS_INTEGRAL
EVI_MWLR_DY_FRS_LENGTH
EVI_MWLR_DY_FRS_SOSFRS
EVI_MWLR_DY_FRS_EOSFRS
EVI_MWLR_DY_FRS_INTEGRAL
NDVI_MWLR_DY_RMS_LENGTH
NDVI_DL_DY_WS_INTEGRAL
NDVI_MWLR_DY_WS_SOWS
NDVI_MWLR_DY_WS_PEAK
NDVI_DL_DY_WS_PEAK
NDVI_DL_DY_WS_LENGTH
NDVI_DL_DY_WS_SOWS
NDVI_MWLR_DY_WS_LENGTH
NDVI_MWLR_DY_WS_EOWS
NDVI_MWLR_DY_WS_INTEGRAL
NDVI_MWLR_DY_VS_MEAN
NDVI_DL_DY_VS_MEAN
NDVI_DL_DY_VS_LENGTH
NDVI_DL_DY_VS_INTEGRAL
NDVI_MWLR_DY_VS_LENGTH
NDVI_MWLR_DY_VS_INTEGRAL
NDVI_DL_DY_RMS_MEAN
NDVI_DL_DY_RMS_LENGTH
NDVI_DL_DY_RMS_INTEGRAL
NDVI_MWLR_DY_RMS_INTEGRAL
NDVI_MWLR_DY_FRS_MEAN
NDVI_DL_DY_FRS_MEAN
NDVI_DL_DY_FRS_LENGTH
NDVI_DL_DY_FRS_SOFRS
NDVI_DL_DY_FRS_EOFRS
NDVI_DL_DY_FRS_INTEGRAL
NDVI_MWLR_DY_FRS_LENGTH
NDVI_MWLR_DY_FRS_SOFRS
NDVI_MWLR_DY_FRS_EOFRS
NDVI_MWLR_DY_FRS_INTEGRAL
0 0,1 0,2 0,3 0,4 0,5 0,6 0,7 0,8 0,9
R2 (soiltype 4 (n = 35))
R2 (soiltype 3 (n = 20))
R2 (variety 9 and soiltype 4 (n = 15))
R2 (variety 2 (n = 12))
R2 (variety 9 (n = 23))
R2 (n = 73)

11.7 Full result of the multiple regression analyses

11.7.1 Haryana

Table 9: *Results of the multiple regressions for Haryana. If more than one end variables are used, the R^2 is the adjusted R^2.*

Haryana	*NDVI_DL*	*NDVI_MWLR*	*EVI_DL*	*EVI_MWLR*
WS_VI				
Start variables	4	5	4	5
R^2	0.04	0.08	0.02	0.01
End variables	1	4	1	1
Significant	0.01	0.001	0.08	0.18
WS_VI_BiasC.				
Start variables	8	8	7	8
R^2	0.11	0.18	0.15	0.07
End variables	2	3	2	1
Significant	2.00E-03	1.00E-05	1.80E-04	0.01
Phases_VI				
Start variables	12	13	12	13
R^2	0.05	0.08	0.02	0.04
End variables	1	2	1	4
Significant	2.06E-03	2.40E-04	0.06	0.02
Phases_VI_BiasC				
Start variables	15	16	15	16
R^2	0.13	0.17	0.16	0.24
End variables	2	6	2	6
Significant	7.80E-04	8.60E-04	1.10E-04	1.76E-05
No_VI				
Start variables		16		
R^2		0.12		
End variables		2		
Significant		1.33E-03		
All_VI_Fixed				
Start variables		5		
R^2		0.05		
End variables		1		
Significant		3.20E-05		
All_VI_Fixed_BiasC				
Start variables		8		
R^2		0.15		
End variables		3		
Significant		1.36E-04		

11.7.2 Odisha

Table 10: *Results of the multiple regressions for Odisha. If more than one end variables are used, the R^2 is the adjusted R^2.*

Odisha	*NDVI_DL*	*NDVI_MWLR*	*EVI_DL*	*EVI_MWLR*
WS_VI				
Start variables	4	5	4	5
R^2	0.08	0.06	0.24	0.14
End variables	1	1	3	2
Significant	0.02	0.04	5.80E-05	2.20E-03
WS_VI_BiasC.				
Start variables	7	8	7	8
R^2	0.29	0.27	0.42	0.37
End variables	4	4	6	5
Significant	1.46E-05	3.59E-05	1.54E-07	6.19E-07
Phases_VI				
Start variables	12	13	12	13
R^2	0.14	0.17	0.24	0.31
End variables	1	7	3	5
Significant	0.14	0.01	8.23E-05	1.22E-05
Phases_VI_BiasC				
Start variables	15	16	15	16
R^2	0.39	0.30	0.53	0.49
End variables	4	4	9	8
Significant	2.34E-07	1.19E-05	4.05E-09	1.23E-08
No_VI	-		-	-
Start variables		7		
R^2		0.17		
End variables		3		
Significant		9.98E-04		

11.8 Full results of the RF classification

11.8.1 Haryana

Table 11: *Haryana: Full results of the RF Classifications for the different groups of variables.*

Haryana	*NDVI_DL*	*NDVI_MWLR*	*EVI_DL*	*EVI_MWLR*
WS_VI				
OFS	1	4	2	2
CVS	0.25	0.31	0.26	0.33
WS_VI_BiasC.				
OFS	8	4	2	2
CVS	0.37	0.31	0.26	0.33
Phases_VI				
OFS	8		3	
CVS	0.32		0.43	
Phases_VI_BiasC				
OFS	9	10	8	16
CVS	0.32	0.35	0.33	0.37
No_VI				
OFS		8		
CVS		0.46		
All_VI_Fixed				
OFS		5		
CVS		0.25		
All_VI_Fixed_BiasC				
OFS		6		
CVS		0.32		

11.8.2 Odisha

Table 12: *Odisha: Full results of the RF Classifications for the different groups of variables.*

Odisha	*NDVI_DL*	*NDVI_MWLR*	*EVI_DL*	*EVI_MWLR*
WS_VI				
OFS	3	2	3	4
CVS	0.42	0.4	0.38	0.36
WS_VI_BiasC.				
OFS	3	4	3	5
CVS	0.44	0.41	0.37	0.34
Phases_VI				
OFS	5	12	4	
CVS	0.39	0.46	0.47	
Phases_VI_BiasC				
OFS	14	14	5	7
CVS	0.45	0.47	0.46	0.44
No_VI				
OFS		2		
CVS		0.40		

11.9 Linear regression – No VI variables

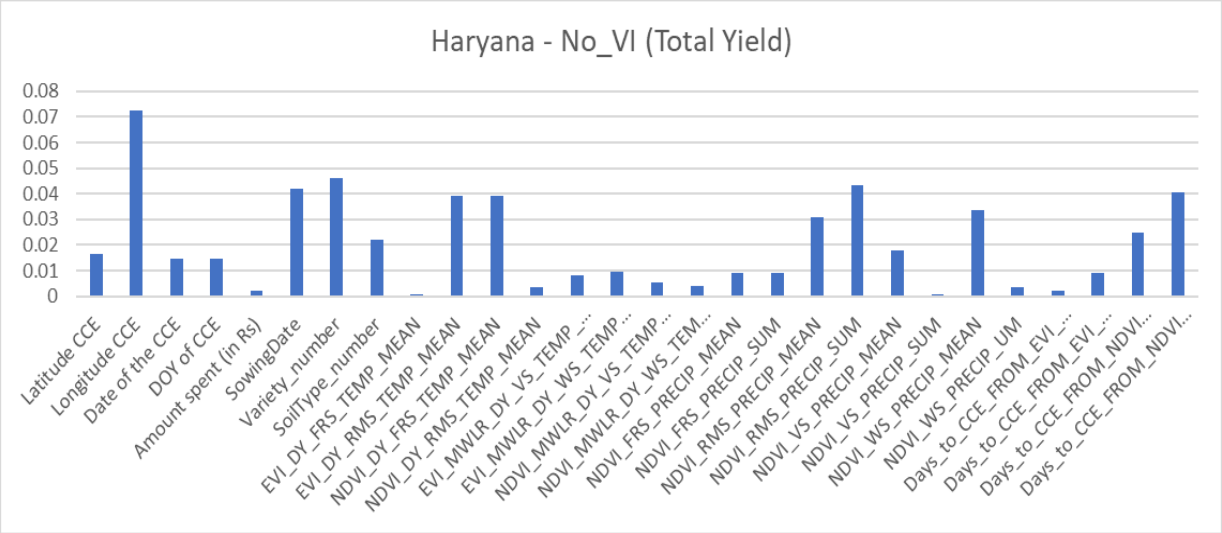


Figure 52: Results of the linear regression between the Non-VI-variables and total yield for Haryana.

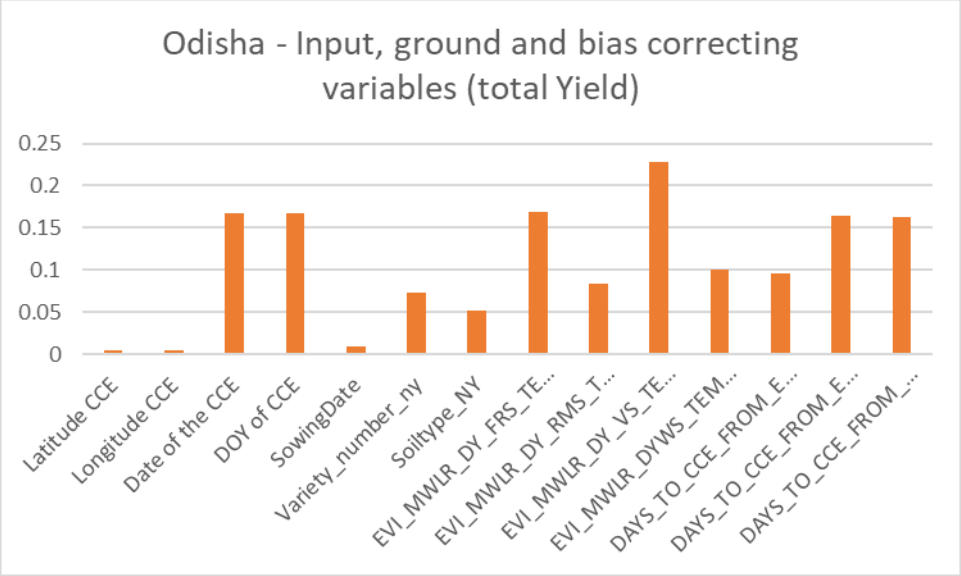


Figure 53: Results of the linear regression between the Non-VI-variables and total yield for Odisha.

11.10 CCE method

Crop cut method. IFPRI, 2019 - Protocol for Crop Cutting Experiments: Gp-Level Yield Estimation Through Smartphone Based Near-Surface Sensing Approach

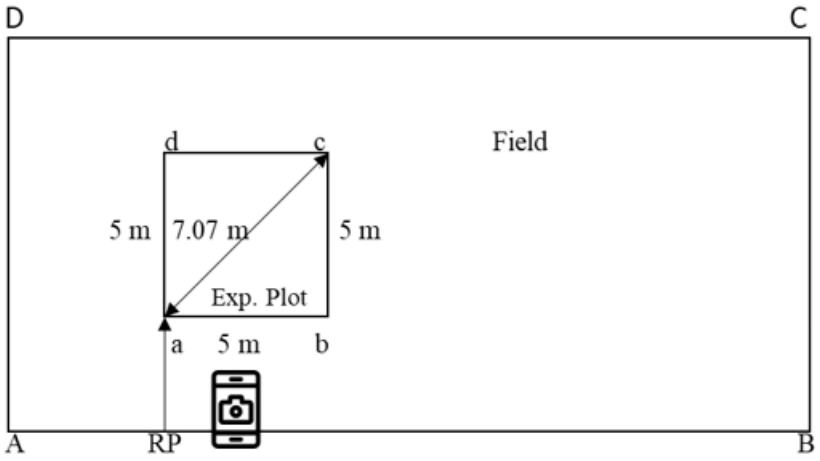


Figure 1: Randomly demarcating the CCE plot

Figure 54: Instructions for the CCEs.

11.11 GEE script: Data preparation and the temporal aggregation

The script is generally similar for all the different variables with MWLR. However, the step when finding SOS and EOS will be different if the variables is to be found for a phase. Similarly, the triggering measure will differ. In the end will be shown examples of how the DL smoothing, precipitation, temperature and fixed seasons differ.

```
Imports (4 entries)
  var Sent2a: ImageCollection "Sentinel-2 MSI: MultiSpectral Instrument, Level-2A"
  var AOI: Polygon, 4 vertices
  var ROI: Point (76.64, 30.17)
  var table: Table users/n-b-p/Berber/CCE_luget_export
////////////////////////////////////////////////
///////From timeseries to image//////////////////
////////////////////////////////////////////////

///// Filtering the imagecollection
var s2a = Sent2a
  .filterBounds(ROI)
  .filterDate('2019-05-15', '2019-12-01');
Map.addLayer(table);
Map.addLayer(s2a);

var timeField = 'system:time_start';


///// Adds NDVI band

// function to add NDVI and time bands to image collection
var addDataBands = function(image) {
  var ndvi = image.normalizedDifference(['B8', 'B4']).rename('NDVI');
  return image.addBands(ndvi)
              .addBands(image.metadata('system:time_start').divide(1e18).rename('time'));
};

var data = s2a
  .map(addDataBands);



///// clips the timeseries to the region of interest
function klipper (image){
  return image.clip(AOI);
}
var data = data.map(klipper);



///// First smoothing of the timeseries
function smoother(t){

  function applyFit(img){
      return img.select('time').multiply(fit.select('scale')).add(fit.select('offset'))
                .set('system:time_start',img.get('system:time_start')).rename('NDVI');
  }
  t = ee.Date(t);

  var window = data.filterDate(t.advance(-windowSize,'day'),t.advance(windowSize,'day'));

  var fit = window.select(['time','NDVI'])
    .reduce(ee.Reducer.linearFit());

  return window.map(applyFit).toList(5);
}


function reduceFits(t){
  t = ee.Date(t);
  return fitIC.filterDate(t.advance(-windowSize,'day'),t.advance(windowSize,'day'))
              .mean().set('system:time_start',t.millis()).rename('NDVIsmooth1');
}

```

```
61
62  var dates = ee.List(data.aggregate_array('system:time_start'));
63
64  var windowSize = 6; //days on either side
65
66  var fitIC = ee.ImageCollection(dates.map(smoother).flatten());
67
68  var smoothed = ee.ImageCollection(dates.map(reduceFits));
69
70
71
72  ///// Joins the smoothed timeseries to the original
73
74  var filter = ee.Filter.equals({
75    leftField: 'system:time_start',
76    rightField: 'system:time_start'
77  });
78
79  // Create the join.
80  var simpleJoin = ee.Join.inner();
81
82  // Inner join
83  var innerJoin = ee.ImageCollection(simpleJoin.apply(data, smoothed, filter));
84
85  var data = innerJoin.map(function(feature) {
86    return ee.Image.cat(feature.get('primary'), feature.get('secondary'));
87  });
88
89
90  //// Prepares for second smoothing
91
92  /// creates variable (sneak), that contains the highest value of the original or smoothed timeseries
93
94  var førstel = function(image) {
95    var diff = image.select('NDVIsmooth1').subtract(image.select('NDVI')).rename('diff');
96    return image.addBands(diff);
97    };
98
99
100 var andenl =  function(image) {
101   var absdiff = image.select('diff').abs().rename('absdiff');
102   return image.addBands(absdiff);
103   };
104
105
106 var tredjel =  function(image) {
107   var sneak = image.select('NDVI').add(image.select('absdiff').add(image.select('diff')).divide(2)).rename('sneak');
108   return image.addBands(sneak);
109   };
110
111 var data = data
112   .map(førstel)
113   .map(andenl)
114   .map(tredjel);
```

```
115
116 ///// Then do the second smoothing on the variable "sneak".
117
118
119 // Function to smooth time series
120 // stacks windows of linear regression results
121 // requires that a variable 'data' exists with NDVI and time bands
122 function smoother2(t){
123   // helper function to apply linear regression equation
124   function applyFit(img){
125       return img.select('time').multiply(fit.select('scale')).add(fit.select('offset'))
126                 .set('system:time_start',img.get('system:time_start')).rename('NDVI2');
127   }
128   t = ee.Date(t);
129
130   var window = data.filterDate(t.advance(-windowSize,'day'),t.advance(windowSize,'day'));
131
132   var fit = window.select(['time','sneak'])
133     .reduce(ee.Reducer.linearFit());
134
135   return window.map(applyFit).toList(5);
136 }
137
138 // function to reduce time stacked linear regression results
139 // requires that a variable 'fitIC' exists from the smooter function
140 function reduceFits2(t){
141   t = ee.Date(t);
142   return fitIC1.filterDate(t.advance(-windowSize,'day'),t.advance(windowSize,'day'))
143                .mean().set('system:time_start',t.millis()).rename('NDVIsmooth2');
144 }
145
146
147
148 var dates2 = ee.List(data.aggregate_array('system:time_start'));
149
150
151 var fitIC1 = ee.ImageCollection(dates2.map(smoother2).flatten());
152
153 var smoothed2 = ee.ImageCollection(dates2.map(reduceFits2));
154
155 ///// joins the new smoothing to the original
156
157 // Inner join
158 var innerJoin2 = ee.ImageCollection(simpleJoin.apply(data, smoothed2, filter));
159
160 var data = innerJoin2.map(function(feature) {
161   return ee.Image.cat(feature.get('primary'), feature.get('secondary'));
162 });
163
164 ///// Repeated until 10 smoothings
165
166
167
```

```
885
886 ///// Inserting a value every day/////
887
888 /// add day of year
889 var addDate_ny = function(image){
890   var doy = image.date().getRelative('day', 'year');
891   var doyBand = ee.Image.constant(doy).float().rename('doy');
892
893   return image.set('doy', doyBand);
894 };
895
896 /////vælger et bånd og snævrer tiden. og tilføjer doy som property, så den kan bruges til at joine med
897
898 var smoothedndvi = data.select('NDVIsmooth10').map(addDate_ny);
899
900
901 ///shifts ndvismooth10 with 5 days back and forward,and then merge
902 //Back
903 function gs_dummy_back_f (image) {
904   var time_start = ee.Number(image.get('system:time_start'));
905   var time_start2 = time_start.subtract(5*86400000);
906   var NDVIsmooth10 = image.select('NDVIsmooth10').rename('NDVIsmooth10_back');
907   return image.set('system:time_start', time_start2).addBands(NDVIsmooth10);
908 }
909
910 var gs_dummy_back = smoothedndvi.select('NDVIsmooth10').map(gs_dummy_back_f).map(addDate_ny);
911
912
913 //forward
914 function gs_dummy_forward_f (image) {
915   var time_start = ee.Number(image.get('system:time_start'));
916   var time_start2 = time_start.add(5*86400000);
917   var NDVIsmooth10 = image.select('NDVIsmooth10').rename('NDVIsmooth10_forward');
918   return image.set('system:time_start', time_start2).addBands(NDVIsmooth10);
919 }
920
921 var gs_dummy_forward = smoothedndvi.select('NDVIsmooth10').map(gs_dummy_forward_f).map(addDate_ny);
922 print('gs_dummy_forward', gs_dummy_forward);
923
924
925 ///Joining to one timeseries
926
927 var filter_ny = ee.Filter.equals({
928   leftField: 'doy',
929   rightField: 'doy'
930 });
931
932 // Inner join
933 var innerJoin_ny1 = ee.ImageCollection(simpleJoin.apply(smoothedndvi, gs_dummy_back, filter_ny));
934
935 var smoothedndvi_merged_1 = innerJoin_ny1.map(function(feature) {
936   return ee.Image.cat(feature.get('primary'), feature.get('secondary'));
937 });
938 var innerJoin_ny2 = ee.ImageCollection(simpleJoin.apply(smoothedndvi_merged_1, gs_dummy_forward, filter_ny));
939
940 var smoothedndvi_merged = innerJoin_ny2.map(function(feature) {
941   return ee.Image.cat(feature.get('primary'), feature.get('secondary'));
942 });
943
944
```

```
945
946  /////Then need to insert the daily value. two on rither side.
947  //first two to the left
948
949  function timeturner1 (image) {
950    var time_start = ee.Number(image.get('system:time_start'));
951    var time_start2 = time_start.subtract(86400000);
952    var NDVIsmooth10 = image.select('NDVIsmooth10').subtract(((image.select('NDVIsmooth10')
953    .subtract(image.select('NDVIsmooth10_forward'))).divide(5)).multiply(1)).rename('NDVIsmooth10');
954    return image.set('system:time_start', time_start2).addBands(NDVIsmooth10);
955  }
956
957  var dummy1 = smoothedndvi_merged.map(timeturner1);
958
959  function namechange (image) {
960    var NDVIsmooth10 = image.select('NDVIsmooth10_3').rename('NDVIsmooth10');
961    return image.addBands(NDVIsmooth10);
962  }
963  var dummy1 = dummy1.select('NDVIsmooth10_3');
964  var dummy1 = dummy1.map(namechange);
965  var dummy1 = dummy1.select('NDVIsmooth10');
966  print('dummy1', dummy1);
967
968  ///another
969  function timeturner2 (image) {
970    var time_start = ee.Number(image.get('system:time_start'));
971    var time_start2 = time_start.subtract(2*86400000);
972    var NDVIsmooth10 = image.select('NDVIsmooth10').subtract(((image.select('NDVIsmooth10')
973    .subtract(image.select('NDVIsmooth10_forward'))).divide(5)).multiply(2)).rename('NDVIsmooth10');
974    return image.set('system:time_start', time_start2).addBands(NDVIsmooth10);
975  }
976
977  var dummy2 = smoothedndvi_merged.map(timeturner2);
978
979
980  var dummy2 = dummy2.select('NDVIsmooth10_3');
981  var dummy2 = dummy2.map(namechange);
982  var dummy2 = dummy2.select('NDVIsmooth10');
983
984
985  /////then two to the right
986  function timeturner3 (image) {
987    var time_start = ee.Number(image.get('system:time_start'));
988    var time_start2 = time_start.add(86400000);
989    var NDVIsmooth10 = image.select('NDVIsmooth10').subtract(((image.select('NDVIsmooth10')
990    .subtract(image.select('NDVIsmooth10_back'))).divide(5)).multiply(1)).rename('NDVIsmooth10');
991    return image.set('system:time_start', time_start2).addBands(NDVIsmooth10);
992  }
993
994  var dummy3 = smoothedndvi_merged.map(timeturner3);
995
996  function namechange (image) {
997    var NDVIsmooth10 = image.select('NDVIsmooth10_3').rename('NDVIsmooth10');
998    return image.addBands(NDVIsmooth10);
999  }
1000 var dummy3 = dummy3.select('NDVIsmooth10_3');
1001 var dummy3 = dummy3.map(namechange);
1002 var dummy3 = dummy3.select('NDVIsmooth10');
1003
```

```
///another
function timeturner4 (image) {
  var time_start = ee.Number(image.get('system:time_start'));
  var time_start2 = time_start.add(2*86400000);
  var NDVIsmooth10 = image.select('NDVIsmooth10').subtract(((image.select('NDVIsmooth10')
  .subtract(image.select('NDVIsmooth10_back'))).divide(5)).multiply(2)).rename('NDVIsmooth10');
  return image.set('system:time_start', time_start2).addBands(NDVIsmooth10);
}

var dummy4 = smoothedndvi_merged.map(timeturner4);


var dummy4 = dummy4.select('NDVIsmooth10_3');
var dummy4 = dummy4.map(namechange);
var dummy4 = dummy4.select('NDVIsmooth10');


var merged = dummy1.merge(smoothedndvi).merge(dummy2).merge(dummy3).merge(dummy4);


var smoothedndvi_time = merged.filterDate('2019-05-01', '2019-12-15');

////////shortening the season from min til min////////////

var addDate = function(image){
  var doy = image.date().getRelative('day', 'year');
  var doyBand = ee.Image.constant(doy).float().rename('doy');

  return image.addBands(doyBand);
};

var data = smoothedndvi_time.select('NDVIsmooth10').map(addDate);

var maxndvi = data.select('NDVIsmooth10').max().rename('maxndvi');
//finding date for peak
///function that removes all values less than max
function for_max_date (image) {
  var smoorth = image.select('NDVIsmooth10');
  return image
  .updateMask(smoorth.gte(maxndvi.select('maxndvi')));
}
/// get maxdoy
var maxndvi_date = data.map(for_max_date).select('doy').max();

function for_forstedel (image) {
  var smoorth = image.select('doy');
  return image
  .updateMask(smoorth.lte(maxndvi_date.select('doy')));
}

var forste_del = data.map(for_forstedel);
```

```
//finding date for min in first half

var forste_del_min = forste_del.select('NDVIsmooth10').min();

function for_forste_del_min_dato (image) {
  var smoorth = image.select('NDVIsmooth10');
  return image
  .updateMask(smoorth.lte(forste_del_min));
}
//dette skal så være start
var forste_del_min_dato = forste_del.map(for_forste_del_min_dato).select('doy').min();
print('forste_del_min_dato', forste_del_min_dato);


//finding for second half

function for_andendel (image) {
  var smoorth = image.select('doy');
  return image
  .updateMask(smoorth.gte(maxndvi_date.select('doy')));
}

var anden_del = data.map(for_andendel);

//finding min for second half of the season

var anden_del_min = anden_del.select('NDVIsmooth10').min();
print('anden_del_min',anden_del_min);

function for_anden_del_min_dato (image) {
  var smoorth = image.select('NDVIsmooth10');
  return image
  .updateMask(smoorth.lte(anden_del_min));
}

var anden_del_min_dato = anden_del.map(for_anden_del_min_dato).select('doy').min();

///Then make the window

function for_vindue (image) {
  var smoorth = image.select('doy');
  return image
  .updateMask(smoorth.gte(forste_del_min_dato))
  .updateMask(smoorth.lte(anden_del_min_dato));
}
var smoothedndvi_time = data.map(for_vindue);

```

```
1106  ///////determining SoS and EOS based on when the VI value reaches a certain percentage of the increase or decrease
1107
1108  var maxndvi = smoothedndvi_time.select('NDVIsmooth10').max().rename('maxndvi');
1109
1110
1111  var minndvi = smoothedndvi_time.select('NDVIsmooth10').min().rename('minndvi');
1112
1113
1114  var addDate = function(image){
1115    var doy = image.date().getRelative('day', 'year');
1116    var doyBand = ee.Image.constant(doy).float().rename('doy');
1117
1118    return image.addBands(doyBand);
1119  };
1120
1121  var smoothedndvi_time = smoothedndvi_time.map(addDate);
1122
1123
1124  //// SOS is when reaching 20% of the increase
1125  var SOS_threshold = ((maxndvi.subtract(minndvi)).multiply(0.2)).add(minndvi); ///for the phases the specific values will be used
1126  var EOS_threshold = ((maxndvi.subtract(minndvi)).multiply(0.2)).add(minndvi) ;
1127
1128
1129  ///function that removes all values less than SOS
1130  function for_SOS_date (image) {
1131    var smoorth = image.select('NDVIsmooth10');
1132    return image
1133    .updateMask(smoorth.gte(SOS_threshold.select('maxndvi')));
1134  }
1135  /// get SOSdoy
1136  var SOSdate = smoothedndvi_time.map(for_SOS_date).select('doy').min();
1137
1138  ///function that removes all values less than EOS
1139  function for_EOS_date (image) {
1140    var smoorth = image.select('NDVIsmooth10');
1141    return image
1142    .updateMask(smoorth.gte(EOS_threshold.select('maxndvi')));
1143  }
1144  /// get SOSdoy
1145  var EOSdate = smoothedndvi_time.map(for_EOS_date).select('doy').max();
1146
1147  ////function that removes values before SOS and after EOS
1148  function filter_SOS_EOS (image) {
1149    var doyselect = image.select('doy');
1150    return image
1151      .updateMask(doyselect.gte(SOSdate))  ///remove obs. before SOS
1152      .updateMask(doyselect.lte(EOSdate));
1153  }
1154
1155  var growing_season = smoothedndvi_time.map(filter_SOS_EOS).select('NDVIsmooth10');
1156
1157  //////
1158  //Integral triggering measure
1159  var accndvi = growing_season.sum().subtract((((maxndvi.subtract(minndvi))
1160  .multiply(0.2)).add(minndvi)).multiply(growing_season.count()));  ////for other triggering measures this will differ
1161
1162  Export.image.toAsset({
1163    image: accndvi,
1164    description: 'Whole_Season',
1165    assetId: '08_06/Edday_ndvi10s_nord1_w6_midtmaj-dec_usky_s_fraction_WS_02_inds',
1166    scale: 10,
1167    region: AOI,
1168    pyramidingPolicy: {'.default': 'sample'}
1169  });
```

11.11.1 For EVI

```
Imports (4 entries)
  var Sent2a: ImageCollection "Sentinel-2 MSI: MultiSpectral Instrument, Level-2A"
  var AOI: Polygon, 4 vertices
  var ROI: Point (76.64, 30.17)
  var table: Table users/n-b-p/Berber/CCE_luget_export
```

```
////////////////////////////
//////  EVI  ////////////////
////////////////////////////

//////filtering the timeseries
var s2a = Sent2a
  .filterBounds(ROI)
  .filterDate('2019-05-15', '2019-12-01');
//  .sort('CLOUD_COVER');


var timeField = 'system:time_start';

var simpleJoin = ee.Join.inner();

/// Adds EVI band to the timeseries (calls it NDVI so it matches with the rest of the script)

var addDataBands = function(image) {
   var evi = image.expression(
    '2.5 * ((NIR - RED) / (NIR + 6 * RED - 7.5 * BLUE + 1))', {
      'NIR': image.select('B8').divide(10000),
      'RED': image.select('B4').divide(10000),
      'BLUE': image.select('B2').divide(10000)
}).rename('evi');
    return image.addBands(evi)
                .addBands(image.metadata('system:time_start').divide(1e18).rename('time'));
};

var addDataBands2 = function(image) {
  var evil = image.expression( '1 - evil', {
    'evil': image.select('evi')}).rename('evil');
   return image.addBands(evil);
};


var addDataBands3 = function(image) {
  var ndvi = image.expression(
      '((abs + (1 - evi)) / (1 - evi)) * evi / 2', {
        'abs': image.select('evil').abs(),
        'evi': image.select('evi')
       }).rename('NDVI_a');
   return image.addBands(ndvi);
};

////Make all values less than 0 to 0.
function addDataBands4 (image) {
  var ndviabs = image.select('NDVI_a').abs();
  var ndvi = ndviabs.add(image.select('NDVI_a')).divide(2).rename('NDVI');
  return image.addBands(ndvi);
}
```

```
55
56  var a_addDate_ny = function(image){
57    var doy = image.date().getRelative('day', 'year');
58    var doyBand = ee.Image.constant(doy).float().rename('doy');
59
60    return image.set('doy', doyBand);
61  };
62
63  var data = s2a
64    .map(addDataBands)
65    .map(addDataBands2)
66    .map(addDataBands3)
67    .map(addDataBands4)
68    .map(a_addDate_ny);
69
70
71  ///////////////
72
73  ////values of 0 are made an average between the two neighbours
74  //Back 5 days
75  function z_gs_dummy_back_f (image) {
76    var time_start = ee.Number(image.get('system:time_start'));
77    var time_start2 = time_start.subtract(5*86400000);
78    var NDVIsmooth5 = image.select('NDVI').rename('NDVI_back_z');
79    return image.set('system:time_start', time_start2).addBands(NDVIsmooth5);
80  }
81
82  var z_gs_dummy_back = data.select('NDVI').map(z_gs_dummy_back_f).map(a_addDate_ny);
83
84  //Back 10 days
85  function bz_gs_dummy_back_f (image) {
86    var time_start = ee.Number(image.get('system:time_start'));
87    var time_start2 = time_start.subtract(10*86400000);
88    var NDVIsmooth5 = image.select('NDVI').rename('NDVI_back_10z');
89    return image.set('system:time_start', time_start2).addBands(NDVIsmooth5);
90  }
91
92  var bz_gs_dummy_back = data.select('NDVI').map(bz_gs_dummy_back_f).map(a_addDate_ny);
93
94  //forward 5 days
95  function z_gs_dummy_forward_f (image) {
96    var time_start = ee.Number(image.get('system:time_start'));
97    var time_start2 = time_start.add(5*86400000);
98    var NDVIsmooth5 = image.select('NDVI').rename('NDVI_forward_z');
99    return image.set('system:time_start', time_start2).addBands(NDVIsmooth5);
100 }
101
102 var z_gs_dummy_forward = data.select('NDVI').map(z_gs_dummy_forward_f).map(a_addDate_ny);
103
104 //forward 10 days
105 function bz_gs_dummy_forward_f (image) {
106   var time_start = ee.Number(image.get('system:time_start'));
107   var time_start2 = time_start.add(10*86400000);
108   var NDVIsmooth5 = image.select('NDVI').rename('NDVI_forward_10z');
109   return image.set('system:time_start', time_start2).addBands(NDVIsmooth5);
110 }
111
112 var bz_gs_dummy_forward = data.select('NDVI').map(bz_gs_dummy_forward_f).map(a_addDate_ny);
113
```

```
115
116 ///////joiner
117 var filter_ny = ee.Filter.equals({
118   leftField: 'doy',
119   rightField: 'doy'
120 });
121
122 // Inner join
123 //gs and back5
124 var innerJoin_z1 = ee.ImageCollection(simpleJoin.apply(data, z_gs_dummy_back, filter_ny));
125
126 var growing_season_merged_1z = innerJoin_z1.map(function(feature) {
127   return ee.Image.cat(feature.get('primary'), feature.get('secondary'));
128 });
129
130
131 //and forward5
132 var innerJoin_z2 = ee.ImageCollection(simpleJoin.apply(growing_season_merged_1z, z_gs_dummy_forward, filter_ny));
133
134 var growing_season_merged_2z = innerJoin_z2.map(function(feature) {
135   return ee.Image.cat(feature.get('primary'), feature.get('secondary'));
136 });
137
138 //and forward10
139 var innerJoin_z3 = ee.ImageCollection(simpleJoin.apply(growing_season_merged_2z, bz_gs_dummy_forward, filter_ny));
140
141 var growing_season_merged_3z = innerJoin_z3.map(function(feature) {
142   return ee.Image.cat(feature.get('primary'), feature.get('secondary'));
143 });
144
145 //and back 10
146 var innerJoin_z4 = ee.ImageCollection(simpleJoin.apply(growing_season_merged_3z, bz_gs_dummy_back, filter_ny));
147
148 var growing_season_merged_4z = innerJoin_z4.map(function(feature) {
149   return ee.Image.cat(feature.get('primary'), feature.get('secondary'));
150 });
151
152
153
154 ////////
155 function forevi_z (image) {
156   var ndvi = image.select('NDVI');
157   var værdi_hvis_nul = (image.select('NDVI_back_z').add(image.select('NDVI_forward_z'))
158   .add(image.select('NDVI_forward_10z')).add(image.select('NDVI_back_10z'))).divide(4).rename('værdi_hvis_nul_z');
159   var ny_ndvi = ((((ndvi.subtract(0.00001).multiply(-1)).add((ndvi.subtract(0.00001))
160   .abs()).multiply(1/0.00002))).multiply(værdi_hvis_nul).add(ndvi)).rename('Ny_NDVI_z');
161   return image.addBands(ny_ndvi);
162 }
163
164 var data = growing_season_merged_4z.map(forevi_z);
165
166
```

```
//////////
//////Shift new timeseries back and forth 5 and 10 days
//back
//5
function a_gs_dummy_back_f (image) {
  var time_start = ee.Number(image.get('system:time_start'));
  var time_start2 = time_start.subtract(5*86400000);
  var NDVIsmooth5 = image.select('Ny_NDVI_z').rename('NDVI_back');
  return image.set('system:time_start', time_start2).addBands(NDVIsmooth5);
}

var a_gs_dummy_back = data.select('Ny_NDVI_z').map(a_gs_dummy_back_f).map(a_addDate_ny);

//10
function b_gs_dummy_back_f (image) {
  var time_start = ee.Number(image.get('system:time_start'));
  var time_start2 = time_start.subtract(10*86400000);
  var NDVIsmooth5 = image.select('Ny_NDVI_z').rename('NDVI_back_10');
  return image.set('system:time_start', time_start2).addBands(NDVIsmooth5);
}

var b_gs_dummy_back = data.select('Ny_NDVI_z').map(b_gs_dummy_back_f).map(a_addDate_ny);

//print('b_gs_dummy_back', b_gs_dummy_back)

//forward 5
function a_gs_dummy_forward_f (image) {
  var time_start = ee.Number(image.get('system:time_start'));
  var time_start2 = time_start.add(5*86400000);
  var NDVIsmooth5 = image.select('Ny_NDVI_z').rename('NDVI_forward');
  return image.set('system:time_start', time_start2).addBands(NDVIsmooth5);
}

var a_gs_dummy_forward = data.select('Ny_NDVI_z').map(a_gs_dummy_forward_f).map(a_addDate_ny);



//forward 10
function b_gs_dummy_forward_f (image) {
  var time_start = ee.Number(image.get('system:time_start'));
  var time_start2 = time_start.add(10*86400000);
  var NDVIsmooth5 = image.select('Ny_NDVI_z').rename('NDVI_forward_10');
  return image.set('system:time_start', time_start2).addBands(NDVIsmooth5);
}

var b_gs_dummy_forward = data.select('Ny_NDVI_z').map(b_gs_dummy_forward_f).map(a_addDate_ny);

```

```
///Joins

var filter_ny = ee.Filter.equals({
  leftField: 'doy',
  rightField: 'doy'
});

// Inner join
//gs and back
var innerJoin_ny1 = ee.ImageCollection(simpleJoin.apply(data, a_gs_dummy_back, filter_ny));

var growing_season_merged_1 = innerJoin_ny1.map(function(feature) {
  return ee.Image.cat(feature.get('primary'), feature.get('secondary'));
});

//and forward
var innerJoin_ny2 = ee.ImageCollection(simpleJoin.apply(growing_season_merged_1, a_gs_dummy_forward, filter_ny));

var growing_season_merged = innerJoin_ny2.map(function(feature) {
  return ee.Image.cat(feature.get('primary'), feature.get('secondary'));
});


//and 10 back
var b_innerJoin_ny1 = ee.ImageCollection(simpleJoin.apply(growing_season_merged, b_gs_dummy_back, filter_ny));

var growing_season_merged = b_innerJoin_ny1.map(function(feature) {
  return ee.Image.cat(feature.get('primary'), feature.get('secondary'));
});

//and 10 forward
var b_innerJoin_ny2 = ee.ImageCollection(simpleJoin.apply(growing_season_merged, b_gs_dummy_forward, filter_ny));

var growing_season_merged = b_innerJoin_ny2.map(function(feature) {
  return ee.Image.cat(feature.get('primary'), feature.get('secondary'));
});
```

```
//////////calculates and apply the EVI outlier removal measure

function forevi (image) {
  var middel_uden = (image.select('NDVI_back').add(image.select('NDVI_forward')).add(image.select('NDVI_back_10').add(image.select('NDVI_forward_10')))).divide(4);
  var varians_uden = ((image.select('NDVI_back').subtract(middel_uden)).pow(2)).add((image.select('NDVI_forward').subtract(middel_uden)).pow(2))
  .add((image.select('NDVI_forward_10').subtract(middel_uden)).pow(2)).add((image.select('NDVI_back_10').subtract(middel_uden)).pow(2)).rename('varians_uden');
  var middel_med = (image.select('Ny_NDVI_z').add(image.select('NDVI_back')).add(image.select('NDVI_forward')).add(image.select('NDVI_back_10')
  .add(image.select('NDVI_forward_10')))).divide(5);
  var varians_med = ((image.select('Ny_NDVI_z').subtract(middel_med)).pow(2)).add((image.select('NDVI_back').subtract(middel_med)).pow(2))
  .add((image.select('NDVI_forward').subtract(middel_med)).pow(2)).add((image.select('NDVI_forward_10').subtract(middel_med)).pow(2))
  .add((image.select('NDVI_back_10').subtract(middel_med)).pow(2)).rename('varians_med');
  var ndvi_middel_abs = (image.select('Ny_NDVI_z').subtract(middel_uden)).abs(); ///hvis abs() er på tager den også hvis nogle værdier er for lave
  var cutoff = 1;  //includes only those with negative measure larger than this.
  var measure = (((varians_uden.subtract(varians_med)).divide(varians_uden)).multiply(ndvi_middel_abs.divide(middel_uden))).add(cutoff).rename('Measure');
  var gør_nul = (((measure.add(measure.abs())).divide(2)).multiply(image.select('Ny_NDVI_z'))).divide((measure.add(measure.abs())).divide(2)).rename('Nul_NDVI');
  var værdi_hvis_nul = (image.select('NDVI_back').add(image.select('NDVI_forward')).add(image.select('NDVI_forward_10'))
  .add(image.select('NDVI_back_10'))).divide(4).rename('værdi_hvis_nul');
  var ny_ndvi = ((((gør_nul.subtract(0.00001).multiply(-1)).add((gør_nul.subtract(0.00001)).abs()).multiply(1/0.00002)))
  .multiply(værdi_hvis_nul).add(gør_nul)).rename('Ny_NDVI');
  //var ny2_ndvi = ((gør_nul.multiply(-1).add(værdi_hvis_nul)).add((gør_nul.multiply(-1).add(værdi_hvis_nul)).abs())).divide(2).add(gør_nul).rename('Ny2_NDVI')
  return image.addBands(measure).addBands(gør_nul).addBands(værdi_hvis_nul).addBands(ny_ndvi).addBands(varians_med).addBands(varians_uden);//.addBands(ny2_ndvi);
}


var growing_season_merged = growing_season_merged.map(forevi);

var data = growing_season_merged.select(['NDVI','Ny_NDVI_z', 'Nul_NDVI', 'Ny_NDVI', 'Measure', 'varians_med', 'varians_uden']);


```

This process is then repeated once more.

11.11.2 For the DL smoothings:

This code is inserted after the 10 MWLR smoothings and shortening of the season, and before finding the SOS and EOS.

```
////////Double logistic

var fitty_op = ((sæson.select('NDVIsmooth10').max()).subtract(forste_del_min)).multiply(0.5).add(forste_del_min);
var fitty_ned = ((sæson.select('NDVIsmooth10').max()).subtract(anden_del_min)).multiply(0.5).add(anden_del_min);

function for_s (image) {
  var smoorth = image.select('NDVIsmooth10');
  return image
  .updateMask(smoorth.gte(fitty_op));
}
function for_a (image) {
  var smoorth = image.select('NDVIsmooth10');
  return image
  .updateMask(smoorth.gte(fitty_ned));
}
var s = sæson.map(for_s).select('doy').min();
var a = sæson.map(for_a).select('doy').max();


//////finding slopes

//Increaing part
function for_stigning (image) {
  var smoorth = image.select('doy');
  return image
  .updateMask(smoorth.gte(s.subtract(8)))
  .updateMask(smoorth.lte(s.add(8)));
  }

var stigning = sæson.map(for_stigning);

var linearfit_stigning = stigning.select(['doy','NDVIsmooth10']).reduce(ee.Reducer.linearFit());

//decreasing part
function for_fald (image) {
  var smoorth = image.select('doy');
  return image
  .updateMask(smoorth.gte(a.subtract(8)))
  .updateMask(smoorth.lte(a.add(8)));
  }

var fald = sæson.map(for_fald);


var linearfit_fald = fald.select(['doy','NDVIsmooth10']).reduce(ee.Reducer.linearFit());

//// the actual DL smoothing

function for_dl (image) {
  var wndvi1 = forste_del_min.select('NDVIsmooth10');
  var wndvi2 = anden_del_min.select('NDVIsmooth10');
  var en = ee.Image(1);
  var minus_ms = linearfit_stigning.select('scale').multiply(-8);
  var ma = linearfit_fald.select('scale').multiply(-8);
  var doy = image.select('doy');
  var dl = wndvi1.add(maxndvi.subtract(wndvi1)).multiply((en.divide(en.add((minus_ms.multiply(doy.subtract(s))).exp()))))
  .add((maxndvi.subtract(wndvi2)).multiply(en.divide(en.add((ma.multiply(doy.subtract(a))).exp())).subtract(en))).rename('dl');
  return image.addBands(dl);
}

var smoothedndvi_time = sæson.map(for_dl);
```

11.11.3 For fixed:

Instead of using the VI value to find SOS and EOS, the fixed dates are used:

```
var smoothedndvi_time = merged.filterDate('2019-08-15', '2019-10-10');
```

11.11.4 For precipitation:

```
1209
1210  /////////Adds precip in the shortened window/////////
1211  var precip = Chirps.filterBounds(ROI);
1212  var precip = precip.map(addDate).map(for_vindue);
1213
```

Finds the EOS and SOS though the VI values as normal but uses the boundary dates as window for the precipitation. And also use triggering measure on precipitation

```
1282  //var growing_season = smoothedndvi_time.map(filter_SOS_EOS).select('NDVIsmooth10');
1283
1284  ///////////Shortems Precip with the found EOS and SOS///////////
1285  var growing_season = precip.map(filter_SOS_EOS).select('precipitation');
1286
1287
1288
1289  //////
1290  //then aggregate
1291  var accndvi = growing_season.mean();  ///triggering measure is here the mean
1292
1293  Export.image.toAsset({
1294    image: accndvi,
1295    description: 'Whole_Season_precip_mean',
1296    assetId: '08_06/Edday__nord1_w6_midtmaj-dec_usky_s_fraction_WS_02_inds_Precip_mean',
1297    scale: 10,
1298    region: AOI,
1299    pyramidingPolicy: {'.default': 'sample'}
1300  });
1301  Map.addLayer(table);
1302
```

11.11.5 For temperature:

Temperature is inserted after the smoothings. And it is smoothed once.

```
//////////Adds temperature and adds the doy as band ////////////
var dota = Temp.filterBounds(ROI).filterDate('2019-01-01', '2019-12-31');
var dota = dota.map(addDate);


//////then smoothingen
var timeField = 'system:time_start';
// function to add NDVI and time bands to image collection
var addDataBands = function(image) {
  return image.addBands(image.metadata('system:time_start').divide(1e18).rename('time'));
};
var dota = dota.map(addDataBands);
function osmoother(t){
  // helper function to apply linear regression equation
  function oapplyFit(img){
      return img.select('time').multiply(ofit.select('scale')).add(ofit.select('offset'))
                .set('system:time_start',img.get('system:time_start')).rename('T');
  }
  t = ee.Date(t);

  var window = dota.filterDate(t.advance(-windowSize,'day'),t.advance(windowSize,'day'));

  var ofit = window.select(['time','LST_Day_1km'])
    .reduce(ee.Reducer.linearFit());

  return window.map(oapplyFit).toList(5);
}

// function to reduce time stacked linear regression results
// requires that a variable 'fitIC' exists from the smooter function
function oreduceFits(t){
  t = ee.Date(t);
  return ofitIC.filterDate(t.advance(-windowSize,'day'),t.advance(windowSize,'day'))
               .mean().set('system:time_start',t.millis()).rename('T_smooth');
}


var dates = ee.List(dota.aggregate_array('system:time_start'));

var windowSize = 20; //days on either side

var ofitIC = ee.ImageCollection(dates.map(osmoother).flatten());

var osmoothed = ee.ImageCollection(dates.map(oreduceFits));

//// then joining
var filter = ee.Filter.equals({
  leftField: 'system:time_start',
  rightField: 'system:time_start'
});

// Create the join.
var simpleJoin = ee.Join.inner();

// Inner join
var innerJoin = ee.ImageCollection(simpleJoin.apply(dota, osmoothed, filter));

var dota = innerJoin.map(function(feature) {
  return ee.Image.cat(feature.get('primary'), feature.get('secondary'));
});

```

After the EOS and SOS is found by the VI values, they are used to shorten the temperature timeseries and the triggering measure is applied for temporal aggregation:

```
//////////Shortens temp with the found EOS and SOS//////////
var growing_season = dota.map(filter_SOS_EOS).select('T_smooth');

///triggering measure
var accndvi = growing_season.mean();

Export.image.toAsset({
  image: accndvi,
  description: 'Whole_Season_t_mean',
  assetId: '08_06/Edday__nordl_w6_midtmaj-dec_usky_s_fraction_WS_02_inds_Temp_mean',
  scale: 10,
  region: AOI,
  pyramidingPolicy: {'.default': 'sample'}
});

```

11.12 GEE script: Spatial aggregation

Script showing how the created images were turned into variables.

NORD *

```
Imports (7 entries)
  var table: Table users/n-b-p/Berber/CCE_luget_export
  var god: FeatureCollection (262 elements)
  var middel: FeatureCollection (48 elements)
  var tvivlsom: FeatureCollection (4 elements)
  var s2: ImageCollection "Sentinel-2 MSI: MultiSpectral Instrument, Level-2A"
  var ROI: Point (76.95, 29.96)
  var image: Image users/n-b-p/21_05_20/Edday_accndvi10smooth_nord_1_w6_midtmaj-dec_u_
var samlet = god;

////first sptial aggregation (not used, only for comparison)
var reduced = image
  .reduceRegions({
    collection:samlet ,
    reducer:ee.Reducer.mean(),
    scale: 0.1
    });

//// joins to yield datapoint

var joined = table.map(function(feat){
feat = ee.Feature(feat);
var grain_yield = feat.get('Grain_yiel');
var biomass_yield = feat.get('Biomass_yi');
var total_yield = feat.get('Total_yiel');
var Filter = reduced.filterBounds(feat.geometry()).map(function(reduced){
return ee.Feature(reduced)
.set('Grain Yield', grain_yield)
.set('Biomass Yield', biomass_yield)
.set('Total Yield', total_yield);
});
return Filter;
}).flatten();

///// Adds the internal buffer
var addbuffer = function(feature) {
  return feature.buffer(-5)};

var buffer = samlet.map(addbuffer);

/////Spatial aggregation inside buffer
var reduced_small = image
  .reduceRegions({
    collection:buffer,
    reducer:ee.Reducer.median(),
    scale: 0.1
    });

////joiner the two feature collections (all field and internal buffer)
var joined2 = joined.map(function(feat){
feat = ee.Feature(feat);
var grain_yield = feat.get('Grain Yield');
var biomass_yield = feat.get('Biomass Yield');
var total_yield = feat.get('Total Yield');
var mean_poly = feat.get('mean'); /////////det er værdien inden buffer
var Filter = reduced_small.filterBounds(feat.geometry()).map(function(reduced_small){
return ee.Feature(reduced_small)
.set('mean_poly', mean_poly)
.set('Grain Yield', grain_yield)
.set('Biomass Yield', biomass_yield)
.set('Total Yield', total_yield);
});
return Filter;
}).flatten();

```

```
///// creating the scatterplots with yield and VI value

//for total yield
var chart_t = ui.Chart.feature.byFeature ({
      features: joined2,
      xProperty: 'Total Yield',
      yProperties: ['median'] /////mean er værdien med buffer
})
     .setChartType('ScatterChart')
     .setOptions({
       title: 'Correlation in buffer',
       hAxis: { title: 'Total Yield'},
       vAxis: { title: 'VI measure' },
       lineWidth: 1,
       pointSize: 3,
       trendlines: { 0: {showR2: true, visibleInLegend: true}, 1: {showR2: true, visibleInLegend: true}}
});

print(chart_t);

// for grain yield
var Bornhold = ui.Chart.feature.byFeature ({
      features: joined2,
      xProperty: 'Grain Yield',
      yProperties: ['median'] /////mean er værdien med buffer
})
     .setChartType('ScatterChart')
     .setOptions({
       title: 'Correlation in buffer',
       hAxis: { title: 'Grain Yield'},
       vAxis: { title: 'VI measure' },
       lineWidth: 1,
       pointSize: 3,
       trendlines: { 0: {showR2: true, visibleInLegend: true}, 1: {showR2: true, visibleInLegend: true}}
});

print(Bornholm);

```

11.13 Spyder script – RF Classification

```
import csv
import numpy as np
from sklearn.ensemble import RandomForestClassifier
import matplotlib.pyplot as plt
from sklearn.model_selection import StratifiedKFold
from sklearn.feature_selection import RFECV

input = r"C:\Users\kwill\Desktop\EAST_All_total_5kategorier.csv"

dataTrain = np.loadtxt(input, dtype='str', delimiter=',')

labels = dataTrain[0,:]
x = dataTrain[1:, :-1]
y = dataTrain[1:, -1]

classifier = RandomForestClassifier(n_estimators=500, random_state=1, bootstrap=True)

rfecv = RFECV(estimator=classifier, step=1, cv=StratifiedKFold(n_splits=10, shuffle=True, random_state=1), scoring='accuracy')
rfecv.fit(x, y)

with open(r"C:\Users\kwill\Desktop\EAST_All_total_5kategorier_Routput_RFECV.csv", 'w', newline= '') as f:
    thewriter = csv.writer(f)

    thewriter.writerow(["Feature name"])
    thewriter.writerow(labels)
    thewriter.writerow(["Part of optimal combination?"])
    thewriter.writerow(rfecv.support_.tolist())
    thewriter.writerow([])
    thewriter.writerow(["Optimal combination - Feature and importance score"])
    label_collector = []
    for element in range(len(labels.tolist())-1):
        if str(rfecv.support_.tolist()[element])=='True':
            label_collector.append(labels.tolist()[element])
    thewriter.writerow(label_collector)

    thewriter.writerow(rfecv.estimator_.feature_importances_.tolist())


plt.figure()
plt.xlabel("Number of features selected")
plt.ylabel("Cross validation score (nb of correct classifications)")
plt.plot(range(1, len(rfecv.grid_scores_) +1), rfecv.grid_scores_)
plt.show()
```