Yield estimation in potato using WOFOST, satellite and ancillary data

MSc Thesis Plant Production Systems

Souravi Saha

October, 2024

WAGENINGEN
UNIVERSITY & RESEARCH

Yield estimation in potato using WOFOST, satellite and ancillary data

MSc Thesis Plant Production Systems

Name Student:	*Souravi Saha*
Registration Number:	*1217534*
Study:	*MSc Plant Sciences – Specialization Crop Sciences*
Chair group:	*Plant Production Systems (PPS)*
Code Number:	*PPS-80436*
Date	*October, 2024*
Supervisors:	*Tom Schut*
	Lammert Kooistra
Examiners:	*Gerrie van de Ven*

Disclaimer: this thesis report is part of an education program and hence might still contain (minor) inaccuracies and errors.

Correct citation: Souravi Saha, 2024, Yield estimation in potato using WOFOST, satellite and ancillary data, MSc Thesis Wageningen University, 41 p.

Contact office.pps@wur.nl for access to data, models and scripts used for the analysis

WAGENINGEN
UNIVERSITY & RESEARCH

Contents

Abbreviations used in this study

Y_p - Potential yield

Y_w - Water-limited yield

Y_a - Actual yield

avg_cnp_cvr - Average canopy cover values

NDVI- Normalised Difference Vegetation Index

ndvi_interpolated - Interpolated NDVI values

elevation - Ground elevation

clay_est - Estimate of clay percentage in soil

irri_quantity – total irrigation applied for the growing season

cluster_name - Name of each cluster

adj R^2 – adjusted R^2

r.m.s.e. - root mean square error

Abstract

Potato is an economically important crop in the Netherlands. Crop models are used often to estimate yields. WOFOST is one such well-known model commonly used to simulate potential (Yp) and water-limited (Yw) yields. Till date, a lot of variation in actual yield (Ya) cannot be captured by WOFOST. This study evaluates how much variation in fields can be captured by combining WOFOST, remote sensing and ancillary data.

This study addresses four research questions. Simple linear regression models were developed to understand how the WOFOST results can be improved by incorporating remote sensing and ancillary data. Efforts were made to study the models for different field problem categories. Three field categories were studied in detail: fields with (almost) no limiting or reducing factors, fields with reducing factors and fields from all categories combined together. Ancillary data was added incrementally to the regression models. Finally, a random forest algorithm was used to evaluate how much field variation could be captured with and without ancillary data.

I conclude that rational inclusion of satellite and ancillary data to WOFOST results can improve the estimation of actual yields (Ya), although it differed for each field problem category. This research clearly shows the potential of adding more data in future studies to further explain the variation in the actual yields (Ya).

1. Introduction

1.1 Status of potato production in the Netherlands

The potato crop originates from the mountains of South America, with the earliest records dating back to 2500BC in modern day Peru and Bolivia ("Potatoes on Mars," n.d.), made its way to Europe in the 16th century (Beukema and Van Der Zaag, 1990). Currently, potato is the third most important global food crop after rice and wheat which is consumed by over a billion people (Gómez et al., 2019).

Dutch agriculture changed dramatically from the 1960s with specialisation and intensification becoming the means in terms of production to achieve food self-sufficiency in the EU (Vos, 1992). The Netherlands produced 6.7 million tonnes of potato in 2021 from 159 thousand ha of land (FAO, 2022), the ninth position among the top potato producing countries in the world. The Netherlands is also the leading exporter of 'certified' seed potatoes, accounting for 26.2% of the total potato production in the country (Goffart et al., 2022). Potato production occurs in 16% of the country's arable land, which is the highest allocation fraction to potato crops among different countries of north-western Europe (Goffart et al., 2022). The provinces of Drenthe, Groningen, North Brabant, Zeeland, and Flevoland contribute to 70% of the potato production in the Netherlands. Ware potatoes are meant for direct human consumption, owing to their low starch content, high water content and firmer texture. Ware potato is predominantly grown on the sandy and loamy soils in the Polders and the south-east of the country (Vos, 1992). The global potato chips market is expected to increase at the rate of almost 4% per year (Edelenbosch and Munnichs, 2020), emphasizing a need for increasing the amount of current potato production.

The farming of ware potato in the Netherlands is a highly productive crop with large among field heterogeneity (Ravensbergen et al., 2024). This indicates that there is a potential for improving yields and thus farmer's income (Ravensbergen, 2024).

1.2 Potato crop yield estimation

Yield is often recognised as an indicator for the economic pillar for measuring sustainability of farming systems (as yield is directly related to the income of the farm). The often interchangeably used terms, estimation, and prediction, have a subtle difference in between them. 'Estimation' refers to the process of determining the most appropriate values for model parameters or coefficients, such as regression coefficients in statistical models. 'Prediction' refers to the value that is produced by the model. 'Estimation' of an event happens after the occurrence of the event whereas 'prediction' of an event happens before the event occurred. On the other hand, 'forecasting' is a term which predicts an event definitely on the temporal scale. Yield estimation relies on current season crop monitoring using satellite data whereas yield prediction mostly on historical data. Yield estimation is often more accurate as it is based on the current field conditions. Yield estimation can be used for evaluating agricultural performance of crop, planning the post-harvest activities, etc.

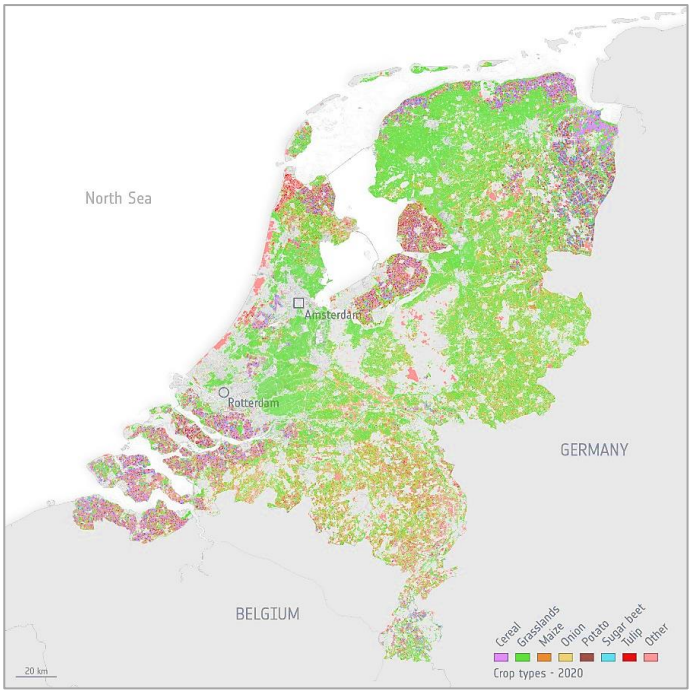


Figure 1: Major crops grown in the Netherlands (ESA, 2020) The brown regions represent the potato growing regions in the Netherlands.

The difference between potential yield (without any water or nutrient constraint) and the actual yield (in farmer's field conditions) or the difference between water-limited yield (water constraint) and the actual yield represent the yield gap. This is important to identify to understand the reducing factors responsible for reducing the water-limited yield and rectify to minimize the latter yield gap. The actual yield can be improved by yield increasing (using non-substitutable inputs like water and nutrients) and yield protecting (using substitutable inputs like pesticides, to some extent) measures (Van Ittersum and Rabbinge, 1997).

1.3 Production ecology

In production ecology, there are three levels of production factors which influence the production. These factors that play an important role in the crop growth process.

Growth defining factors determine the maximum production (i.e., potential yield) that can be achieved in a given physical environment and for a plant species. Such governing factors include radiation intensity, carbon dioxide concentrations and temperature (Van Ittersum et al., 2003). Potential Yield is referred to as Y_p in this study.

Growth limiting factors determine production possible with limitation in factors like water and/or nutrients (attainable yield) in a particular physical environment for a plant species. Proper management of water and nutrients can help to reach potential production levels. Water-limited Yield is referred to as Y_w in this study.

Growth reducing factors determine production levels which can reduce or hamper growth (actual yield). These factors are biotic include pests and diseases, weeds (biotic) and pollution. Crop protection measures can help reach the attainable yields. Actual Yield is referred to as Y_a in this study.

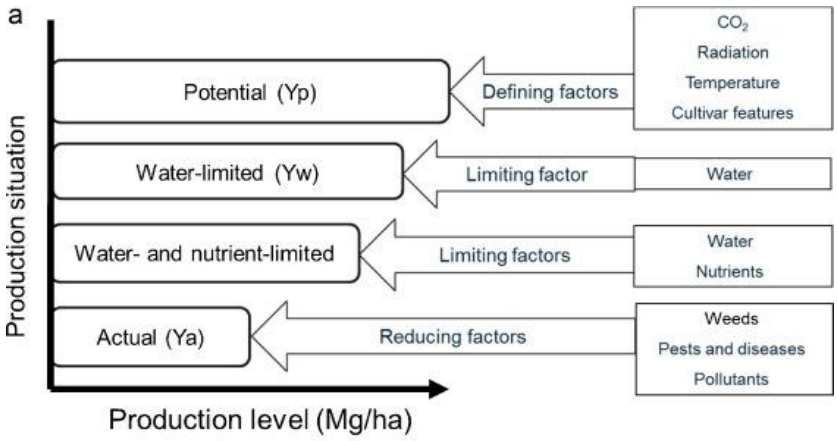


Figure 2: Growth factors in the three production situations in production ecology (Van Ittersum et al., 2013)

These production ecological situations form the basis for crop models like WOFOST. Currently, WOFOST can provide simulated yield values for the potential, water limited and nitrogen limited yields.

1.4 Crop yield models and WOFOST

1.4.1 Crop Yield Models

Various approaches have been applied in history to estimate yield. Statistical models require fewer inputs than mechanistic models making it more suitable for data-limited applications. However, statistical models fail to explain yield due to the absence of knowledge about the underlying biological processes involved within the plant system (Divya et al., 2020). The mechanistic nature of crop simulation models works as a solution to this shortcoming. According to (Bouman et al., 1996), "Crop simulation models consists of non-linear mathematical equations and logic to analyse the crop production system". The word 'logic' in the previous statement is of significant importance as it differentiates between statistical and mechanistic models. Thus to incorporate the information of both types of models, the integration of statistical and simulated outputs from mechanistic crop models requires further investigation.

Crop growth models can capture field processes including soil water and nutrient dynamics, plant water and nutrient uptake, biomass development and final crop yield with much accuracy if average soil information is fed into the models. Models with nutrient dynamics are not as commonly found as with water dynamics and it is nitrogen that is simulated in these models.

There are many crop models have been developed over the last few decades like DSSAT, APSIM, EPIC. There are also specialised crop models for potato like Lintul-Potato, INFOCROP-POTATO, SUBSTOR, POTATO, SWACO (Beukema and Van Der Zaag, 1990), etc. Recently , the WOFOST input parameters for new potato cultivars have been calibrated for WOFOST (Den et al., 2022). This was the reason for choosing WOFOST over other models for this study.

1.4.2 WOFOST

WOFOST is based on the principles of SUCROS model. WOFOST is one of the many models developed by the Wageningen 'School of de Wit' models. This model has been used in the MARS crop yield forecasting system for more than 25 years (de Wit et al., 2019).

WOrld FOod STudies (WOFOST) model was originally developed by the Centre of the World Food Studies in Wageningen for developing countries (Diepen et al., 1989). Eventually, it was utilised for Europe also (WOFOST version 6.0) as the biophysical part of the model remains the same in principle (Diepen et al., 1989). WOFOST explains daily growth on the basis of underlying processes like photosynthesis, respiration and their interaction with the environment. WOFOST has crop, soil and weather and ASTRO modules. Water limited yield production simulation of WOFOST is attached to a soil water balance model to keep track of the moisture content of the soil. Over the entire growing period, the attainable crop growth, biomass and water use can be simulated.

WOFOST uses a time step of one day for its own calculations (de Wit et al., 2019). It generates output data for an entire growing season, right from emergence till maturity (Diepen et al., 1989). The potential production simulation output results are listed for every tenth day of the growth cycle. WOFOST has only one spatial dimension and is considered as a point analysis (Diepen et al., 1989). It is, therefore, essential to scale all the parameters to the same spatial resolution before inputting them in the crop model. In table 1, it is possible to see which approach was used in the WOFOST model to simulate different yield levels. For the nutrient limited yield simulation, QUEFTS is applied to the simulation results and hence not part of the WOFOST model directly.

Table 1: The three production situations in WOFOST and their main governing principle (Van Ittersum et al., 2003)

Production situation	**Governing approach**
Potential	Photosynthesis
Water limited	Tipping bucket
NPK limited	QUEFTS

SUCROS was mainly used for research purposes, WOFOST was designed more rigorous, having clear version control and proper documentation. WOFOST can simulate different crop types by changing parameter values external to the model itself (de Wit et al., 2019). The WOFOST model is being developed continuously and there will be an effort to make the simulations are close as Y_a.

1.5 Satellite data to estimate yield

The sole use of crop growth models for yield estimation is often costly and time consuming owing to the requirement of large amount of inputs for the model employment, which relies on extensive data collection (Luo et al., 2020), (Kasampalis et al., 2018). Crop models can be used for tactical decision making (Bouman et al., 1996). Linear regression models along with crop model outputs, and remote sensing (RS) indicators are extensively used in crop forecasting studies due to their simplicity and interpretability (Paudel et al., 2023). Computational modelling for estimating yield is much dependent on the past data and is unable to detect when there are sudden changes in climate, soil, irrigation, and cultivation. Moreover, WOFOST does not take into account the yield variability due to pest attack or different agronomic management changes (Diepen and de Wit, n.d.). Remotely sensed data can provide valuable information to crop models to improve yield predictions (Kasampalis et al., 2018). Remote sensing data can give more accurate description of the crop's actual condition during different stages of the crop growing period (Kasampalis et al., 2018). This can be attributed to multiple reasons including agronomic management practices, pests and diseases, which are not simulated by crop models like WOFOST. Remote sensing is attractive to the scientific world owing to its 'non-destructive, high -throughput, and having large spatial coverage' (Lin et al., 2023). Handheld multispectral sensors, used for yield prediction, remains a challenge when yield variability of fields needs to be studied (Lin et al., 2023). Potato crop biomass is a function of Leaf Area Index (LAI). Lower LAI values indicate reduction in yield due to low defining or limiting factors or some form of stress.

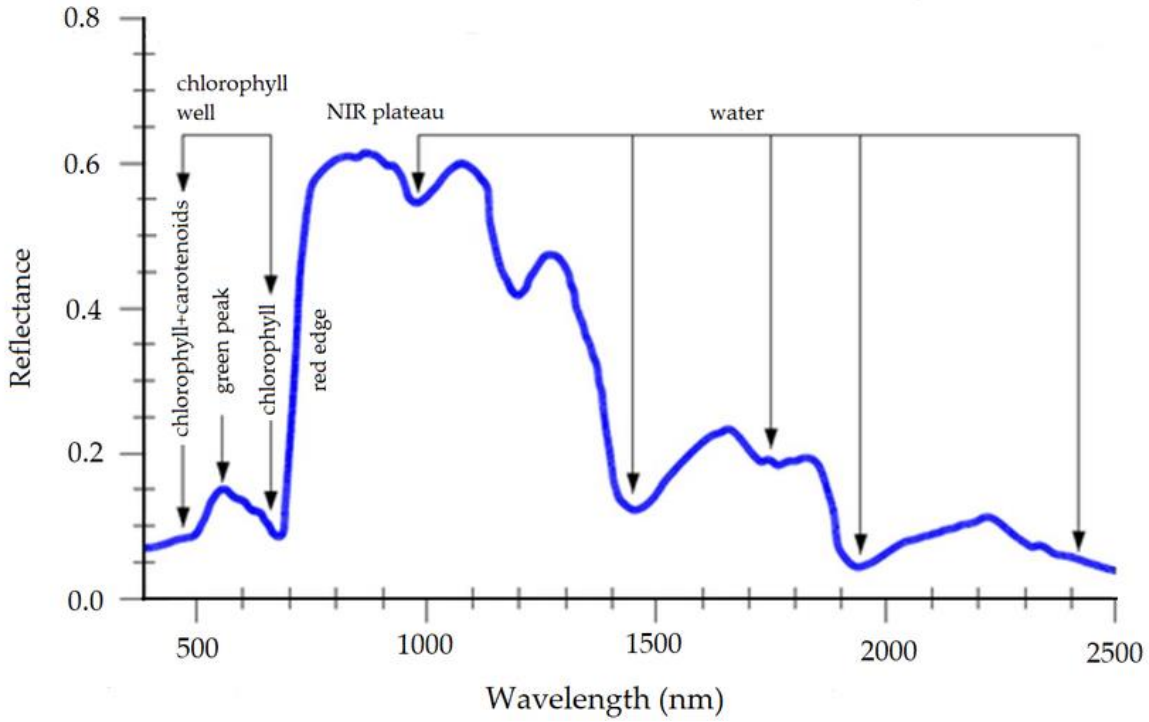


Figure 3: Spectral response of typical vegetation (Moroni et al., 2019)

LAI is an important biophysical parameter that controls canopy processes (Herrmann et al., 2011). Normalised Difference Vegetation Index (NDVI) is strongly related to various biophysical characteristics of plant like leaf area index (LAI), chlorophyll content, fractional cover, dry and wet biomass and physiological processes which depend on light interception like yield, net and gross primary productivity (Glenn et al., 2011). (Herrmann et al., 2012) studied Partial Least Square analysis for predicting LAI, where Sentinel-2 NDVI predicted LAI with high r-value of 0.89 for potato crop. The LAI prediction was evaluated by the correlation coefficient (r) value of the relation between the predicted and the observed LAI values. Satellite data can capture the variability for wheat yield at farm level (Nain et al., 2012). (Nguyen et al., 2022) used NDVI and NDWI indices derived from the Sentinel 2 data to predict yield in canola. However, the limitation of using NDVI is that it reaches 'saturation' with LAI more than 2. Thus, it might be impossible to predict very high values of LAI using NDVI (Herrmann et al., 2012). At the same time, even if NDVI has been used as a proxy for ground cover in many studies (Schut et al., 2018), there are more processes that determine yield which needs to be combined with crop models to capture processes at play during the grain filling stage.

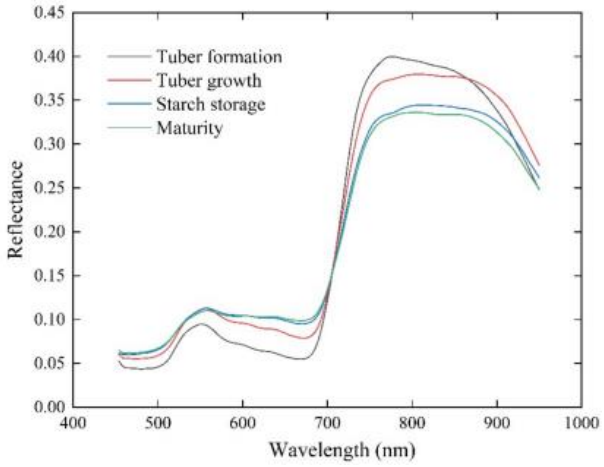


Figure 4: The variation in reflection spectra for four potato crop phenological stages (Liu et al., 2022). This explains the usefulness of NIR bands to predict yield.

NDVI is highly sensitive to soil background, playing a large role at low LAI and becomes less sensitive to soil background when LAI is usually around three when soil coverage is nearly complete (Wu et al., 2007). Although soil-adjusted vegetation index can reduce the effect of soil variability for low LAI and increase the sensitivity of high LAI, determining the optimal value of soil adjustment factor L in the formula requires the field information about LAI or vegetation density.

Several spectral indices have been used to study agricultural crops in different aspects. In table 2, NDVI is used to study overall crop vigour while NDWI is used to study water stress. NDWI is sensitive to changes in liquid water content of vegetation canopies (Gao, 1996). Lower NDWI values indicate water stress. There are specific spectral indices being developed for potato crop. (Gómez et al., 2021) developed a Potato Productivity index (PPI) which represents photosynthetic activity and water stress together. The water band in the PPI index is band 9 of Sentinel 2 (R_{945}). Higher PPI values indicate less stressing conditions for potato plants.

Table 2: Spectral indices, their functions, equations and their relationship to WOFOST. Note: B, G , R, NIR, RE1, WA represents Blue, Green, Red, Near infrared, Red Edge 1, and Water bands respectively.

Satellite data vegetation indices	Function	Equation	WOFOST parameter/output related to it
NDVI	General crop vigour	$NDVI = \frac{NIR - R}{NIR + R}$	Biomass through LAI
NDWI	Water stress	$NDWI = \frac{G - NIR}{G + NIR}$	Water limited crop growth
PPI	Photosynthetic ability and water stress	$PPI = \frac{G}{B + R} + \frac{WATER}{NIR} \times \frac{G}{RE\,1}$	Combined effect of NDVI and NDWI on WOFOST

(Al-Gaadi et al., 2016) utilised satellite vegetation indices NDVI and Soil Adjusted Vegetation Index (SAVI) to develop empirical models to determine yield. A lower prediction error range was demonstrated by Sentinel 2 (3.8-10.2%) with 10m spatial resolution, as compared to Landsat 8 (7.9-13.5%), having 30m spatial resolution, for predicting potato yield although there was not any improvement in the R^2 value for the models.

Potato spectral reflectance has been used as a representation for plant health in earlier studies (Po et al., 2010). Red and NIR are the commonly used bands for vegetation spectral indices as they are related to the photosynthetic process of the crop. Differences in vegetation cover can be often related to the difference in planting dates. Spectral bands can capture a difference of 2 week in planting dates in two fields (Po et al., 2010). Thus, inclusion of satellite data can add information through various means.

1.6 Ancillary Data

Ancillary data refers to any supplementary data apart from the primary data. In this study, ancillary data refers to all data apart from crop model outputs and satellite data. Integration of agronomic, management and meteorological information with remote sensing data can improve yield estimation capability (Lin et al., 2023). For example, potato yield depends on the characteristics of soil significantly. A lot of data has been used previously to understand the drivers of yield gap in the Netherlands (Silva et al., 2020). Soil compaction has been mentioned to be one of the significant drivers of yield heterogeneity. This can be included in the models to improve the estimation results.

Yield was less responsive to irrigation when the rainfall is relatively high and the supplemental irrigations were late (Porter et al., 1999). According to (Po et al., 2010), the reason for the earlier senescence in one of the fields is not only because of early planting but also due to moisture stress. Potato tuber is a commodity with a high moisture content and hence soil moisture and precipitation patterns remain as important factors attributed to yield variability (Po et al., 2010). Potato crop differs from other crops as the economically useful harvest is found

underground (Lin et al., 2023). This makes the study challenging as optical satellite data can capture information for aboveground objects.

1.7 Problem definition

SWAP-WOFOST model has been used to simulate potato yield in the Netherlands. However, some sources of temporal and spatial variability cannot be captured by crop growth models. Remotely sensed vegetation indices, such as NDVI respond to variations caused by pest attack or agronomic management practices that are not accounted for in SWAP-WOFOST like crop growth models. This study will investigate whether remote sensing and ancillary data can help to capture information which cannot be explained by SWAP-WOFOST.

This study will also investigate whether non-linear models can really bring in light things that cannot be explained by simple regression models. Random forest is such a way among many others that might help us to find out informative insights in the data and help to understand the drivers even better. However, it is important to analyse whether the need of non-linear models is required, or linear models are enough to represent the data. If non-linear models are used when not required, there can be overfitting issues which can negatively affect the model performance (Desloires et al., 2023).

1.8 Aim and Research questions

The aim of this research is to better capture yield variation in fields using combinations of SWAP-WOFOST, satellite and ancillary data (soil, irrigation and elevation data).

This study addresses four key research questions as follow:

RQ 1. Can remote sensing data explain the variability in actual yield which cannot be explained by SWAP-WOFOST outputs alone?

RQ 2. Can ancillary data explain the variability in actual yield which cannot be explained by SWAP-WOFOST outputs alone?

RQ 3. Can remote sensing and ancillary data explain the variability in actual yield which cannot be explained by SWAP-WOFOST outputs alone?

RQ 4. Can non-linear model like Random Forest improve the linear regression model containing SWAP-WOFOST, remote sensing and ancillary data?

1.9 Hypothesis

Hypothesis 1- Satellite derived NDVI values are strongly correlated with field measured ground cover values.

Hypothesis 2- The regression model involving the SWAP-WOFOST outputs, the remotely sensed data and the ancillary data is expected to perform best (have the highest R^2 and the lowest r.m.s.e.) among the regression models tested.

2. Materials and methods

2.1 Location description

The data collection took place from 96 established commercial ware potato farms in the years 2020-21. The fields were spread across the six most important potato growing regions of the Netherlands, namely, Tholen/ West Brabant, Zuid Holland, Flevoland, North Brabant, Drenthe, and Limburg. Eight potato fields per region per year (a total of 48 fields per year) were selected to maintain homogeneity in the number of sample fields in each region. Fields in the first three regions had a clay soil texture where the potato variety grown was Innovator while the latter three regions have sandy soil with the variety Fontane. The varieties were chosen as they were the main variety used in each soil texture. Two of the farms did not have crop registration data, hence, could not be simulated with SWAP-WOFOST. Fields were visited once every two weeks from planting to harvest, hence 10-13 visits per field in one growing season (Ravensbergen, 2024).

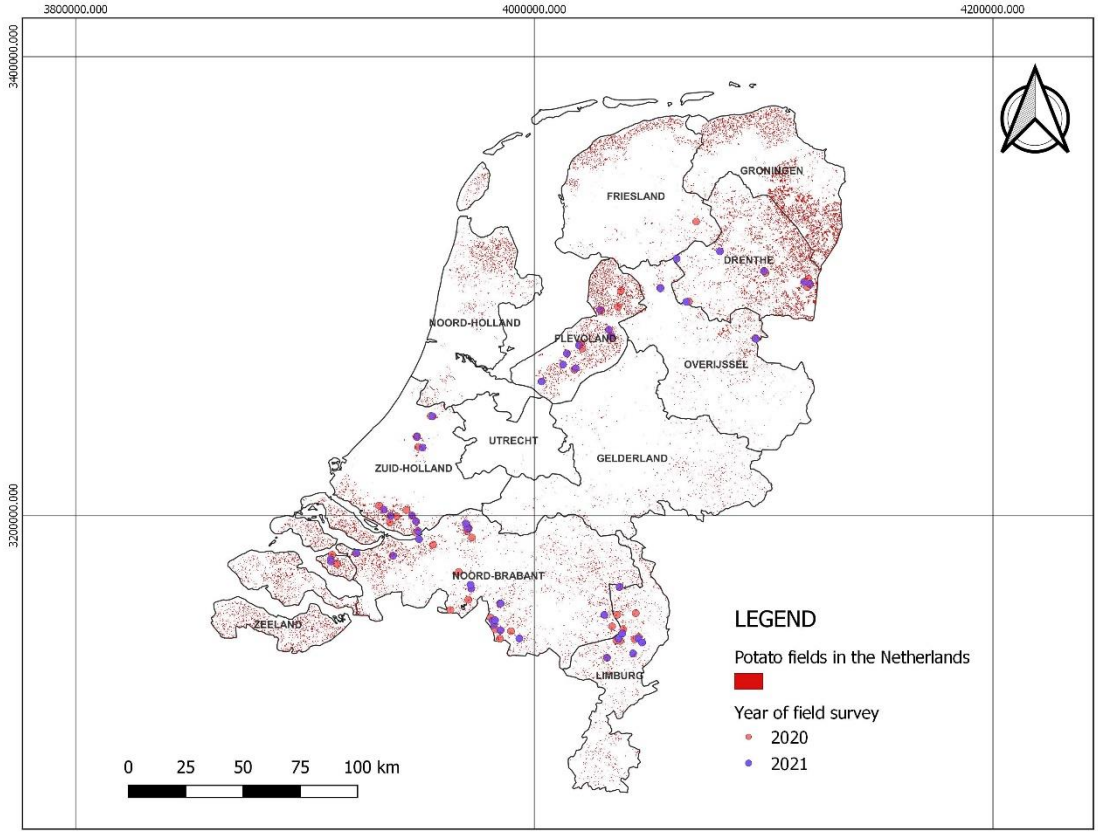


Figure 5: Map of the study area. Potato growing fields in the Netherlands and the field locations. The year in the legend indicate the year in which the field survey was conducted.

2.2 Materials used

Three types of data was used for the study- field , satellite and ancillary data. Field survey was conducted by Paul Ravensbergen as a part of his thesis work (Ravensbergen, 2024). The SWAP-WOFOST crop simulation yield outputs were also done by him.

Field data: Actual yield data, irrigation data, cultivar data and ground cover data were collected from field measurements. There was total 96 fields studied in the research, out of which 94 were used for SWAP-WOFOST simulations due to lack of crop registration information about the two fields. There were four replicates for each farm to measure actual yield and ground cover by potato crop. The agronomic management practices were the same for all the four replicates. The yield was measured during and at the end of the growing season. In this study, only the final yield was used as the actual yield. The measurement for the yield at final harvest followed

the same procedure for both the years. Final yield sampling was done on four $3m^2$ area after haulm killing or natural senescence, or just before the harvesting by the farmer in case the haulms had not senesced (Ravensbergen, 2024). A six kg subsample per plot to measure underwater weight, which was then recalculated to dry matter concentration. The data used is the Dry weight marketable yield and the unit is tonnes/hectare. Dry weight of actual yield data was used as water content might not be the best representation of the modelled yields. Dry weight eliminates the chances of fluctuations of varying moisture content of the crop, leading to more confusion in the dataset.

Both the varieties used in the study are for primarily used for French fry processing. The yellow-fleshed Fontane is the most important variety in the Flanders region, with 50% of the potato yield production volume. Fontane once replaced Bintje due to its resistance to potato cyst nematode *Globodera rostochiensis.* Innovator, the white fleshed potato, is resistant to another potato cyst nematode *Globodera pallida* to which Fontane is sensitive. However, Innovator cannot be used everywhere in the Netherlands due to its low underwater weight. Innovator is not recommended for sandy soils as it has a high chance of developing rust and brown spots. Innovator is also drought sensitive which can affect the yield (Demin, 2021).

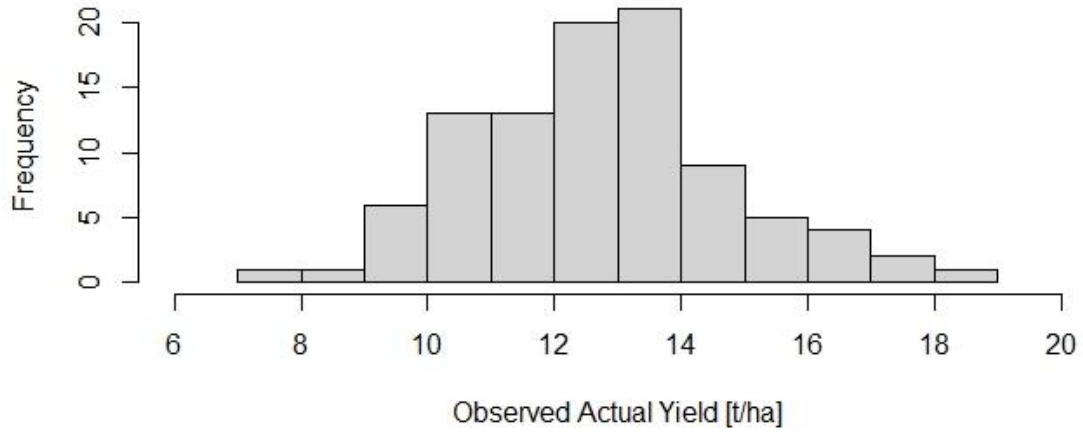


Figure 6: Histogram of the field observed actual yields. (Source: Field work from (Ravensbergen, 2024))

Ground cover was measured over the entire growing season at an interval of 15 days with four measurements per plot. The four plots were averaged for each field in this study. So, first the four measurements per plot was averaged followed by the average of the four plots to derive an average Canopy Cover value for each date. The number of emerged plants per plot was counted before full canopy closure. Crop health was scored from 1 to 5 by visual interpretation. A score of 5 represents healthy crop while a score of 1 is a diseased crop. The crop health score was averaged for the entire season.

Irrigation data was collected from the farmers. The irrigation schedule including the amount of irrigation provided at each time interval (irrigation event) were asked from the farmer. The amount of irrigation provided were summed up for the entire season for each field which was used as ancillary data. The percentage of clay in the soil was also included as ancillary data.

Satellite data: Pre-processed Sentinel 2 data available on Google Earth Engine (GEE) platform were used. The Sentinel-2 image collection named "COPERNICUS/S2_SR" was used for surface reflectance data. Excel (.csv) file containing all the field geographical coordinates were converted to point shapefile before it was ingested on GEE as an asset. Firstly, the clouds were masked in the Sentinel-2 images followed by calculating the NDVI band. Lastly the NDVI values for the field locations. It was necessary to choose the starting and the ending date of the required satellite data. The earliest planting date for 2020 and 2021 were 26th March (field ID44) and 31st March (field ID 115) respectively. In some fields, haulm killing occur days or weeks before the harvesting of the potatoes and for some of the fields, plants are killed during the harvest process. Hence, the harvesting date had been preferred over the crop end date as the variable for choosing the ending date of the satellite data. The latest harvesting date (which can be beyond the crop end date) for 2020 and 2021 were 7th November and 10th November respectively. Hence, to maintain homogeneity of the datasets, the datasets for NDVI extraction were collected

between 26th March and 11th November for both the years. The NDVI values were exported in csv. format. This was used in R software for further data analysis.

The mean elevation data is extracted from the AHN (*Actueel Hoogtebestand Nederland*) DEM (Digital Elevation Model) of the Netherlands 0.5m resolution Lidar data available on GEE platform. The dataset has only one band named 'elevation' where the unit is metre. It contains ground level samples with all other items above ground (such as buildings, bridges, trees etc.) removed. Elevation data was extracted for all the field locations and exported as .csv file for further analysis in R software.

Crop model data: Simulated SWAP-WOFOST outputs for 94 fields had been used. The outputs comprised of both the potential and the water limited yields for all the fields over the entire growing period. The final yield value for the fields were extracted from the entire WOFOST output over the growing season for both Y_p and Y_w and used for the research. In this study, irrigation is included in the water-limited yield. To understand the data, preliminary statistical study of the data was done as presented in Tables 3, 4 and Figure 7.

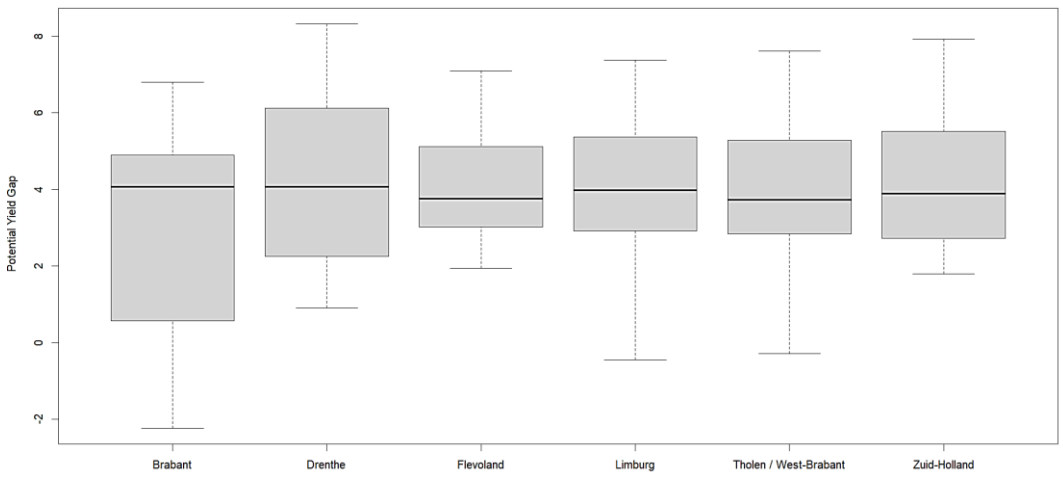
It is seen that the median potential yield gap (Figure 7(a)) is almost similar for all clusters (around 4t/ha) as compared to water-limited yield gap. In Figure 7(b), the median yield gaps for Flevoland and Zuid-Holland is almost around zero, which means that the Yw is almost near Ya.

Table 3: Average dry matter yields of Fontane and Innovator varieties of potato on SWAP-WOFOST modelled data and field observed data. (Source: Field work from (Ravensbergen, 2024))

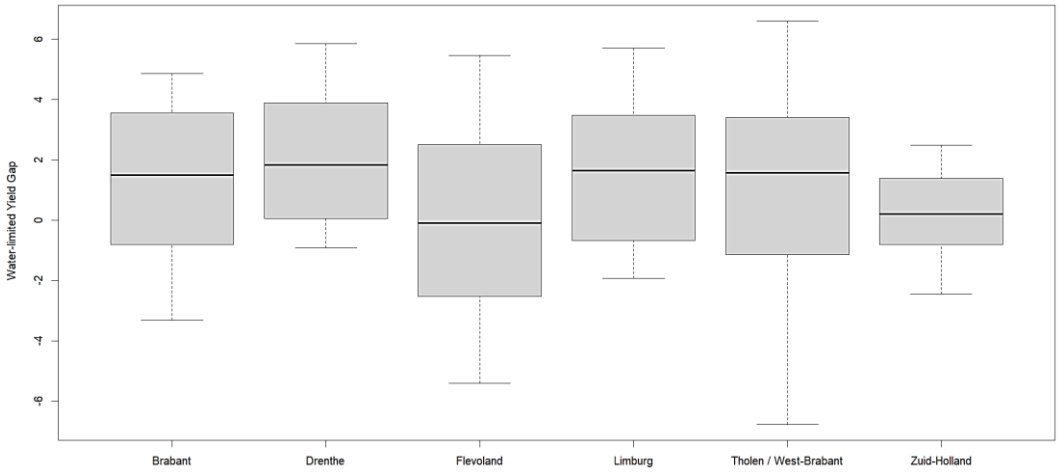
Variety	Yp (WOFOST) (t/ha)	Yw (WOFOST) (t/ha)	Ya (Field observation) (t/ha)
Innovator	16.06	12.34	12.01
Fontane	17.10	15.09	13.39

Table 4: Yield gap statistics of the six clusters on Yp and Yw SWAP-WOFOST modelled yields (t/ha). (Source: Field work from (Ravensbergen, 2024))

Cluster name (16 fields per clusters)	Mean YieldGap (Y_p-Y_a)	Mean YieldGap (Y_w-Y_a)	Standard Deviation YieldGap (Y_p-Y_a)	Standard Deviation YieldGap (Y_w-Y_a)
Brabant	2.99	1.25	2.90	2.76
Drenthe	4.18	2.21	2.17	2.24
Flevoland	4.05	-0.04	1.52	3.37
Limburg	3.99	1.65	2.08	2.59
Tholen / West-Brabant	3.84	0.84	2.07	3.44
Zuid-Holland	4.26	0.16	1.78	1.45



(a)



(b)

Figure 7: Boxplots of Yield gaps of (a) potential (Y_p-Y_a) and (b) water-limited (Y_w-Y_a) yields for each cluster. (Source: Field work from (Ravensbergen, 2024))

Ancillary data: Total irrigation quantity (mm) data, clay estimate (percentage) and elevation data (m) have been used as ancillary data in the study. The data collection for the ancillary data has been discussed earlier as part of the field data collection.

Table 5: Materials used for the study

Data	Material used	Variable attribute
Satellite data	Sentinel-2 (European Space Agency)	Spectral index NDVI
	AHN (*Actueel Hoogtebestand Nederland*) DEM (Digital Elevation Model) of the Netherlands	Elevation data
Crop model data	SWAP-WOFOST data	WOFOST potential and water limiting yields
Field data	Soil data	Estimated clay percentage based on farmer survey or soil map
	Actual yield (Y_a) data	Field observation
	Irrigation data	Field observation

2.3 Methodology

2.3.1 Method

This study will analyse relationships between a set of independent variables and actual yield as the dependent variable. Firstly, a simple linear regression model was developed between SWAP-WOFOST yields and actual yields (model 1). Colours and shapes were used to differentiate between the six clusters and the two varieties respectively in Figure 8. The clusters are the different regions of the Netherlands representing differences in soil texture and weather conditions. The same model was used to distinguish between five field situation categories discussed in the next paragraph. This was necessary to have a general understanding of the data and how well SWAP-WOFOST outputs correlated to Y_a. Next, as the goal of the study was to improve SWAP-WOFOST estimations using other forms of data, the fields were divided into categories based on their field situations. Firstly, the five categories were divided into two major categories-normal and problematic fields. Normal fields included fields which had Ya either close to Yw or Yp. Problematic fields, on the other hand, comprised of two categories which are either not yet modelled by SWAP-WOFOST or is erroneous. We are interested in improving the estimation of yields of problematic fields which are currently not well represented by the SWAP-WOFOST outputs.

There were five categories of yield reducing situations observed in fields. The category 'Accurate potential' refers to the situation when modelled potential yield was equal to the field observed actual yield. 'Accurate oxygen' refers to the situation when the modelled water-limited yield was equal to the field observed actual yield. The limitation in this case was the deficiency of oxygen. Oxygen stress is created by waterlogging which is common in clayey soils. 'Accurate drought' refers to the situation when the modelled water-limited yield was equal to the field observed actual yield. The limitation in this case was the stress by drought. 'Reducing factor' category refers to any other production factor limitation for the situation when the modelled water-limited yield was equal to the field observed actual yield. This can refer to factors which reduces the yield other than drought stress and oxygen deficiency like pests, diseases, weeds, etc. Lastly, 'erroneous' category refers to situations when field observed actual yield was higher than modelled yield (both potential and water-limited). Minor differences of 0.1 ton were neglected. There was a total of 94 fields where 48 fields were part of 2020 and 46 fields were part of 2021.

Table 6: Field problem category with number of fields in each year

Problem category	**2020**	**2021**
Accurate potential	8	7
Accurate oxygen	6	14
Accurate drought	12	1
Reducing factors	11	19
Erroneous	11	5
Total (All_fields)	**48**	**46**

It is possible to simulate the first three categories using SWAP-WOFOST with apperciable accuracy. However, the fourth category of reducing factors is not well simulated by SWAP-WOFOST and hence can NDVI can be helpful here. As mentioned earlier, the NDVI is thought to be able to provide information about the real conditions on the field which cannot be captured by SWAP-WOFOST solely. The fourth category of field problems represents factors reducing actual yields. This can be weeds, diseases, etc. The fifth category is the erroneous category which was also studied a bit in this study to see if NDVI can predict the behaviour of the fields in this category. There is all_fields cateogory used in the models which combines all the categories together.

The Canopy Cover evolution for a growing season was studied for the different field problem categories. Firstly, the Canopy Cover evolution was done for normal and problematic categories and then for all the problem categories seperately. . Normal field include the 'accurate_potential', 'accurate_drought' and 'accurate_oxygen' categories while problematic fields include the 'reducing_factors' and 'erroneous' categories. This was studied to identify if it possible to differentiate between the categories on the basis of Canopy Cover change over time.

The measured Canopy Cover is then added to the SWAP-WOFOST outputs to see if the Canopy Cover variable has some added information for the actual yield estimate (Model 2). The Canopy Cover from the entire season is included in the model. The rationale here is that Canopy Cover is the representation of the actual conditions of the field which are not addressed by the modelled water limited yields.

The next step involves the derivation of the NDVI values. The rationale is that NDVI values are easier to derive than ground cover due to its availability from remotely sensed data rather than laborious and time-consuming field measurements of ground cover. Figures 12 and 13 contain all the datapoints for all dates of canopy capture. As it is not possible to always have an NDVI for a date when there was a Canopy Cover observation, NDVI values were interpolated for all the dates in between the actual NDVI values when the NDVI values could be captured. NDVI may not be available on certain days because there may not be satellite taking photo on that day or that day was too cloudy and that part of the image was masked.

This step was followed by the question whether ground cover can be replaced by NDVI by running a correlation analysis (the higher the correlation, the better the replaceability). The hypothesis at this stage is that NDVI and ground cover are highly related to each other as NDVI is highly dependent on the reflection in the NIR and the R bands, which act as unique indicator of vegetation in remote sensing studies. Once this is established, model

with Y_a and Canopy Cover was compared with model with Y_a and interpolated NDVI. This was done to see if model with NDVI work similar or better as compared to model with Canopy Cover.

If ground cover can be replaced by NDVI, then the reasons behind this significant relationship will be investigated. One of the reasons behind significant NDVI relationship which can be attributed to is the number of missing plants. The number of missing plants is often created by some form of stress. The number of missing plants can be related to the ground cover as the greenness decreases. The number of missing plants is a sum of the missing plants per plot and the number of plants which were still there but had a low chance of surviving.

After differentiating the datapoints on the field category basis, the datapoints were also differentiated cluster wise and year wise to understand if the fields are distinguishable in those regard. We combine SWAP-WOFOST data and ancillary data into a regression model to check how much of the variability in Y_a can be explained by them.

Lastly, we try to merge all the data to find how much of the Y_a can be accurately predicted by the combination of the three types of data (crop model outputs, remotely sensed data, and ancillary data). In lot of the graphs in the result section, non-linear regression line was used instead of linear regression line to show the potential of non-linear models to estimate actual yield. There are two types of regression lines found in the graphs. The geom_smooth() function in R was used for the lines of the graphs. The 'lm' (linear model) method was used for the straight lines and the 'loess' (Locally Estimated Scatterplot Smoothing) method was used the polynomial lines.

As the main focus of the study is to improve the estimation of the reducing category, Random Forest was performed only on the reducing_factor category. The independent variables chosen for this model were Yw, elevation, clay_est, ndvi_interpolated and irri_quantity. The two hyperparameters-mtry and ntree were optimised for the model. The hyperparameter ntree refers to the number of trees in the model. There should be enough trees to stabilise the error but not more than required. The hyperparameter mtry is how many variables will be included in the first split.r.m.s.e. was used for comparision in accuracy with the other models. Firstly, the r.m.s.e., values were calculated for different ntree values keeping mtry constant and then for the ntree deriving the lowest value of r.m.s.e., mtry was manually examined (Figure 17). Feature importance was also studied to identify the variables selected by the model for better prediction. 'IncNodePurity' is the short form for Increase in Node Purity. This calculates how much inclusion of a predictor variable improves the model's ability to predict. Higher ranked variables are more important features. The entire study was done on R-software version 4.3.1.

2.3.2 Description of the models

Model 1 (SWAP-WOFOST): The SWAP-WOFOST Y_p and Y_w are regressed with the actual yield field data using simple linear regression. Clusters, variable known as cluster_name, were used a factor.

Model 2 (SWAP-WOFOST and NDVI): The SWAP-WOFOST and the NDVI values are used as explanatory variables to see if there is any improvement in the accuracy with the inclusion of the NDVI values. Different variables have been added to the model 1 and these were referred to as sub-models. Multiple linear regression was used to estimate the Y_a using SWAP-WOFOST simulated outputs, Canopy Cover and NDVI. Firstly, Canopy Cover was used for the adding information to the model 1. This was followed by using NDVI as a replacement for Canopy Cover as the relationship between Canopy Cover and NDVI is strong from earlier results (Figure 12). NDVI alone performed better than Canopy Cover and hence can be used as a replacement for Canopy Cover as an explanatory variable. Then, both Canopy Cover and NDVI was used for improving the estimations. Avg_cnp_cvr refers to the Canopy Cover variable while ndvi_interpolated refers to the interpolated NDVI values.

Model 3 (SWAP-WOFOST and ancillary data): A model is developed using different factors that might have a significant relationship with actual yield (Y_a) like elevation, percentage of clay and quantity of irrigation. Ancillary data consisted of clay percentage, total irrigation amount and elevation data. Although clay percentage and irrigation amount are inputs to SWAP-WOFOST model, these variables were included in the model improvement. This is because it is assumed that these variables can add information to the model in such a way that SWAP-WOFOST fails to acknowledge. Linear regression models to estimate Y_a using SWAP-WOFOST and ancillary data were developed for the three field situation categories. Irri_quantity refers to the total amount of irrigation applied to the field while clay_est refers to the clay percentage.

Model 4 (SWAP-WOFOST, NDVI and ancillary data): Data from Model 2 and 3 (SWAP-WOFOST, NDVI, ancillary data) are then combined together into Model 4. It is expected that the model 4 will perform the best among all the models. The inclusion of remote sensing and crop models to the ancillary data in model 3 will check if the hypothesis 2 stands correct. Random forest are implemented for the model 4 with the idea to further improve the accuracy of the predicted yields.

2.3.3 Evaluation of the models

There are many ways for model performance measures for comparing model predictions and field observations. The models are analysed first graphically and then statistically. The quantitative measures used to meaure model performance in this study are R^2 and r.m.s.e. R^2 will explain how well the X variable explains Y variable (Wang and Jain, 2003) while r.m.s.e. stands for the square root of the sum of the squared differences between the predicted and observed values divided by the number of observations. The larger the R^2 value, the better the model while the smaller the r.m.s.e. value, the better the model. However, it must be dealt with caution as overfitting curves often have r.m.s.e. smaller than measurement and sampling error. The four main models had a critical graphical and quantitative analysis of the model performance. Adjusted R^2 and root mean square error (r.m.s.e.) has been used for quantitative analysis of the model performance. There are other models where only the graphical analysis of the model is discussed. A p-value of less than 0.05 makes all the models statistically significant.

2.3.4 Cross validation approach in Random Forest model

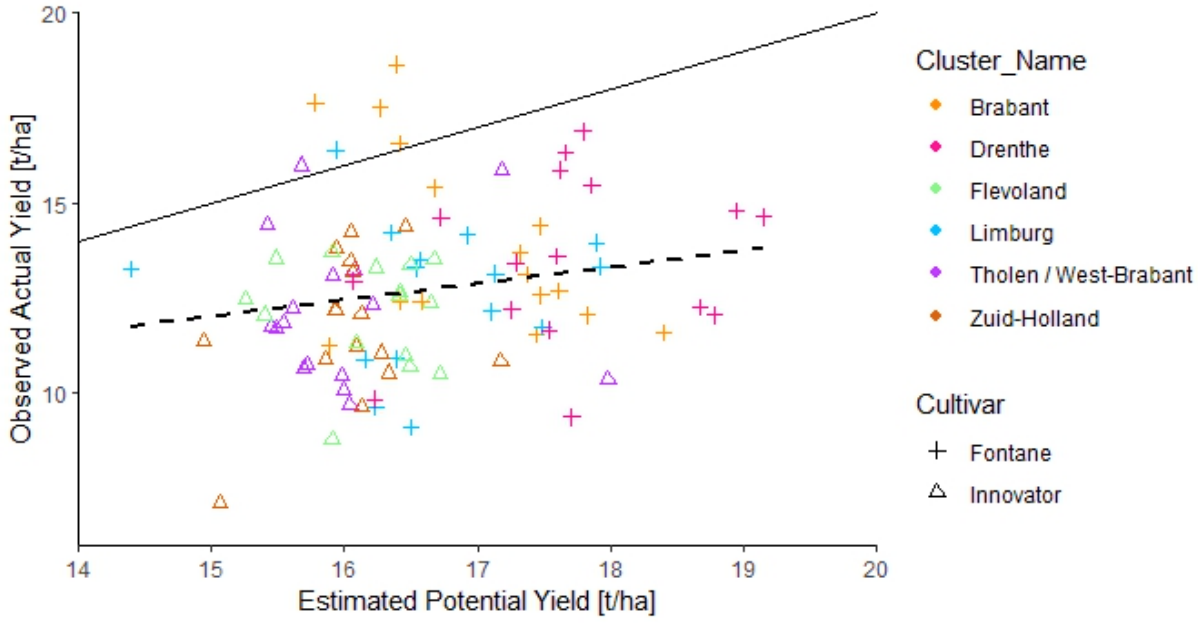
Linear regression is performed for all the four models against Y_a data in R software. Random forest was performed only in the last model to see if there can be any further improvement in the model using advanced regression technique. For this model, cross validation was performed to test accuracy and prediction of the Y_a. As the n<1000 for our dataset (n=94 fields), every observation is valuable to be seperated into training and testing data. Hence, the approach of leave one out cross validation becomes useful for the dataset. The leave out approach involves the use of one specific entity as the validation dataset while the rest of the observations are used as training dataset. The leave year out approach, the leave cluster out approach and leave field out approach are the three options available for our dataset. Leave year out approach tests the temporal generalisation of the model and accuracy and prediction in a new year while leave site out approach tests how accurate a model is when applied on a new cluster in the same year. As, we have two years of data in our dataset, there can be a huge variability in our model owing to variability in the weather conditions. On the other hand, leaving a cluster out results in huge variability owing to differences in soil conditions and management practices. In this study, the interest was to predict each field instead of year and cluster. Hence, the leave field out cross validation is used for this study. This ensures that every field is tested on the performance of the rest of the fields. In this way, the variability is much homogeneous. In this study, this LOOCV was run for all fields separately and then average r.m.s.e. was used for analysis.

3. Results

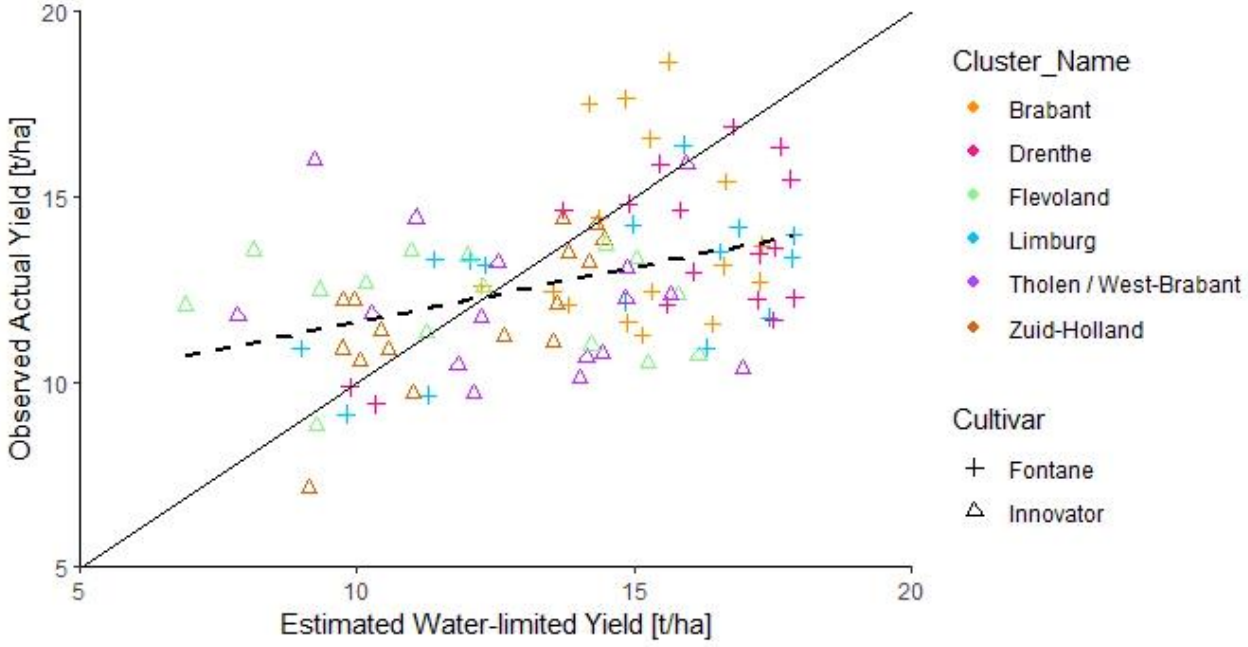
3.1 Actual yield estimation using SWAP-WOFOST model only

3.1.1 Simple Linear regression

The simple linear regression models between the actual field observed yields and SWAP-WOFOST yield estimates are represented in Figure 8 and Appendix 1 and 2. The actual yields from field data collected are represented by Observed Actual Yield (Y_a) while the SWAP-WOFOST model potential and water limited output yields are represented by Estimated Potential Yield (Y_p) and Estimated Water-limited Yield (Y_w) respectively.



(i)

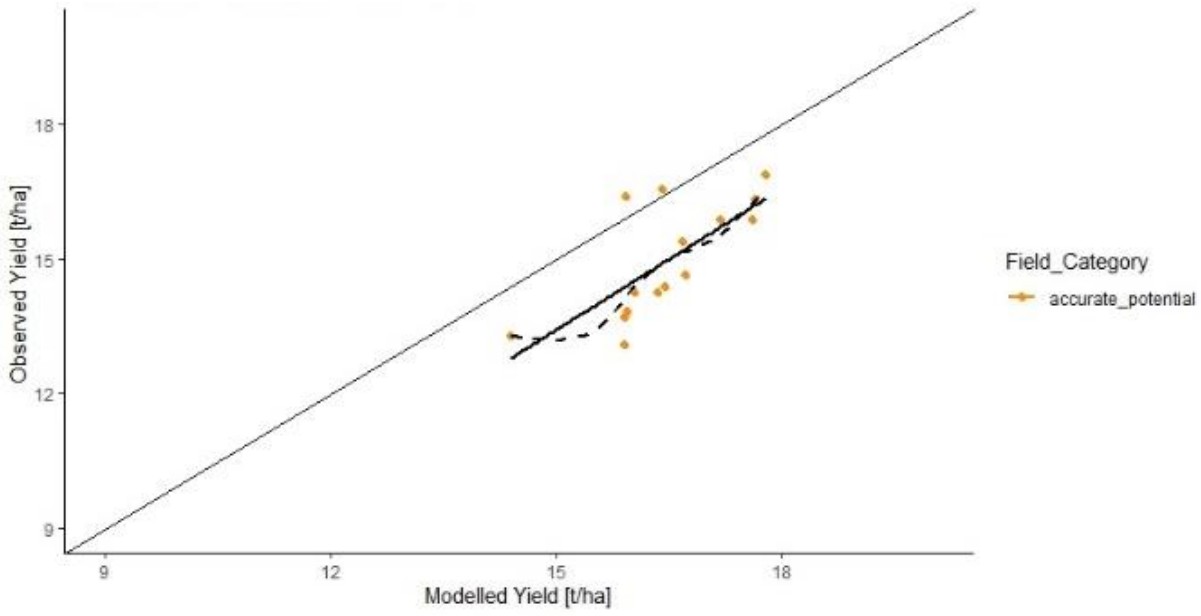


(ii)

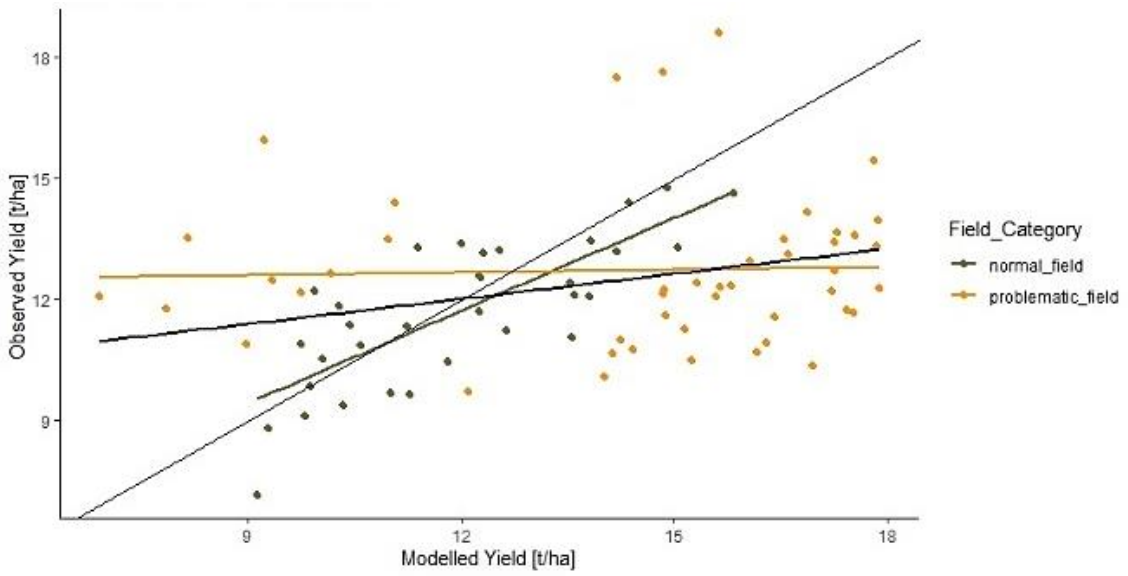
Figure 8: Relationship between (i) the Yp represents the SWAP-WOFOST modelled potential yield and (ii) the Yw represents the SWAP-WOFOST modelled water limited yield on the X-axis and Observed Actual yields on the Y axis. The solid and the dotted black lines represent the 1:1 line and the linear regression line respectively.

Potential yield values are much higher than actual yield values in Figure 8 (i) The spread of the datapoints was more around the 1:1 line in Figure 8(ii). Thus, water-limited yields are a better representation of actual yields as compared to potential yields. Fontane variety has higher overestimation as compared to Innovator in general In Figure 8 (i). Fontane also has a wider spread of datapoints while Innovator is more clustered. This can be due to the soil characteristics which is probably not well simulated by SWAP-WOFOST potential yields. Five out of six points which are on or above the 1:1 line belongs to the Fontane category. Drenthe, Limburg and Zuid-Holland are highly significant and their counter interactions with Y_p are also highly significant in Appendix 1. No significant relationships are found in Appendix 2. However, the median of the model residuals for Appendix 2 (0.02 t/ha) is much lower than Appendix 1 (0.17 t/ha). Although the overall adjusted R^2 improves in the water limited yield as compared to the potential yield, the R^2 value is only 0.21 proving there is a lot of space for improvement.

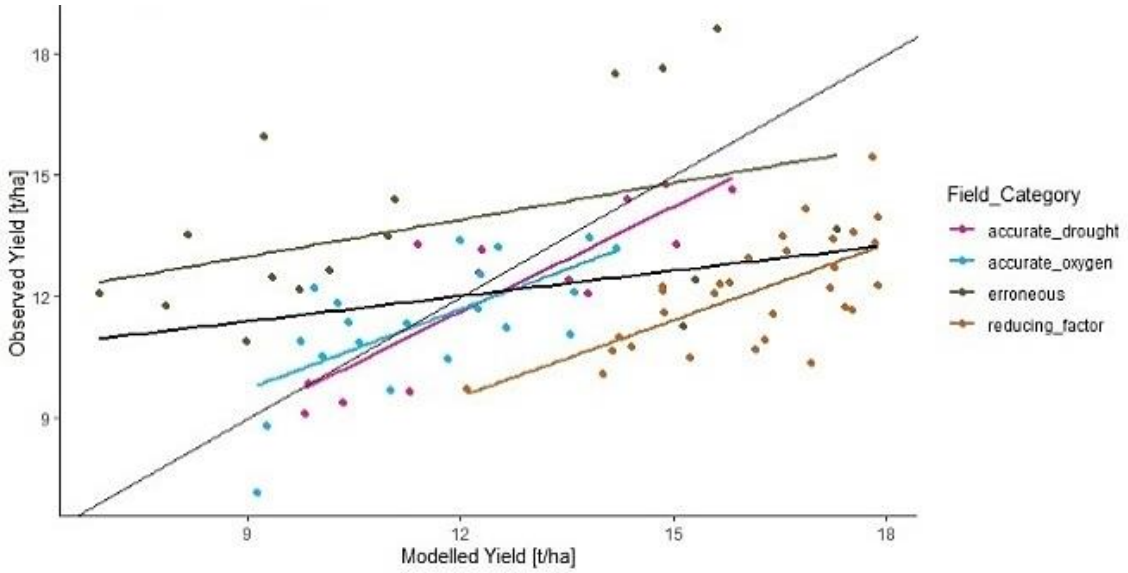
In Figure 9, the dataset was divided into normal and problematic fields as per the field category As Yp can be a better representation for accurate_potential category, Figure 9 (i) represents relationship of the two. However, for the rest of the categories, the actual yield values were compared with the water limited modelled yield values hence, Figure 9 (ii) and (iii) are represent the other categories. In Figure 9 (ii), the categories are combined into the normal and the problematic fields.



(i)



(ii)



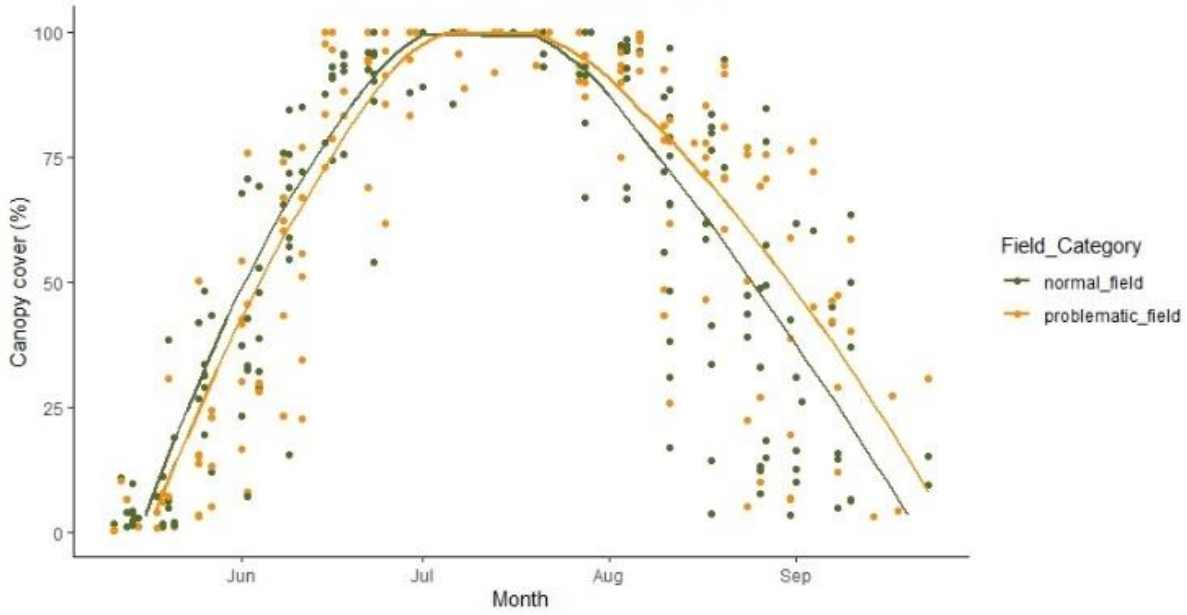
(iii)

Figure 9: Relationship between (i) Yp representing the SWAP-WOFOST modelled potential yields for accurate_potential category and (ii) Yw representing the SWAP-WOFOST modelled water limited yields for two problem categories and (iii) Yw representing the SWAP-WOFOST modelled water limited yields for five problem categories are on the X-axis and Observed Actual yields on the Y-axis respectively. The dotted black line represents the non-linear relationship in Figure 9 (i). The diagonal black line of the graph represents the 1:1 line while the other black line in the graph represents the regression line for all the values in the graph.

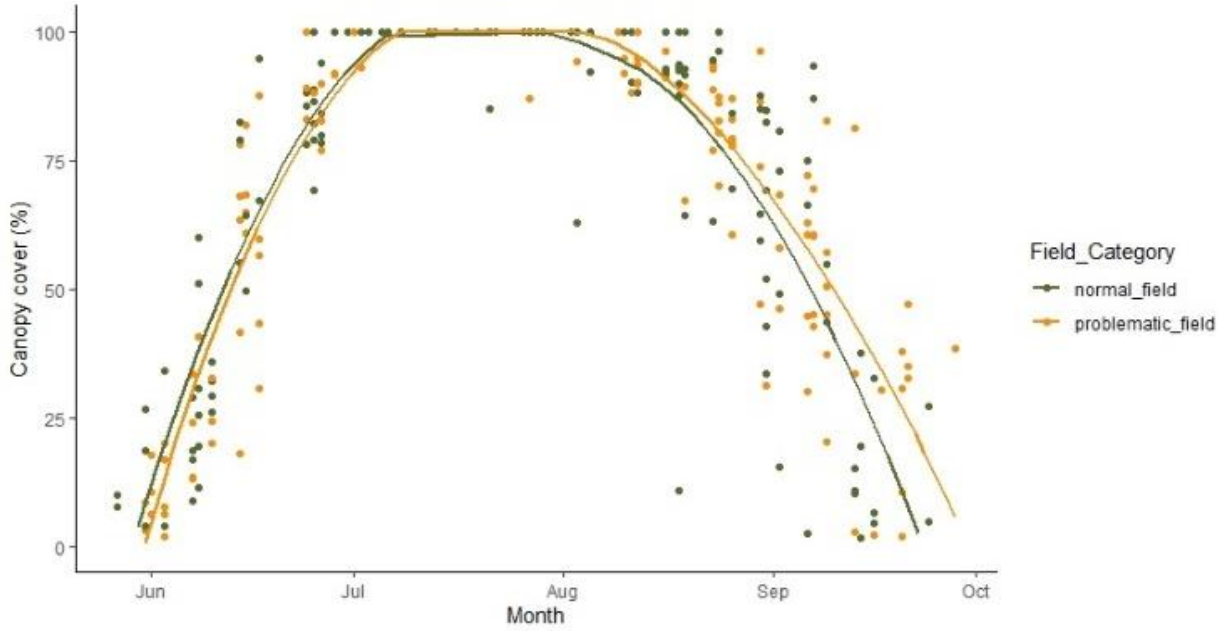
Figure 9 (i) shows Modelled (Y_p) Vs Observed yield and has only one category 'accurate potential', where the model overestimates the yield. Figure 9 (ii) shows that the normal fields are randomly scattered around the 1:1 line while the problematic fields are found scattered further away from the 1:1 line. When Figure 9 (ii) is compared with the Figure 9(iii), it can be seen that the underestimated yield values of the problematic category of Figure 9(ii) belong to the 'erroneous' category while the overestimated yield values of the problematic category of Figure 9(ii) belong to the 'reducing_factor' category.

3.1.2 Variation of Canopy Cover (%) over time

Canopy Cover (%) evolution is presented Figure 10 and 11. This is Canopy Cover captured on the field over two years. Similar evolution in the Canopy Cover can be observed in both the years. In both the years, the maximum vegetative stage is reached in July-August. It is difficult to differentiate between the two field categories at any phenological stage.

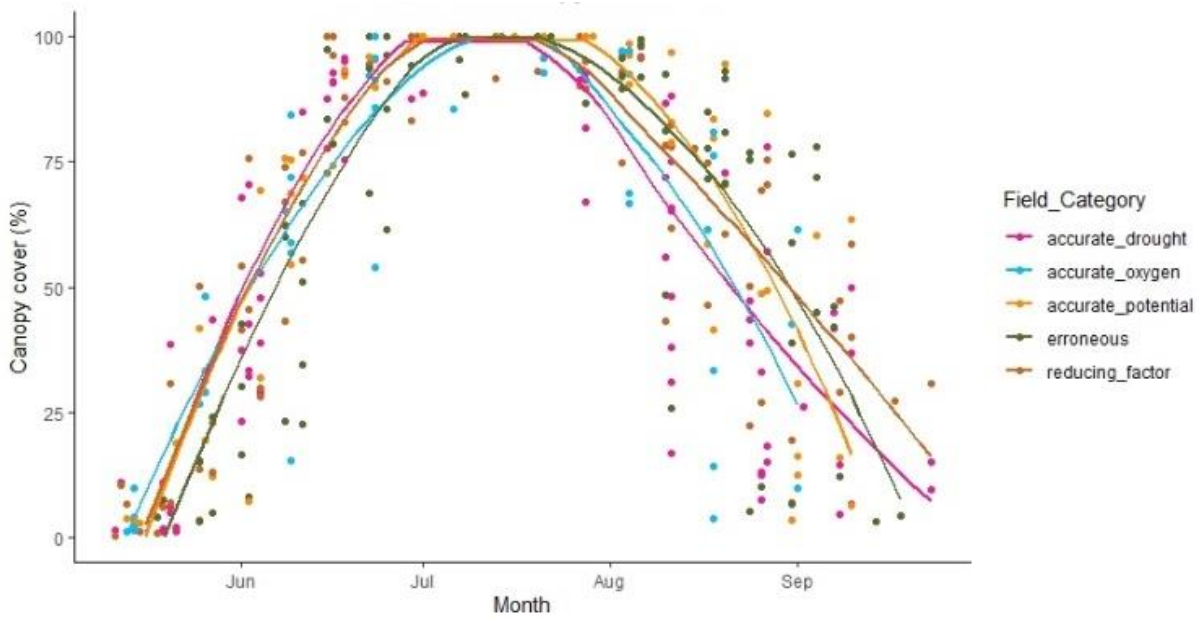


(i)

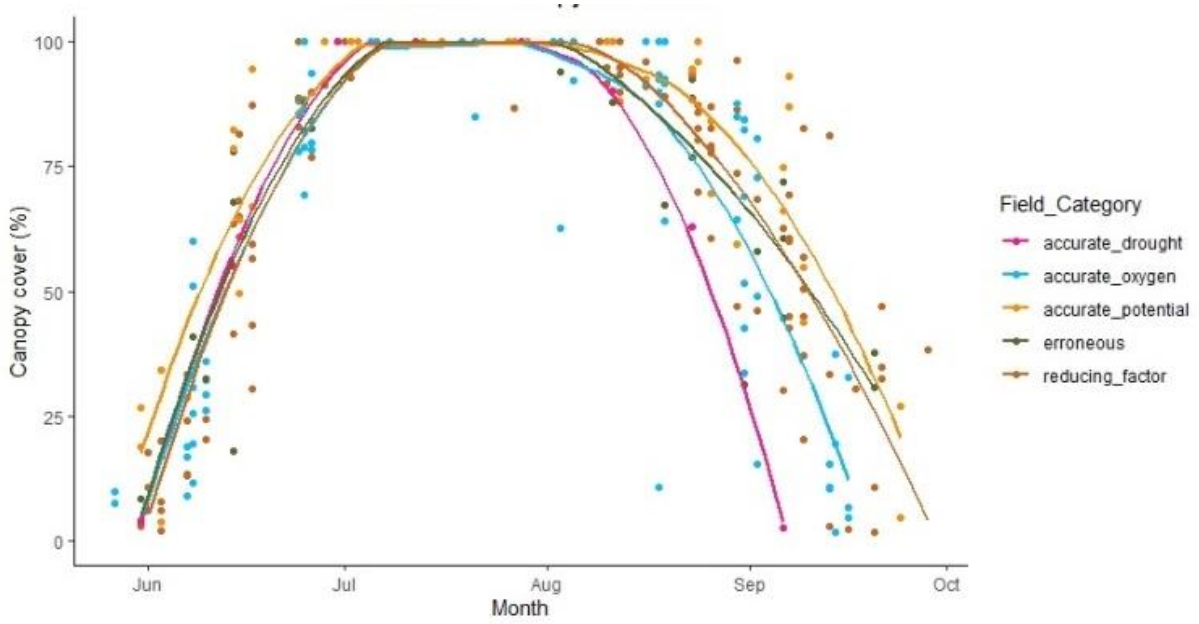


(ii)

Figure 10: Time series of field measured Canopy Cover for (a) 2020 and (b) 2021 respectively for normal and problematic fields. The normal fields include fields of the accurate_potential, accurate_oxygen and accurate_drought categories and the problematic fields include the fields of the the reducing_factor and erroneous categories.



(i)



(ii)

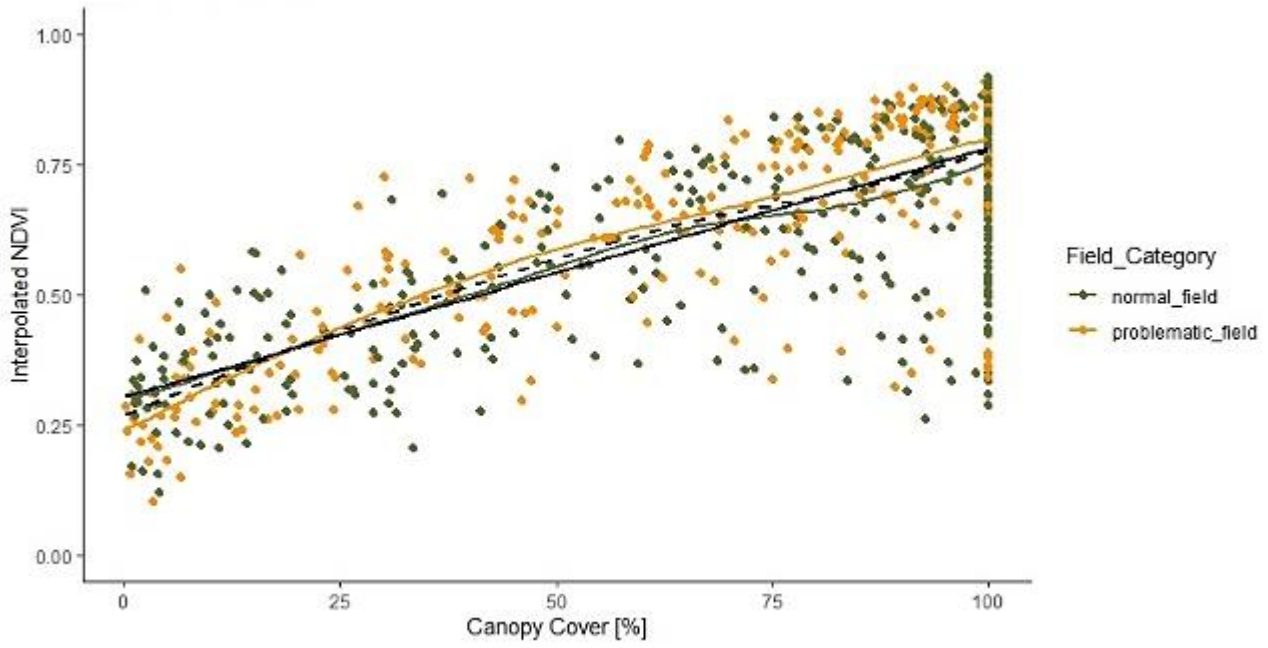
Figure 11: Time series of field measured Canopy Cover for (a) 2020 and (b) 2021 respectively for the five categories of the field problems.

More variation in the datapoints of different categories is observed after the maximum vegetation stage in both Figures 10 and 11. This shows that the differences in the categories are observed from the grain filling stage till the harvest stage when observed with Canopy Cover. The two categories accurate_drought and accurate_oxygen in Figure 11 (ii) have a reduction in canopy cover faster than the other categories in Figure 11. Reducing_factor category extends beyond accurate_potential category in 2020 by a lot as compared to 2021.

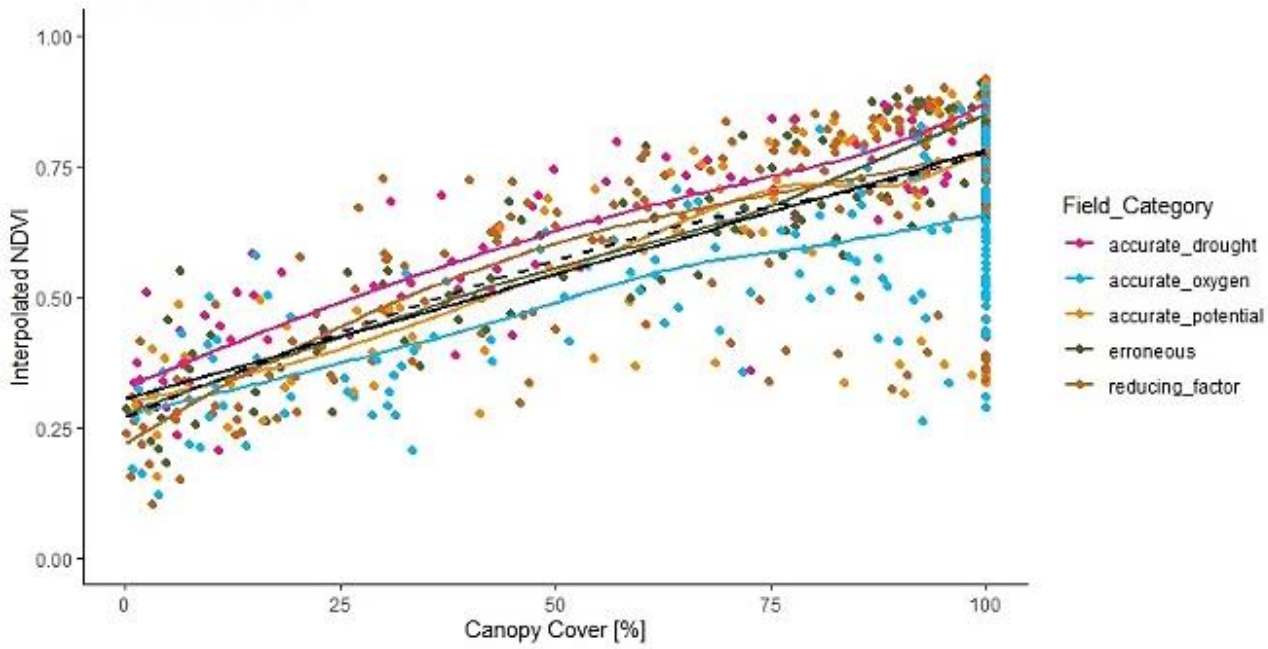
3.1.3 Relationship between interpolated NDVI and Canopy Cover

Figure 12 and Appendix 3 provide an overview of the relationship between interpolated NDVI and Canopy Cover. There is a general upward trend which proves that with increasing Canopy Cover, NDVI also increases. However,

towards the right-bottom part of the graphs, it is seen that a lot of datapoints which have been captured as almost full Canopy Cover do not have very NDVI (NDVI <0.5).



(i)



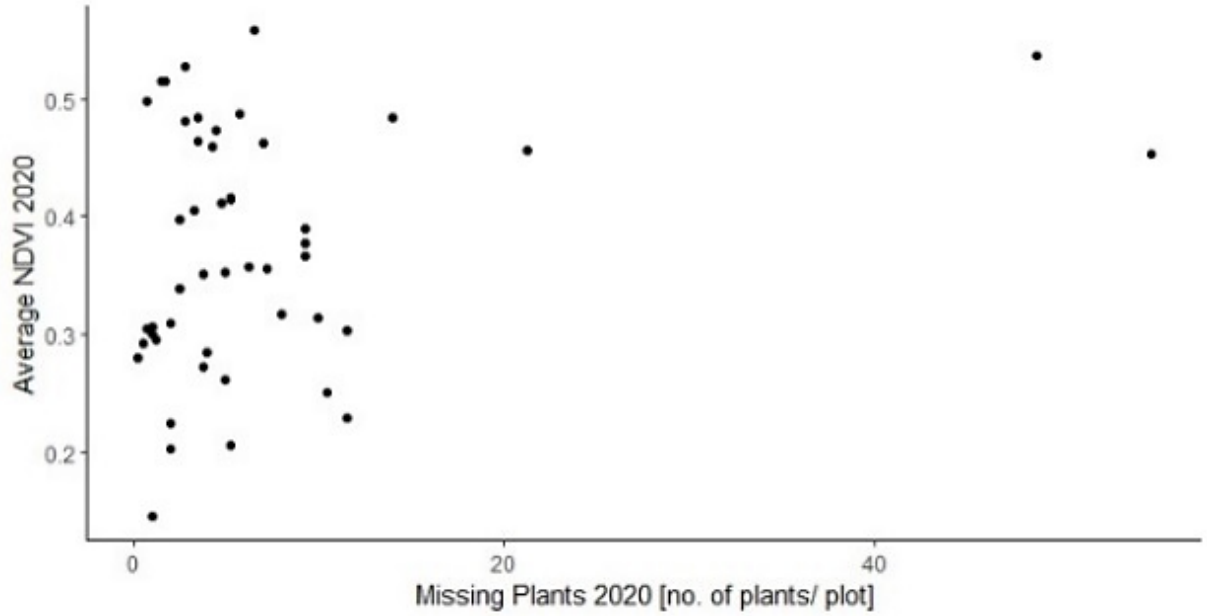
(ii)

Figure 12: The relationship between interpolated ndvi and field observed Canopy Cover for (i)normal and problematic fields and (ii) the five different categories.

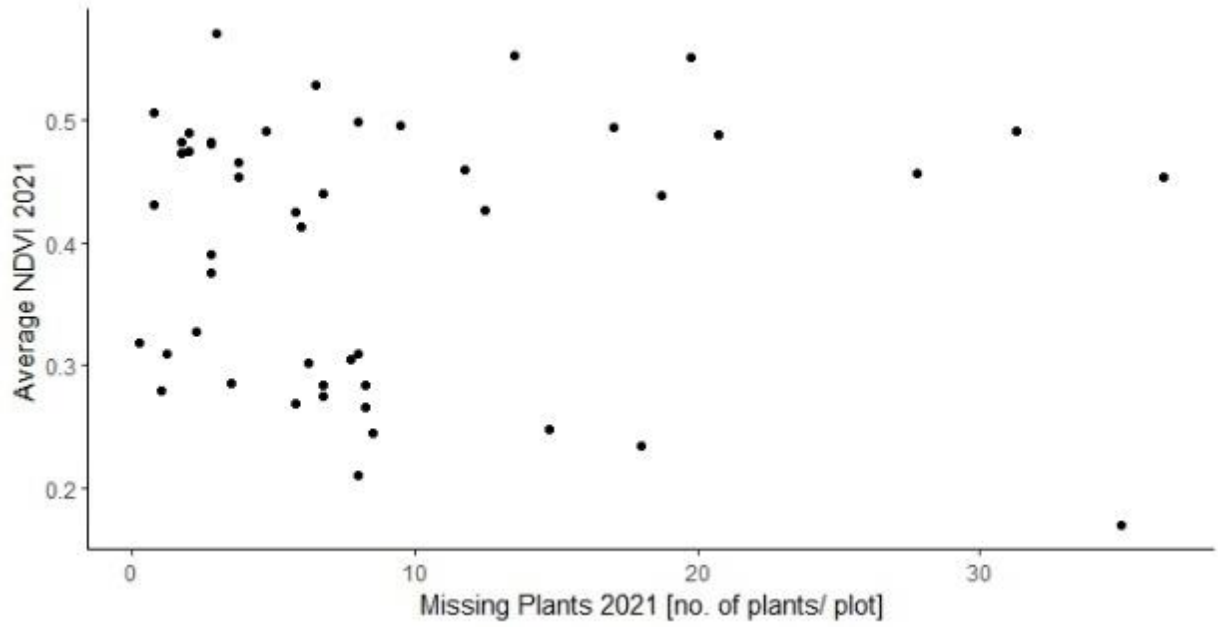
The ground Canopy Cover percentage is related to the NDVI to see if NDVI can be a replacement for it. Hence, the NDVI was studied with respect to each field problem category for both the years. In Appendix 3, a clear distinction is seen between the two varieties. Datapoints with Fontane variety is found mostly towards the top part of the graphs as compared to datapoints with Innovator variety , which is found in the lower part of the graphs. This can be related to the algorithm for the differences in canopy architecture of the different varieties and soil characteristics of the respective cluster.

3.1.4 Reasons for reduction of yield

The reduction of yield studied were missing number of plants and crop health. The relationship between the NDVI and the reason was studied for all fields together. For Figure 13 (i), the points are clustered on the left side of the graph as the number of missing plants hovered between zero and 20. However, there was random scattering observed in NDVI with increasing number of missing plants. Same number of missing plants can have different NDVI values. In Figures 14 (i) and (ii), the fields with crop health in ascending order shows a constant in (i) or decrease in (ii) in NDVI values. It is possible to say that the minute difference in crop health across fields could not be captured by NDVI. Also, the crop health was qualitative and was graded on visual introspection.

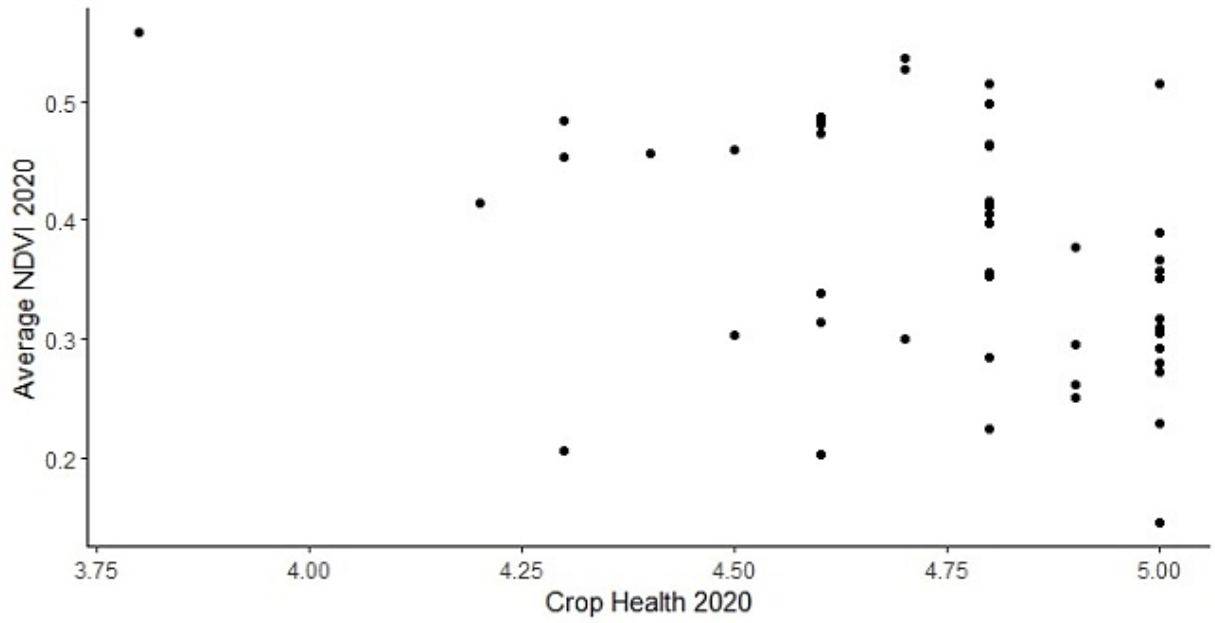


(i)

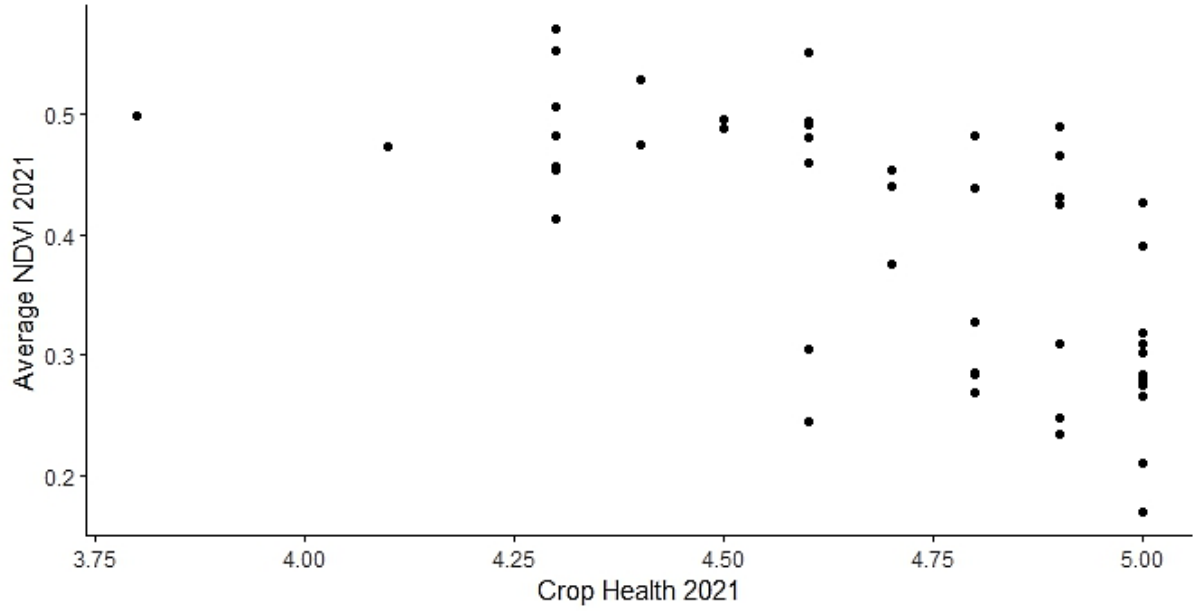


(ii)

Figure 13: Relationship between avg NDVI and number of missing plants for all fields in (i) 2020 and (ii) 2021.



(i)



(ii)

Figure 14: Relationship between NDVI and crop health for all fields in (i) 2020 and (ii) 2021.

The correlation coefficients in missing plants reason for all the fields category and the reducing factory category fields were 0.15 and 0.11 respectively. The correlation coefficients in crop health reason for all the fields and the reducing factory category fields were -0.54 and -0.48 respectively.

3.2 Actual yield estimation using WOFOST model output, average Canopy Cover and NDVI

Multiple linear regression models were developed for including the information of NDVI and canopy cover. The results of the combined Canopy Cover and NDVI (column 3 of each table) were slightly improved in Tables 7-9 as compared to their only NDVI counterpart (column 4 of each table). The best R^2 and r.m.s.e. values were found when the clusters were used as factors. This can be because clusters have different soil and weather conditions. NDVI_interpolated improves the R^2 for accurate_potential, reducing_factor and all the categories combined together categories as compared to avg_cnp_cvr. In accurate_potential and reducing_factor categories, the highest adjusted R^2 is for the submodel- Y_p + ndvi_interpolated*cluster_name. All the models have p-value very small, making them highly significant. However for all_fields category, there are two best performing submodels- (a) Yw + ndvi_interpolated*cluster_name and (b) Yw + ndvi_interpolated + avg_cnp_cvr *cluster_name in terms of adjusted R^2. Thus, it is possible to say that ndvi_interpolated can enough to improve the model with cluster_name as factor.

Table 7: Comparision of the quantitative analysis of different models for estimating Ya for model 1 (using SWAP-WOFOST) and model 2 (SWAP-WOFOST+NDVI) for accurate_potential category.

x=independent variables	Yp	Yp +avg_cnp_cvr	Yp + ndvi_interpolated	Yp +avg_cnp_cvr + ndvi_interpolated	Yp+ndvi_interpolated*cluster_name	Yp+ndvi_interpolated+avg_cnp_cvr*cluster_name
multiple R^2	0.54	0.55	0.55	0.61	0.75	**0.77**
adj R^2	0.54	0.54	0.55	0.61	**0.75**	0.74
r.m.s.e.	0.82	0.82	0.80	0.75	0.60	**0.59**
p-value	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16

Table 8: Comparision of the quantitative analysis of different models for estimating Ya for model 1 (using SWAP-WOFOST) and model 2 (SWAP-WOFOST+NDVI) for reducing_factor category.

x=independent variables	Yw	Yw +avg_cnp_cvr	Yw + ndvi_interpolated	Yw +avg_cnp_cvr + ndvi_interpolated	Yw + ndvi_interpolated * cluster_name	Yw + ndvi_interpolated + avg_cnp_cvr * cluster_name
multiple R^2	0.45	0.45	0.46	0.50	**0.55**	0.54
adj R^2	0.45	0.44	0.46	0.50	**0.54**	0.52
r.m.s.e.	0.98	0.97	0.97	0.92	0.89	**0.89**
p-value	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16

Table 9: Comparision of the quantitative analysis of different models for estimating Ya for model 1 (using SWAP-WOFOST) and model 2 (SWAP-WOFOST+NDVI) for all fields.

x=independent variables	Yw	Yw +avg_cnp_cvr	Yw + ndvi_interpolated	Yw +avg_cnp_cvr + ndvi_interpolated	Yw + ndvi_interpolated * cluster_name	Yw + ndvi_interpolated + avg_cnp_cvr * cluster_name
multiple R^2	0.15	0.16	0.16	0.18	0.20	**0.22**
adj R^2	0.15	0.15	0.16	0.18	0.20	**0.20**
r.m.s.e.	1.92	1.89	1.92	1.87	1.87	**1.83**
p-value	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16

3.3 Actual yield estimation using SWAP-WOFOST model output and other secondary data

Secondary data refers to clay percentage, elevation and total iriigation amount data. Irri_quantity variable reduced the R^2 value of the SWAP-WOFOST outputs for both the accurate_potential and the reducing_factor category. This means that unnecessary addition of variable might not always add useful information to the model but can add more confusion to the model. However, when added to the model with all other ancillary variables and the cluster factor, this improved the R^2 for all the three categories. This can be due to the differences in the irrigation requirement of the clusters.

It is interesting to observe that SWAP-WOFOST can estimate yields with high accuracy with the aid of ancillary data for accurate_potential category (Table 10). There is overfitting in the last submodel Yp+ elevation+ clay_est+ irri_quantity* cluster_name. When irri_quanity is removed from the overfitted model, Yp+elevation+clay_est*cluster_name performs the best R^2 and r.m.s.e. value in the accurate potential category. For the reducing_factor and all_fields categories, the submodel Yp+elevation+clay_est+irri_quantity*cluster_name performed best in terms of adj R^2 and r.m.s.e.

Table 10: Comparision of the quantitative analysis of different models for estimating Ya for model 3 (using SWAP-WOFOST+ancillary data) for accurate_potential category.

x=independent variables	Yp	Yp+elevation	Yp+clay_est	Yp+irri_quantity	Yp+elevation+clay_est	Yp+elevation+clay_est+irri_quantity	Yp+elevation+clay_est*cluster_name	Yp+elevation+clay_est+irri_quantity*cluster_name
multiple R^2	0.54	0.74	0.64	0.34	0.75	0.68	0.79	**1**
adj R^2	0.54	0.74	0.64	0.34	0.74	0.68	0.78	**1**
r.m.s.e.	0.82	0.62	0.73	0.91	0.61	0.63	0.56	**1.04E-12**
p-value	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16

Table 11: Comparision of the quantitative analysis of different models for estimating Ya for model 3 (using SWAP-WOFOST+ancillary data) for reducing_factor category.

x=independent variables	Yw	Yw +elevation	Yw +clay_est	Yw +irri_quantity	Yw +elevation+clay_est	Yw+elevation+clay_est+irri_quantity	Yw+elevation+clay_est*cluster_name	Yw+elevation+clay_est+irri_quantity*cluster_name
multiple R^2	0.45	0.49	0.50	0.40	0.50	0.48	0.62	**0.74**
adj R^2	0.45	0.48	0.50	0.40	0.50	0.48	0.61	**0.74**
r.m.s.e.	0.98	0.95	0.94	0.94	0.93	0.87	0.82	**0.61**
p-value	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16

Table 12: Comparision of the quantitative analysis of different models for estimating Ya for model 3 (using SWAP-WOFOST+ancillary data) for all fields.

x=independent variables	Yw	Yw +elevation	Yw +clay_est	Yw +irri_quantity	Yw +elevation+clay_est	Yw+elevation+clay_est+irri_quantity	Yw+elevation+clay_est*cluster_name	Yw+elevation+clay_est+irri_quantity*cluster_name
multiple R^2	0.15	0.15	0.15	0.13	0.15	0.14	0.25	**0.30**
adj R^2	0.15	0.15	0.15	0.13	0.15	0.14	0.25	**0.30**
r.m.s.e.	1.92	1.91	1.91	1.98	1.91	1.98	1.80	**1.78**
p-value	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16

3.4 Actual yield estimation using WOFOST model output, NDVI and other secondary data

3.4.1 Multiple Linear regression

Model 4 deals with the improvement in the model performance by incorporating all the variables studied in this scientific work. As already mentioned earlier, SWAP-WOFOST can already be correctly predicted with ancillary data alone for accurate_potential category (Table 10). The final model of Table 10 was already overfitted. Thus the addition of NDVI is not required and has been done solely to maintain homogeneity in the scientific work. In all the categories studied, it is found that addition of all variables with the factor cluster helped to establish the highest R^2 values with the lowest r.m.s.e. values of this study.

Table 13: Comparision of the quantitative analysis of different models for estimating Ya for model 4 (using SWAP-WOFOST+ancillary data+NDVI)) for accurate_potential category.

x=independent variables	Yp+elevation+clay_est+ndvi_interpolated	Yp+elevation+clay_est+ndvi_interpolated*cluster_name	Yp + elevation + clay_est + ndvi_interpolated + irri_quantity	Yp+elevation+clay_est+ndvi_interpolated+irri_quantity*cluster_name
multiple R^2	0.74	0.76	0.67	**1**
adj R^2	0.74	0.75	0.67	**1**
r.m.s.e.	0.61	0.59	0.64	**2.65E-13**
p-value	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16

Table 14: Comparision of the quantitative analysis of different models for estimating Ya for model 4 (using SWAP-WOFOST+ancillary data+NDVI) for reducing_factor category.

x=independent variables	Yw+elevation+clay_est+ndvi_interpolated	Yw+elevation+clay_est+ndvi_interpolated*cluster_name	Yw + elevation + clay_est + ndvi_interpolated + irri_quantity	Yw+elevation+clay_est+ndvi_interpolated+irri_quantity* cluster_name
multiple R^2	0.51	0.56	0.49	**0.74**
adj R^2	0.50	0.55	0.49	**0.74**
r.m.s.e.	0.93	0.88	0.86	**0.61**
p-value	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16

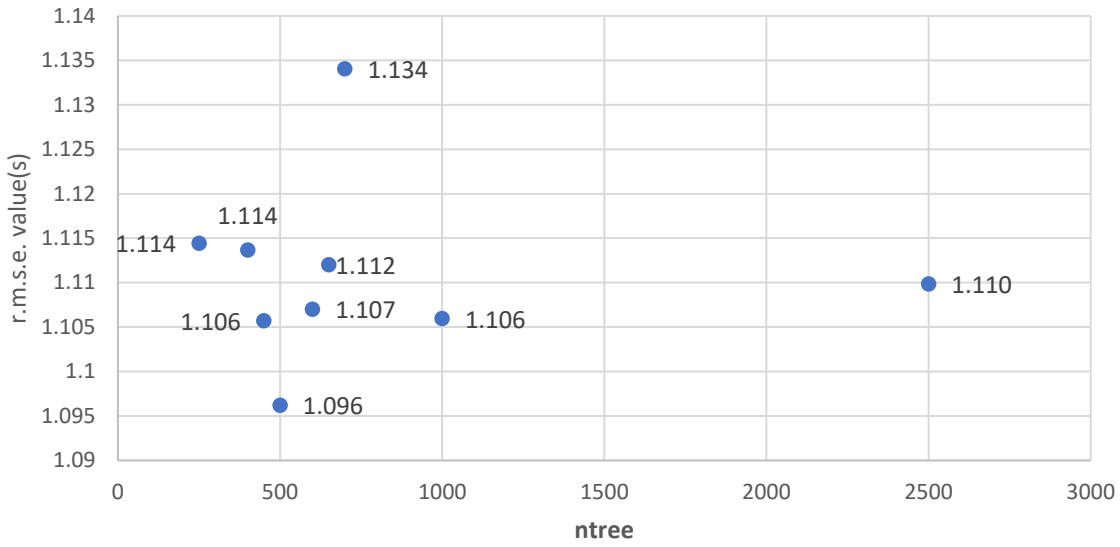
Table 15: Comparision of the quantitative analysis of different models for estimating Ya for model 4 (using SWAP-WOFOST+ancillary data+NDVI) for all fields.

x=independent variables	Yw+elevation+clay_est+ndvi_interpolated	Yw+elevation+clay_est+ndvi_interpolated*cluster_name	Yw + elevation + clay_est + ndvi_interpolated + irri_quantity	Yw+elevation+clay_est+ndvi_interpolated+irri_quantity* cluster_name
multiple R^2	0.16	0.22	0.14	**0.31**
adj R^2	0.16	0.22	0.14	**0.30**
r.m.s.e.	1.92	1.85	1.98	**1.78**
p-value	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16

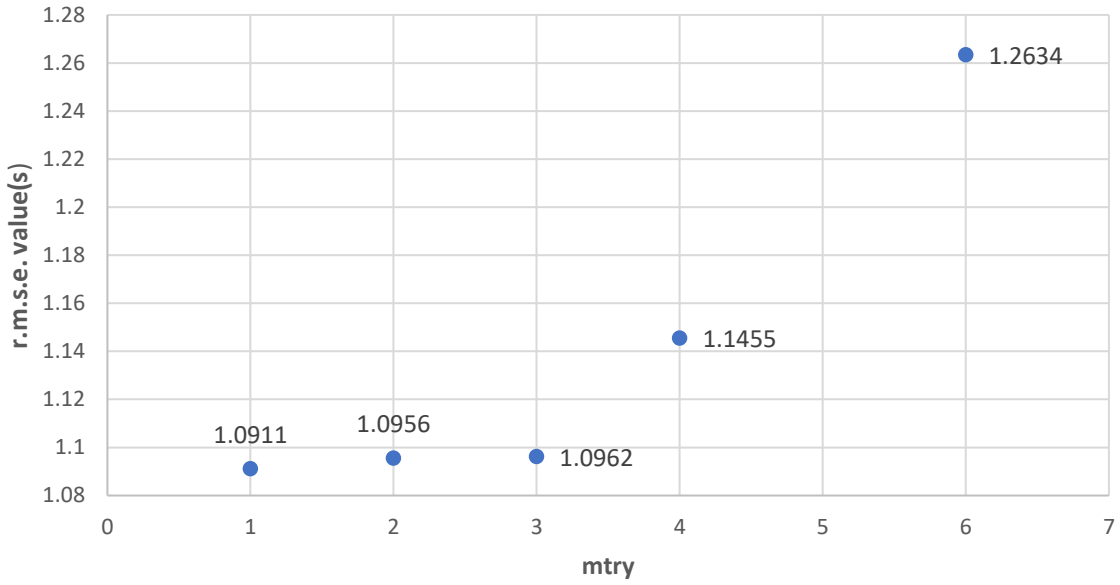
3.4.2 Random Forest Model

Linear regression for the reducing factor category estimated the Y_a with a r.m.s.e. of 0.59 and 0.86 with and without the cluster factor. The best performing model with the most appropriate hyperparameters (ntree=500, mtry=1) had a r.m.s.e. of 1.0911 (Figure 16), which is much higher than the r.m.s.e. of the linear regression model for the same variable combination of 0.86. Figure 15 shows the importance of the feature selection. Although there are subtle differences in the absolute values of the feature importance in each field, overall Y_w is the most important feature chosen by the random forest model.

Figure 15: Feature importance for all the fields in Random Forest model.



(i)



(ii)

Figure 16: Variation in r.m.s.e values with (i) ntree hyperparamter with mtry=3 and (ii) mtry hyperparamter with ntree=500

4. Discussion and Recommendation

4.1 Discussion

In this study, improvement of estimation of Y_a using more relevant data was investigated. In Table 3, the field observed values are closer to the values of Yw as compared to Yp. This can also be inferred from the Figure 8. Figure 8 represents a regression for all fields together. Most of the fields were overestimated in estimating Y_a using Y_p. Only six fields were above the 1:1 line in the Figure 8 (i). The Figure 8 (ii)-Y_w represented the Y_a way better than the Y_p as the spread of the datapoints are around the 1:1 line.

Figure 9 (ii) and (iii) shows that there is space for the model of the reducing factor and the erroneous categories to be improved. The regression lines of the accurate_drought and the accurate_oxygen categories are close to the 1:1 line. Thus on average, the model is able to estimate these two categories accurately. The residuals of these two categories are also small.

Phenology plays a role in the LAI curve. In the early phenological stages, the LAI is low. The LAI reaches a maximum with the maximum vegetative stage and then gradually decreases over time until the plant reaches senescence. This results in a bell-shaped curve over the entire growing season. This Phenomenon has been well captured by Canopy Cover (%) in Figure 10 and 11. Thus, it can be assumed that Canopy Cover can be used a proxy for LAI.

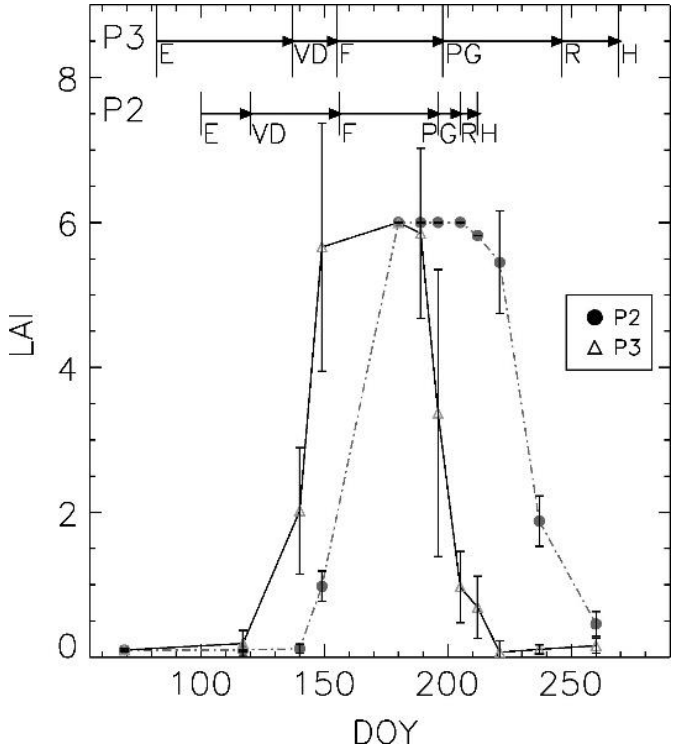


Figure 17: Retrieved LAI of two potato field, namely, P2 and P3. The phenological stages are indicated at the top as E indicating emergence, VD indicating Vegetation Development, F for Flowering, PG for Potato Growing, R for Ripening an H for Harvest (González-Sanpedro et al., 2008)

Figure 10 illustrates the initial growth of the canopy reaching a peak of 100% Canopy Coverage followed by the gradual decrease of the Canopy Cover. Similar inverted bell shaped curves for canopy cover evolution over time has been reported in (Tenreiro et al., 2021). In both the years (i) and (ii), the Canopy Cover was not able to differentiate a lot between the normal and the problematic categories. Removal of the erroneous category from the problematic category did not help to differentiate between the two categories in Figure 10. A deeper look at the two categories at the level of field problems in Figure 11 also could not differentiate among the categories.

Average NDVI might not be the best representation to correlate with the missing plants (Figure 13) or crop health (Figure 14). As a further study, it can be interesting to use a early date of NDVI instead of the average NDVI over the entire growing season to better represent the missing plants. This is because at early phenological stages, the plants do not expand and create 100% soil coverage. In the maximum vegetative stages, the plants expand and cover the entire soil even if there are many missing plants or not. It is not appreciated to use the average NDVI values as the average values will hover around a certain value instead of showing enough variation which is important for capturing ground information.

The NDVI values used were interpolated values and not real satellite data values. This can create some discrepancy in capturing the actual ground truth information on the specific date. There can be differences as there can be sudden changes on ground which might not be really captured by interpolated NDVI. In further studies, it might interesting to see if NDVI has a relation with yield gap values. Also, inclusion of time series for each variable like NDVI and irrigation schedule provides more information and can possibly improve yield estimation.

NDVI was related to ground cover to evaluate its relationship to replace tedious ground cover measurements by satellite data (Figure 12). The Pearson's correlation coefficient for the relationship for all the points together (without categorising) was 0.93. Thus, the first hypothesis is accepted that states satellite derived NDVI values are highly correlated with field measured ground cover values. There are many other studies to show that NDVI and Canopy Cover can be highly correlated. Figure 12 clearly indicates that the data for the normal and the problematic fields is difficult to be differentiated. In Figure 12 (ii), accurate_drought and accurate_oxygen can be differentiated as accurate_oxygen had a lower average NDVI as compared to accurate_drought for the same percentage of Canopy Cover.

As the reasons for actual yield was investigated to identify if reducing factor category can be represented by these factors, the correlation coefficients for all the fields and the reducing factory category fields were calculated. Thus, it is not possible to conclude that reduction in yield either by number of missing plants or crop health is identified by Canopy Cover or NDVI. As the NDVI and Canopy Cover are highly correlated, the graphs were not repeated for the two variables.

Inclusion of measured Canopy Cover improved the R^2 values of the model 1 (Table 7-9). In some cases, the R^2 was maximum for only adding NDVI values. In Table(s) 7-9, the R^2 always improved slightly when NDVI was used as a predictor variable as compared to Canopy Cover. Thus it is possible for NDVI to replace Canopy Cover for all fields. Although the idea was to see if NDVI can replace Canopy Cover, it was interesting to see that both the variables together can be better for estimating the Ya. This means that Canopy Cover can add such information which is not possible for NDVI to add. For example, disease in crop causing yellow leaves can be visible in Canopy Cover but not in satellite data. The results further improve when clusters are used as a factor in all three cases.

Lastly, it is possible to conclude that seperately estimating for each field problem category performs better than estimating for all fields together. Addition of Canopy Cover variable as the explanotory variable in the NDVI+ SWAP-WOFOST model added only meagre information to the model owing to the high correlation between NDVI and Canopy Cover. This indicates NDVI is equivalent to Canopy Cover. Thus, it is possible from Table 8 to say that the addition of remote sensing data to SWAP-WOFOST outputs can improve the Y_a estimation (research question 1). It helps to reduce the yield gap by bringing the Y_w close to Y_a.

All the three categories where Model 3 is examined (Tables 10-12), the inclusion of all the forms of ancillary data to the SWAP-WOFOST results performed as the best models than there counterparts. Accurate potential category had a R^2 of 1 with minimal r.m.s.e. A graphical interpretation can aid to ensure that the model is so perfect. There is overfitting of the model. However, for the other two categories, there was improvement in R^2 and r.m.s.e. values for including all the ancillary data. It is therefore possible to say that ancillary can definitely explain more variation in the data if added to Model 1 (research question 2). There can be two reasons why elevation was highly related to each cluster. Firstly, the elevation of clusters with clayey soils are mainly near the sea and the elevation of these is below the mean sea level as compared to clusters like Limburg which have higher elevation. Also, the groundwater table will vary for each of the clusters.

In Model 4, the three categories behaved differently. The all fields category improved slightly in R^2 while the r.m.s.e. also increased by 0.003 t/ha. The reducing category had a slight decrease of R^2 and a r.m.s.e. increase of

0.002 t/ha. The accurate potential category had a constant R^2 of 1 with a extremely minute decrease in r.m.s.e.. Hence, this shows that adding on information may not necessarily be improving the model and should be concious to what is being added (research question 3). Adding more information to an overfit model is also not the best choice. Hypothesis 2 can be accepted as the model 4 for all categories is almost the same as model 3. The slight differences in R^2 and rmse values is due to the calculations based on four decimal points. However, when R^2 and r.m.s.e. values are rounded to two decimal points, the results of model 4 and model 3 are exactly the same. Cluster and field problem categories when used as factor or separate models respectively improved the understanding of the models.

Reducing_factor category always performed better than all fields category and worse than accurate_potential category. Reducing_factor category fields had dramatic increase in the R^2 and decrease in r.m.s.e. values (Appendix 5-8). Model 3 and 4 had a adjusted R^2 increase of 36.35% and 35.41% and a r.m.s.e. decrease of 31.37% and 31.32% respectively as compared to Model 1. Model 2, on the other hand, had a slight decrease of adjusted R^2 of 4.27% and a slight r.m.s.e. decrease of 0.63% as compared to Model 1.

Random Forest of Model 4 performed worse than Model 1 in terms of r.m.s.e. values. Hence, non-linear models might not be the better option to estimate Y_a for this dataset (research question 4). The model which had the lowest r.m.s.e. was with ntree=500 and mtry=1. Mtry=1 makes trees less diverse due to availability of only one split. Random Forest also used a different approach than the rest (LOOCV method). Feature selection (Figure 15) shows that Yw was the most useful variable while NDVI was the least useful variable when the RF model was developed. This is in tune with the earlier results. Additon of NDVI values to Model 3 to develop Model 4 did not drastically improve the model efficiency. Along with that, cluster was not used as factor in the RF model although the comparision was made with the appropriate counterpart, i.e., linear regression models without cluster as a factor. Deep learning can also be used to estimate Y_a using SWAP-WOFOST results and other data. However, it should be noted that neural networks can function well with ample amount of data, hence care should be taken that the amount of input data for such models are enough.

4.2 Future recommendations that can be applied in further studies

Potato crop is susceptible to extreme dry or wet conditions. The crop has a weak rooting system which is unable to breakthrough the hard impermeable soil layers to reach available water resulting in poor yield. On the other hand, the oversaturated soil can lead to dying roots and rotting tubers. Hence, incorporation of soil moisture conditions using field collected data can provide additional information which can improve the yield estimation. Satellite data which can be used as proxy for soil water content-passive SAR data can also be used as a replacement for field data collected soil moisture content measurements. Inclusion of weather data drastically improved the models for early and late potato yield models explaining the variability present in the yields across northern Belgium (Vannoppen and Gobin, 2022).

Selection of the most appropriate period for estimation of yield is essential as the estimation performance can be highly vary during the different growth periods (Lin et al., 2023) for different varieties and environmental conditions. Average monthly values of Sentinel-2 vegetation indices of beginning of tuberization, early senescence under non-normal conditions such as pests and senescence were used in a study (Gómez et al., 2019). The average maximum LAI for potato crop were found during mid-season (DOY 180) with values varying around 6-8m^2/m^2 while the average minimum was found towards the beginning of the growing season (DOY 140) having values 2-3m^2/m^2 under semi-arid irrigated conditions (Mourad et al., 2020). Tuber expansion stage, which is 70 days after planting, turned out to be the best stage to estimate yield according to (Luo et al., 2020) while (Li et al., 2020) mentions how both the 90 DAP Random Forest (RF) and Partial Least Square (PLS) regression model outperform 60 DAP. Estimation of yield using multiple dates instead of single date improves adjusted R^2 values from 0.7415 to 0.8225 (Luo et al., 2020). Thus, selection of suitable dates can be useful for further improvement of this study.

Random Forest for Model 4 included a lot of information which resulted in slow execution of the process due to computational limitations. Inclusion of more data to the level where it can be called big data can be useful for yield estimation. Big data refers to a voluminous amount of data which are so complex in nature that is often requires newer technologies like artificial intelligence for processing (EU Parliament, 2021). (Silva et al., 2020) studied different region, cultivar, year, soil type, irrigation quantity, rainfall quantity, intercepted PAR, sowing and harvesting dates, available N, P, K applied and field size as drivers of reported Y_a for multiple crops, including

ware, starch, and seed potatoes, in the Netherlands using standard regression models. Many insights, thus, often remain hidden without the use of more advanced processing technologies like artificial intelligence. Thus, although (Silva et al., 2020) provides us an idea about the drivers of yield for potatoes in the Netherlands, involvement of machine learning technique to get more meaningful comprehension of the data can be an appropriate tool. Agronomic parameters like plant height, soil parameters like moisture, conductivity and nutritional parameters have also improved greatly yield estimation. Inclusion of cultivar information to remote sensing data can improve model R^2 values by 56% (Li et al., 2021).

At the same time, this study indicates that more data need not necessarily mean improved models. Thus, addition of data to the models leading to improvement while taking care of computational resources can make the modelling a sustainable choice.

5. Conclusion

This study gave an insight about how remote sensing and ancillary data can aid to SWAP-WOFOST results to estimate actual yields. No strong association was found with the number of missing plants and the crop health with NDVI. Canopy cover and NDVI together can improve the R^2 than alone while when cluster is added as a factor, NDVI alone was enough to have the highest R^2 and lowest r.m.s.e. values. In all fields combined category, there was no significant difference when Canopy Cover and NDVI were added as separate independent variables to the SWAP-WOFOST outputs. However, as NDVI is easier to derive than measuring Canopy Cover, thus, remote sensing can definitely explain better the variability of SWAP-WOFOST outputs. Ancillary data can also help to explain the variability of the SWAP-WOFOST outputs although for the accurate potential category, there is overfitting of the model. When both NDVI and ancillary data with SWAP-WOFOST outputs were combined, the R^2 values and rmse values were similar to the values of the previous SWAP-WOFOST and ancillary data submodels. Thus, combining NDVI with ancillary data does not explain the variability better than the SWAP-WOFOST and ancillary data model. In all the models, cluster as a factor improved the results drastically. Accurate potential category performed better than reducing category and followed by all-fields category. Unnecessarily inclusion of explanatory variables can lead to more confusion than clarity in estimating the yield.

Random Forest did not improve the linear regression model with SWAP-WOFOST, remote sensing and ancillary data combined. Lastly, it is possible to conclude that rational inclusion of satellite and ancillary data can certainly improve the estimation results for reducing field category. The first hypothesis is accepted that states satellite derived NDVI values are highly correlated with field measured ground cover values. The second hypothesis is accepted as well because the results of SWAP-WOFOST, remote sensing and ancillary data model were the same as SWAP-WOFOST and ancillary data model. Further research can improve estimation of actual yield using better techniques and relevant information.

6. Acknowledgements

I would like to thank Tom Schut and Lammert Kooistra for being supportive and kind supervisors. The data used in this study has been collected by Paul Ravensbergen, to whom I would like to thank for sharing his vast knowledge about the topic. Apart from that, people at my department, Plant Production Systems, provided a congenial environment where I could comfortably work on my thesis. Lastly, to all my friends and family, I thank you all for discussing all problems and celebrating my ups.

7. References

Al-Gaadi, K.A., Hassaballa, A.A., Tola, E., Kayad, A.G., Madugundu, R., Alblewi, B., Assiri, F., 2016. Prediction of Potato Crop Yield Using Precision Agriculture Techniques. Plus One. https://doi.org/10.1371/journal.pone.0162219

Beukema, Van Der Zaag, D.E., 1990. Introduction to potato production, Pudoc Wageningen.

Bouman, B.A.M., Van Keulen, H., Van Laar, H.H., Rabbinge, R., 1996. The "School of de Wit" crop growth simulation models: A pedigree and historical overview. Agricultural Systems 52, 171–198. https://doi.org/10.1016/0308-521X(96)00011-X

de Wit, A., Boogaard, H., Fumagalli, D., Janssen, S., Knapen, R., van Kraalingen, D., Supit, I., van der Wijngaart, R., van Diepen, K., 2019. 25 years of the WOFOST cropping systems model. Agricultural Systems 168, 154–167. https://doi.org/10.1016/j.agsy.2018.06.018

Demin, A., 2021. Fontane & Innovator – searching for alternatives [WWW Document]. Potato News. URL https://potatoes.news/fontane-innovator-searching-for-alternatives/

Den, T. ten, van de Wiel, I., de Wit, A., van Evert, F.K., van Ittersum, M.K., Reidsma, P., 2022. Modelling potential potato yields: Accounting for experimental differences in modern cultivars. European Journal of Agronomy 137, 126510. https://doi.org/10.1016/j.eja.2022.126510

Desloires, J., Ienco, D., Botrel, A., 2023. Out-of-year corn yield prediction at field-scale using Sentinel-2 satellite imagery and machine learning methods. Computers and Electronics in Agriculture 209, 107807. https://doi.org/10.1016/J.COMPAG.2023.107807

Diepen, C.A. v., Wolf, J., Keulen, H. v., Rappoldt, C., 1989. WOFOST: a simulation model of crop production. Soil Use And Management 5. https://doi.org/10.1111/j.1475-2743.1989.tb00755.x

Diepen, K. V., de Wit, A., n.d. The WOFOST model simulated processes, main parameters, limitations and calibration needs.

Divya, K.L., Priyank, H.M., Venkatasalam, E.P., Sudha, R., 2020. Crop Simulation Models as Decision-Supporting Tools for Sustainable Potato Production: a Review. Potato Research 64, 387–419. https://doi.org/10.1007/s11540-020-09483-9

Edelenbosch, R., Munnichs, G., 2020. Potatoes are the future - Three scenarios for hybrid potatoes and the global food supply. The Hague: Rathenau Instituut.

ESA, 2020. Crop type for all agricultural parcels in the Netherlands [WWW Document]. URL https://www.esa.int/ESA_Multimedia/Images/2022/02/Crop_type_for_all_agricultural_parcels_in_the_Netherlands

EU Parliament, 2021. Big data: definition, benefits, challenges (infographics) [WWW Document]. URL https://www.europarl.europa.eu/news/en/headlines/society/20210211STO97614/big-data-definition-benefits-challenges-infographics

FAO, 2022. FAOSTAT [WWW Document]. FAO. URL https://www.fao.org/faostat/en/#data/QCL/visualize (accessed 6.29.23).

Gao, B.C., 1996. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sensing of Environment 58, 257–266. https://doi.org/10.1016/S0034-4257(96)00067-3

Glenn, E.P., Neale, C.M.U., Hunsaker, D.J., Nagler, P.L., 2011. Vegetation index-based crop coefficients to estimate evapotranspiration by remote sensing in agricultural and natural ecosystems. https://doi.org/10.1002/hyp.8392

Goffart, J.-P., Haverkort, A., Storey, · Michael, Haase, N., Martin, M., Lebrun, P., Ryckmans, D., Florins, D., Demeulemeester, K., 2022. Potato Production in Northwestern Europe (Germany, France, the Netherlands, United Kingdom, Belgium): Characteristics, Issues, Challenges and Opportunities. Potato Research. https://doi.org/10.1007/s11540-021-09535-8

Gómez, D., Salvador, P., Sanz, J., Casanova, J.L., 2021. New spectral indicator Potato Productivity Index based on Sentinel-2 data to improve potato yield prediction: a machine learning approach. International Journal of Remote Sensing. https://doi.org/10.1080/01431161.2020.1871102

Gómez, D., Salvador, P., Sanz, J., Casanova, J.L., 2019. Potato Yield Prediction Using Machine Learning Techniques and Sentinel 2 Data. Remote Sensing. https://doi.org/10.3390/rs11151745

González-Sanpedro, M.C., Le Toan, T., Moreno, J., Kergoat, L., Rubio, E., 2008. Seasonal variations of leaf area index of agricultural fields retrieved from Landsat data. Remote Sensing of Environment 112, 810–824. https://doi.org/10.1016/J.RSE.2007.06.018

Herrmann, I., Pimstein, A., Karnieli, A., Cohen, Y., Alchanatis, V., Bonfil, D.J., 2012. Ground level LAI assessment of wheat and potato crops by Sentinel-2 bands. European Space Agency, (Special Publication) ESA SP 707 SP.

Herrmann, I., Pimstein, A., Karnieli, A., Cohen, Y., Alchanatis, V., Bonfil, D.J., 2011. LAI assessment of wheat and potato crops by VENμS and Sentinel-2 bands. https://doi.org/10.1016/j.rse.2011.04.018

Kasampalis, D.A., Alexandridis, T.K., Deva, C., Challinor, A., Moshou, D., Zalidis, G., 2018. Imaging Contribution of Remote Sensing on Crop Models: A Review. Journal of Imaging. https://doi.org/10.3390/jimaging4040052

Li, B., Xu, X., Zhang, L., Han, J., Bian, C., Li, G., Liu, J., Jin, L., 2020. Above-ground biomass estimation and yield prediction in potato by using UAV-based RGB and hyperspectral imaging. ISPRS Journal of Photogrammetry and Remote Sensing 162, 161–172. https://doi.org/10.1016/J.ISPRSJPRS.2020.02.013

Li, D., Miao, Y., Gupta, S.K., Rosen, C.J., Yuan, F., Wang, C., Wang, L., Huang, Y., 2021. Improving Potato Yield Prediction by Combining Cultivar Information and UAV Remote Sensing Data Using Machine Learning. Remote Sensing 2021, Vol. 13, Page 3322 13, 3322. https://doi.org/10.3390/RS13163322

Lin, Y., Li, S., Duan, S., Ye, Y., Li, B., Li, G., Lyv, D., Jin, L., Bian, C., Liu, J., 2023. Methodological evolution of potato yield prediction: a comprehensive review. Frontiers in Plant Science 14, 1–25. https://doi.org/10.3389/fpls.2023.1214006

Liu, Y., Feng, H., Yue, J., Fan, Y., Jin, X., Song, X., Yang, H., Yang, G., 2022. Estimation of Potato Above-Ground Biomass Based on Vegetation Indices and Green-Edge Parameters Obtained from UAVs. Remote Sensing 14. https://doi.org/10.3390/rs14215323

Luo, S., He, Y., Li, Q., Jiao, W., Zhu, Y., Zhao, X., 2020. Nondestructive estimation of potato yield using relative variables derived from multi-period LAI and hyperspectral data based on weighted growth stage. Plant Methods 16, 1–14. https://doi.org/10.1186/S13007-020-00693-3/FIGURES/6

Moroni, M., Porti, M., Piro, P., 2019. Design of a remote-controlled platform for green roof plants monitoring via hyperspectral sensors. Water 11, 1–13. https://doi.org/10.3390/w11071368

Mourad, R., Jaafar, H., Anderson, M., Gao, F., 2020. Assessment of Leaf Area Index Models Using Harmonized Landsat and Sentinel-2 Surface Reflectance Data over a Semi-Arid Irrigated

Landscape. Remote Sensing 12, 3121. https://doi.org/10.3390/RS12193121

Nain, A.S., Kersebaum, K.C., Dadhwal, V.K., 2012. Linking crop simulation model and remote sensing for wheat yield forecast. Journal of Agrometeorology 14, 482–490.

Nguyen, L.H., Robinson, S., Galpern, P., 2022. Medium-resolution multispectral satellite imagery in precision agriculture: mapping precision canola (Brassica napus L.) yield using Sentinel-2 time series. Precision Agriculture 1051–1071. https://doi.org/10.1007/s11119-022-09874-7

Paudel, D., De Wit, A., Boogaard, H., Marcos, D., Osinga, S., Athanasiadis, I.N., 2023. Interpretability of deep learning models for crop yield forecasting. Computers and Electronics in Agriculture 206, 107663. https://doi.org/10.1016/j.compag.2023.107663

Po, E.A., Snapp, S.S., Kravchenko, A., 2010. Potato yield variability across the landscape. Agronomy Journal 102, 885–894. https://doi.org/10.2134/agronj2009.0424

Porter, G.A., Opena, G.B., Bradbury, W.B., McBurnie, J.C., Sisson, J.A., 1999. Soil management and supplemental irrigation effects on potato: I. Soil properties, tuber yield, and quality. Agronomy Journal 91, 416–425. https://doi.org/10.2134/agronj1999.00021962009100030010x

Potatoes on Mars [WWW Document], n.d. . International Potato Center. URL https://cipotato.org/potatoes-mars-media-tools/

Ravensbergen, A.P.P., 2024. Exploring variability in yield , resource use efficiency and environmental impact of ware potato production in the Netherlands. Wageningen University and Research.

Ravensbergen, A.P.P., van Ittersum, M.K., Kempenaar, C., Ramsebner, N., de Wit, D., Reidsma, P., 2024. Coupling field monitoring with crop growth modelling provides detailed insights on yield gaps at field level: A case study on ware potato production in the Netherlands. Field Crops Research 308, 109295. https://doi.org/10.1016/J.FCR.2024.109295

Schut, A.G.T., Traore, P.C.S., Blaes, X., de By, R.A., 2018. Assessing yield and fertilizer response in heterogeneous smallholder fields with UAVs and satellites. Field Crops Research 221, 98–107. https://doi.org/10.1016/J.FCR.2018.02.018

Silva, J.V., Tenreiro, T.R., Spätjens, L., Anten, N.P.R., van Ittersum, M.K., Reidsma, P., 2020. Can big data explain yield variability and water productivity in intensive cropping systems? Field Crops Research 255, 107828. https://doi.org/10.1016/J.FCR.2020.107828

Tenreiro, T.R., García-Vila, M., Gómez, J.A., Jiménez-Berni, J.A., Fereres, E., 2021. Using NDVI for the assessment of canopy cover in agricultural crops within modelling research. Computers and Electronics in Agriculture 182, 106038. https://doi.org/10.1016/J.COMPAG.2021.106038

Van Ittersum, M.K., Cassman, K.G., Grassini, P., Wolf, J., Tittonell, P., Hochman, Z., 2013. Yield gap analysis with local to global relevance—A review. Field Crops Research 143, 4–17. https://doi.org/10.1016/J.FCR.2012.09.009

Van Ittersum, M.K., Leffelaar, P.A., Van Keulen, H., Kropff, M.J., Bastiaans, L., Goudriaan, J., 2003. On approaches and applications of the Wageningen crop models. European Journal of Agronomy 18, 201–234.

Van Ittersum, M.K., Rabbinge, R., 1997. Concepts in production ecology for analysis and quantification of agricultural input-output combinations. Field Crops Research 52, 197–208.

Vannoppen, A., Gobin, A., 2022. Estimating Yield from NDVI, Weather Data, and Soil Water Depletion for Sugar Beet and Potato in Northern Belgium. Water 2022, Vol. 14, Page 1188 14, 1188. https://doi.org/10.3390/W14081188

Vos, J., 1992. A Case History: Hundred Years Of Potato Production In Europe With Special Reference To The Netherlands. American Potato Journal 69, 731–751. https://doi.org/10.20595/jjbf.19.0_3

Wang, G.C.S., Jain, C.L., 2003. Regression Analysis: Modeling & Forecasting. Institute of Business Forec, NYC.

Wu, J., Wang, D., Bauer, M.E., 2007. Assessing broadband vegetation indices and QuickBird data in estimating leaf area index of corn and potato canopies. Field Crops Research 33–42. https://doi.org/10.1016/j.fcr.2007.01.003

8. Appendix

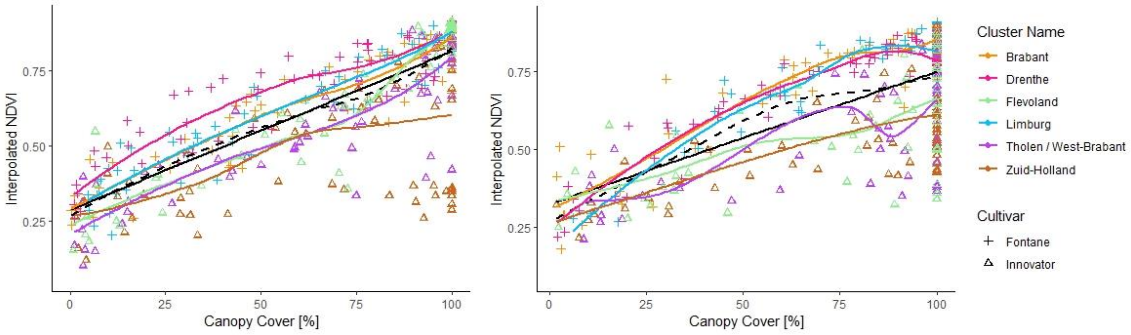
Appendix 1: Model 1 predicting actual yield with WOFOST potential yields only

```
Call:
lm(formula = yld_tot_DM ~ Yp * cluster_name, data = GC_obs_mod_merge_date_merged)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6008 -1.4628  0.1698  1.3069  3.8032

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        43.7382    11.3899   3.840 0.000241 ***
Yp                                 -1.7557     0.6711  -2.616 0.010581 *
cluster_nameDrenthe               -40.3355    15.1392  -2.664 0.009287 **
cluster_nameFlevoland             -27.2606    20.9624  -1.300 0.197089
cluster_nameLimburg               -32.0083    15.1474  -2.113 0.037631 *
cluster_nameTholen / West-Brabant -30.8435    16.4567  -1.874 0.064461 .
cluster_nameZuid-Holland          -49.4134    19.6262  -2.518 0.013758 *
Yp:cluster_nameDrenthe              2.3267     0.8762   2.655 0.009515 **
Yp:cluster_nameFlevoland            1.4865     1.2781   1.163 0.248177
Yp:cluster_nameLimburg              1.8108     0.8999   2.012 0.047468 *
Yp:cluster_nameTholen / West-Brabant 1.7093    1.0004   1.709 0.091290 .
Yp:cluster_nameZuid-Holland         2.8443     1.2017   2.367 0.020297 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.954 on 82 degrees of freedom
Multiple R-squared:  0.2317,    Adjusted R-squared:  0.1286
F-statistic: 2.248 on 11 and 82 DF,  p-value: 0.01912
```

Appendix 2: Model 1 predicting actual yield with WOFOST water limited yields only

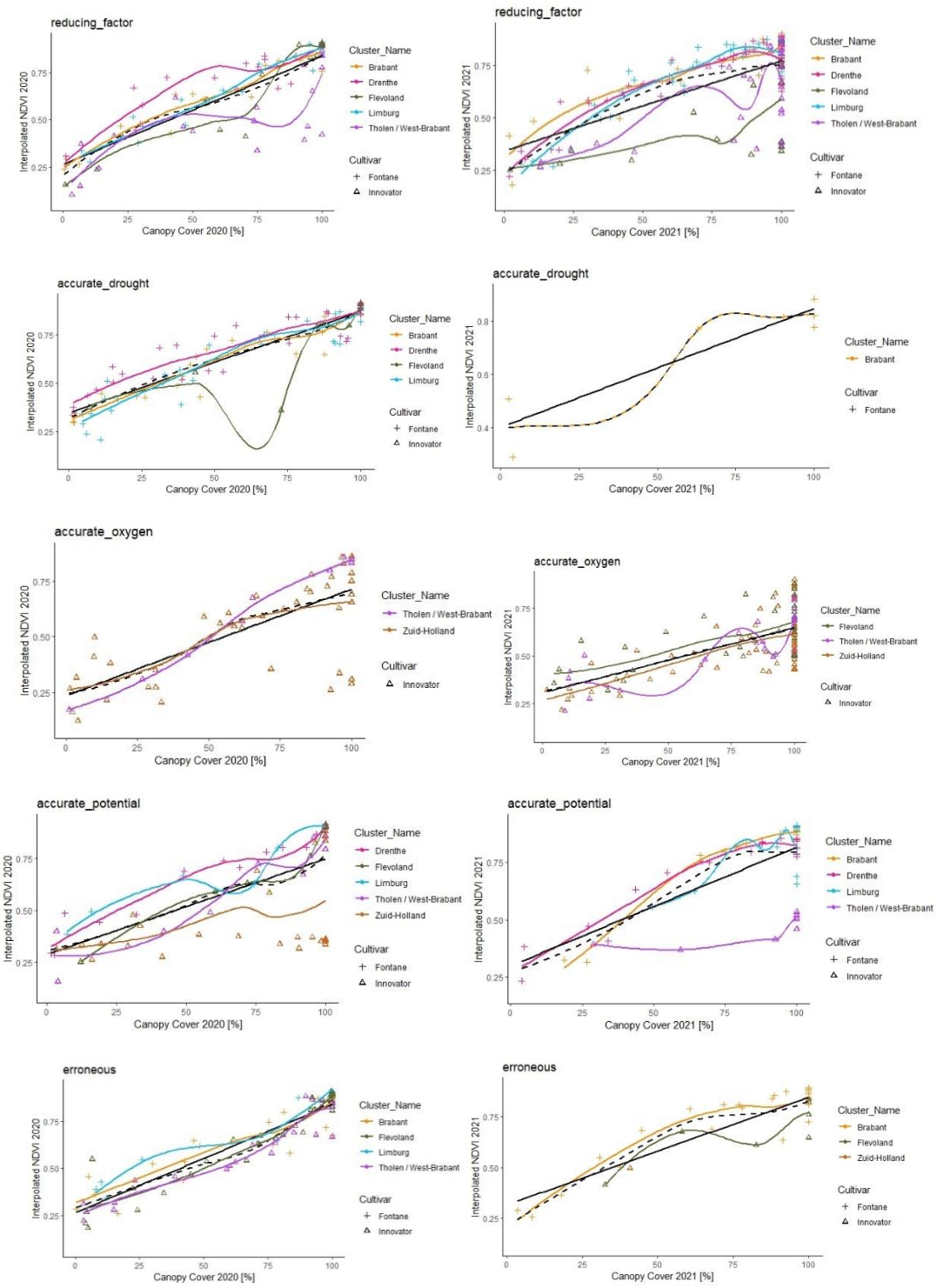
```
Call:
lm(formula = yld_tot_DM ~ Yw * cluster_name, data = GC_obs_mod_merge_date_merged)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4019 -1.3728  0.0182  1.2734  4.6283

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         13.06435    5.19657   2.514   0.0139 *
Yw                                   0.05930    0.34011   0.174   0.8620
cluster_nameDrenthe                 -7.14208    6.03054  -1.184   0.2397
cluster_nameFlevoland               -0.54035    5.59676  -0.097   0.9233
cluster_nameLimburg                 -5.20235    5.71708  -0.910   0.3655
cluster_nameTholen / West-Brabant    0.61842    5.76101   0.107   0.9148
cluster_nameZuid-Holland            -9.35485    5.95400  -1.571   0.1200
Yw:cluster_nameDrenthe               0.42290    0.39080   1.082   0.2824
Yw:cluster_nameFlevoland            -0.09234    0.37903  -0.244   0.8081
Yw:cluster_nameLimburg               0.27551    0.37730   0.730   0.4673
Yw:cluster_nameTholen / West-Brabant -0.17697    0.38860  -0.455   0.6500
Yw:cluster_nameZuid-Holland          0.61636    0.41653   1.480   0.1428
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.857 on 82 degrees of freedom
Multiple R-squared:  0.3056,    Adjusted R-squared:  0.2124
F-statistic:  3.28 on 11 and 82 DF,  p-value: 0.0009269
```

Appendix 3: The relationship between interpolated ndvi and field observed Canopy Cover cluster-wise for (i) 2020 and (ii) 2021. The black lines represent the average of all the clusters. The solid and the dotted lines represent the linear and the non-linear method for regression lines.



(i) (ii)

Appendix 4: The relationship between interpolated ndvi and field observed Canopy Cover field problem-wise for (i) 2020 and (ii) 2021. The black lines represent the average of all the clusters. The solid and the dotted black lines represent the linear and the non-linear method for regression lines.



(i) (ii)

Appendix 5: Model 2 predicting actual yield with Yw and interpolated NDVI for reducing factor category

```
Call:
lm(formula = yld_tot_DM ~ Yw + ndvi_interpolated * cluster_name,
    data = big_data_w_red_fac)

Residuals:
     Min       1Q   Median       3Q      Max
-1.77469 -0.71778  0.05678  0.63865  2.30563

Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                      6.34428    0.39259  16.160  < 2e-16
Yw                                               0.37359    0.02256  16.563  < 2e-16
ndvi_interpolated                               -0.28549    0.25993  -1.098    0.272
cluster_nameDrenthe                              0.14681    0.18464   0.795    0.427
cluster_nameFlevoland                           -1.01017    0.17618  -5.734 1.20e-08
cluster_nameLimburg                              0.11093    0.18708   0.593    0.553
cluster_nameTholen / West-Brabant               -1.42020    0.17538  -8.098 1.18e-15
ndvi_interpolated:cluster_nameDrenthe            0.33816    0.32518   1.040    0.299
ndvi_interpolated:cluster_nameFlevoland          0.53338    0.34349   1.553    0.121
ndvi_interpolated:cluster_nameLimburg            0.45366    0.32858   1.381    0.168
ndvi_interpolated:cluster_nameTholen / West-Brabant  1.42948    0.33131   4.315 1.71e-05

(Intercept)                                     ***
Yw                                              ***
ndvi_interpolated
cluster_nameDrenthe
cluster_nameFlevoland                           ***
cluster_nameLimburg
cluster_nameTholen / West-Brabant               ***
ndvi_interpolated:cluster_nameDrenthe
ndvi_interpolated:cluster_nameFlevoland
ndvi_interpolated:cluster_nameLimburg
ndvi_interpolated:cluster_nameTholen / West-Brabant ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8909 on 1432 degrees of freedom
  (78 observations deleted due to missingness)
Multiple R-squared:  0.5465,    Adjusted R-squared:  0.5433
F-statistic: 172.6 on 10 and 1432 DF,  p-value: < 2.2e-16
```

Appendix 6: Model 2 predicting actual yield with Yw, Canopy Cover and interpolated NDVI for reducing factor category

```
Call:
lm(formula = yld_tot_DM ~ Yw + avg_cnp_cvr + ndvi_interpolated *
    cluster_name, data = big_data_w_red_fac)

Residuals:
     Min       1Q   Median       3Q      Max
-1.73479 -0.63912  0.01867  0.61565  2.46194

Coefficients:
                                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                                      5.299229   1.109680   4.775 3.19e-06
Yw                                               0.433363   0.059306   7.307 4.41e-12
avg_cnp_cvr                                     -0.003828   0.003316  -1.154   0.2496
ndvi_interpolated                                0.231339   0.910860   0.254   0.7997
cluster_nameDrenthe                             -0.347489   0.702592  -0.495   0.6214
cluster_nameFlevoland                           -1.328523   0.697731  -1.904   0.0581
cluster_nameLimburg                             -0.305834   0.714277  -0.428   0.6689
cluster_nameTholen / West-Brabant               -1.680179   0.668404  -2.514   0.0126
ndvi_interpolated:cluster_nameDrenthe            0.939958   0.982975   0.956   0.3400
ndvi_interpolated:cluster_nameFlevoland          1.152736   1.018859   1.131   0.2591
ndvi_interpolated:cluster_nameLimburg            0.903706   0.992725   0.910   0.3636
ndvi_interpolated:cluster_nameTholen / West-Brabant  1.966998   0.970830   2.026   0.0439


(Intercept)                                      ***
Yw                                               ***
avg_cnp_cvr
ndvi_interpolated
cluster_nameDrenthe
cluster_nameFlevoland                            .
cluster_nameLimburg
cluster_nameTholen / West-Brabant                *
ndvi_interpolated:cluster_nameDrenthe
ndvi_interpolated:cluster_nameFlevoland
ndvi_interpolated:cluster_nameLimburg
ndvi_interpolated:cluster_nameTholen / West-Brabant *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9087 on 231 degrees of freedom
  (1278 observations deleted due to missingness)
Multiple R-squared:  0.5419,    Adjusted R-squared:  0.5201
F-statistic: 24.84 on 11 and 231 DF,  p-value: < 2.2e-16
```

Appendix 7: Model 3 predicting actual yield with WOFOST water limited yields and ancillary data for reducing_factor category

```
Call:
lm(formula = yld_tot_DM ~ Yw + elevation + clay_est + irri_quantity *
    cluster_name, data = big_data_w_red_fac)

Residuals:
     Min       1Q   Median       3Q      Max
-1.56550 -0.11976  0.08851  0.35793  1.08004

Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     4.709737   0.541790   8.693  < 2e-16 ***
Yw                                              0.453734   0.022204  20.435  < 2e-16 ***
elevation                                      -0.004009   0.007040  -0.569  0.56920
clay_est                                        0.080370   0.014985   5.363 1.04e-07 ***
irri_quantity                                   0.007473   0.002659   2.810  0.00506 **
cluster_nameDrenthe                            -0.998445   0.207247  -4.818 1.71e-06 ***
cluster_nameLimburg                             1.407678   0.166701   8.444  < 2e-16 ***
cluster_nameTholen / West-Brabant               0.082649   0.381804   0.216  0.82867
irri_quantity:cluster_nameDrenthe               0.003715   0.002442   1.521  0.12850
irri_quantity:cluster_nameLimburg              -0.019022   0.002803  -6.787 2.09e-11 ***
irri_quantity:cluster_nameTholen / West-Brabant -0.045184   0.003468 -13.027  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6157 on 888 degrees of freedom
  (622 observations deleted due to missingness)
Multiple R-squared:  0.7437,    Adjusted R-squared:  0.7408
F-statistic: 257.6 on 10 and 888 DF,  p-value: < 2.2e-16
```

Appendix 8: Model 4 predicting actual yield with WOFOST water limited yields, NDVI and ancillary data for reducing_factor category

```
Call:
lm(formula = yld_tot_DM ~ Yw + elevation + clay_est + ndvi_interpolated +
    irri_quantity * cluster_name, data = big_data_w_red_fac)

Residuals:
     Min       1Q   Median       3Q      Max
-1.58237 -0.11435  0.08668  0.37485  1.09109

Coefficients:
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     4.743806   0.561487   8.449  < 2e-16 ***
Yw                                              0.453492   0.022980  19.734  < 2e-16 ***
elevation                                      -0.005139   0.007178  -0.716  0.47425
clay_est                                        0.082873   0.016295   5.086 4.52e-07 ***
ndvi_interpolated                               0.017192   0.088191   0.195  0.84549
irri_quantity                                   0.007247   0.002737   2.647  0.00826 **
cluster_nameDrenthe                            -1.006927   0.212607  -4.736 2.56e-06 ***
cluster_nameLimburg                             1.373777   0.173103   7.936 6.72e-15 ***
cluster_nameTholen / West-Brabant              -0.034951   0.412117  -0.085  0.93243
irri_quantity:cluster_nameDrenthe               0.003687   0.002499   1.475  0.14048
irri_quantity:cluster_nameLimburg              -0.018328   0.002887  -6.348 3.57e-10 ***
irri_quantity:cluster_nameTholen / West-Brabant -0.044114   0.003766 -11.714  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6186 on 833 degrees of freedom
  (676 observations deleted due to missingness)
Multiple R-squared:  0.7392,    Adjusted R-squared:  0.7357
F-statistic: 214.6 on 11 and 833 DF,  p-value: < 2.2e-16
```