MSc Thesis in

Geoscience and remote sensing

Modeling Sentinel-1 observables for sugarbeet fields using machine learning

Fanhao Kong

MODELING SENTINEL-1 OBSERVABLES FOR SUGARBEET FIELDS USING MACHINE LEARNING

A STUDY ABOUT SAR ASSIMILATION TECHNIQUE

by

Fanhao KONG

to obtain the degree of Master of Science
in Geoscience and Remote Sensing
at the Delft University of Technology,
to be defended publicly on January 23rd, 2023.

Student number:	5213088	
Project duration:	February, 2022 – January, 2023	
Thesis committee:	Dr. Paco López-Dekker,	TU Delft , supervisor
	Prof. Susan Steele-Dunne,	TU Delft
	Dr. Marc Schleiss,	TU Delft
	T. (Tina) Nikaein,	TU Delft

To my parents

PREFACE

This thesis summarizes my work during my graduation project at the Delft University of Technology. During the past eight months, with machine learning technique, I worked on examining the potential of predicting Sentinel-1 backscatter and coherence data by the simulated sugarbeet biophysical variables from DSSAT, a crop model for making system behavior prediction from one location to others where given conditions.

The basics of DSSAT crop model and Sentinel-1 SAR data will be introduced in Literature Review. All the terms will be explained in detail later in this thesis. Letting people from different backgrounds understand the purpose, the methodology and the results of this graduation project is a goal I tried to achieve in this writing process.

Readers that are interested in the methodology that has been used will find it in Methodology and Data. Results and related analysis are shown in Results and Discussion. The insights and related future work this project brought is discussed in Conclusions and Recommendations.

I would like to thank my supervisor Dr. Paco López-Dekker for his instructions throughout the process. It has been a great privilege and joy to study under his guidance and supervision. I also want to thank my daily supervisor Tina Nikaein for her suggestions and constant encouragement. I am also extremely grateful to Prof. Susan Steele-Dunne and Dr. Marc Schleiss, have kindly provided me with assistance and helpful advice on many aspects.

From the very beginning of this study trip to the final presentation, I would like to thank Yibo Wang, Yang Xu, Ke Xu and Yuqi for their company, and my parents' remote support.

Den Haag, January 2023

SUMMARY

Over vegetation-covered areas, C-band SAR signals, including backscattering coefficients(σ^0) and interferometric coherence, are highly varied in both spatial and temporal extents owing to the dynamics in vegetation growth and soil hydraulic characteristics[1]. Having a good understanding of the relationship between them is very useful in many aspects, including agriculture management, hydrology, climate change, etc.

Acquiring adequate vegetation biophysical and soil variables is a challenge, due to it is costly and time-consuming to establish regional frameworks. The crop models, showing notable forecast skills for enlarged time and space scale crop prediction, are employed.

The specific crop model that drives this research is Decision Support System for Agrotechnology Transfer (DSSAT)[2]. DSSAT, a software package that accounts for the interactions between weather, soil and crop management options, is widely used to help users select and compare different options and predict crop-related results. In this paper, the crop simulation is carried out over all the sugarbeet fields in Noord-Brabant, Netherlands. The examination of applying the DSSAT sugarbeet crop model to the study area in the Netherlands is conducted. With the required input data, including daily weather observations, soil profiles, management practices and genotype information(cultivar), the model can automatically simulate and visualize the crop growth and soil variables. Our analysis indicates that the estimates from the current CSM-CERES-Beet model match the general sugarbeet cultivation situations in the Netherlands.

Although crop models can simulate the crop growth process and forecast crop yields, significant uncertainties can result from the unreliability of initial input data and model design. Some external environmental forcing mechanisms, including climate change and human disturbance, would derive unpredictable shifts in crop system responses. Thus, the SAR data begin to be assimilated to detect the anomalies from this step. Abundant available C-band SAR data from Sentinel-1 show a good potential to provide the relative true conditions of surface parameters, such as soil moisture, ground biomass and canopy geometry. Therefore, it opens an innovative perspective for monitoring regional crop conditions by comparing the crop growth estimates and Sentinel-1 C-band SAR observations.

The relation between the crop and soil variables and SAR signals is settled by the random forest regression model, which aims at investigating the correlation between vegetation biophysical variables and the C-band SAR patterns, as well as examining the effect of different feature combinations.

LIST OF ABBREVIATIONS AND SYMBOLS

Table 1: Acronym List

Acronym	Full Name
SM	Soil Moisture
SAR	Synthetic-aperture radar
DSSAT	Decision Support System for Agrotechnology Transfer
CSM	Cropping System Model
LAI	Leaf Area Index
CWAD	Tops weight
SW1D	Soil water content of the top layer(0-5cm)
SW3D	Soil water content of the third layer(15-30cm)
BRP	Basisregistratie Gewaspercelen
RF	Random Forest
PAR	Photosynthetically active radiation (PAR)
VWC	vegetation water content
NDVI	Normalized Difference Vegetation Index
RUE	radiation-use-efficiency
CV	Cross Validation
MSE	Mean Squared Error
ASCAT	Advanced SCATterometers
SMOS	Soil Moisture and Ocean Salinity
SMAP	Soil Moisture Active Passive
IW	Interferometric Wide swath
ASC	Ascending pass
DESC	Descending pass

Table 2: Symbols List

Symbols	Full Name
σ^0	Backscattering coefficients
σ^0_{VH}	Backscattering coefficient VH
σ^0_{VV}	Backscattering coefficient VV
R^2	R-squared
ρ	Spearman correlation coefficient

CONTENTS

1

INTRODUCTION

This chapter introduces the background to this research and the research questions that this thesis aims to address. The last section describes the general structure of this report.

1.1. BACKGROUND

Monitoring crop growth and yield is essential for farmers and government to make reasonable crop management decisions and quick responses to climatic shifts. In this context, there is a widespread demand for agricultural monitoring. The surveying of crop biophysical variables plays a significant role in agricultural scheduling and yield forecasting. Some key variables consist of Leaf Area Index (LAI), Tops weight(CWAD) and also soil water content which control energy and water circulation. LAI and CWAD are direct descriptors of vegetation features, while soil moisture is crucial for crop water detection and for irrigation decisions.

This research aims at linking radar observables with the states of crop and soil moisture. However, because the amount of available ground data is insufficient, we will be using crop model simulations as a proxy for ground data. Crop models are expected to be sensitive to climate and crop management changes and adjust their behaviors according to specific given conditions. The crop models provide us with a more systematic and quick way to understand the crop production process. Establishing a real agricultural framework and waiting for crop growth cycles are expensive and time-consuming. And sometimes particular regional-scale case is limited to assist in making agricultural strategies. With the implementation of crop models, it becomes efficient to collect crop biophysical variables. Researchers and farmers can gain a comprehensive understanding of each variable's influence and importance on the crop growth process by just modifying model parameters with fingertips. Moreover, under the guidance of model simulations, farmers can make adaptations in advance to achieve maximum yield or avoid crop losses derived from climate changes.

DSSAT (Decision Support System for Agrotechnology Transfer) is a software suite that comprises crop simulation models for over 42 crop types[2]. In this research, DSSAT was

performed for sugarbeet in the region of Noord-Brabant for the year 2017 due to that widely-covered sugarbeet can provide enough data for the regression analysis and can be handled with the DSSAT model. Management data generated from the previous field experiments were used to calibrate and evaluate the cropping system model of DSSAT. This guarantees the accuracy of crop simulations the model develops. There are lots of remote sensing approaches allowing to test the simulation performance of the crop model. Here, calculated LAI data from Sentinel-2 satellite are used as an assessment tool to test the validity of the sugarbeet simulations.

Besides using crop models to simulate crop growth, microwave remote sensing signals can be employed to assist in tracking these crop coefficients on a range of scales in the real world. We are interested in linking crop-growth-related parameters to radar observations as they are complementary from an information content point of view to the field and optical observations, and since they are all-weather and all-time available. Several studies suggest that radar data at C-band, which with a high spatial resolution, can reflect the attributes of the soil and vegetation conditions. Mostly, the measured time series of backscattering coefficients σ^0 (VV and VH) are used as an effective tool for crop monitoring due to their already exploited sensitivity to crop biophysical variables.

However, lots of studies[3] these days found that the time series of the coherence reflects in part the temporal evolution of the crop phenology: initial high coherence meets the condition of highly exposed soil, then the lower temporal values coincide with the progressive vegetation growth stage, and the subsequent increase matches the crop maturation. Hence, coherence will also be considered in this correlation examination.

The aim of this study is to gain a better understanding of the relationship between C-band radar observables and vegetation biophysical variables. The analysis was based on C-band backscattering coefficients and coherence data during an entire growing season of sugarbeet in 2017. By using the random forest regression tool, correlation analyses should be conducted between these radar observables and the vegetation features. Since different features have different predictive abilities, the selection of the features will have a significant impact on the regression results. Four relevant vegetation features, tops weight(CWAD), Leaf Area Index (LAI), the soil water content of the top layer(SW1D) and the soil water content of the root zone(SW3D) of the sugarbeet plant, are first selected to develop the regression model, followed by some feature importance analysis to provide suggestions on the further refinement of the selection.

1.2. PROBLEM STATEMENT

This graduation thesis focuses on providing an understanding on the work of detecting the correlation between the C-band SAR backscatter and coherence signals and sugarbeet-related variables in the Noord Brabant, the Netherlands.

To achieve this research goal, multiple sub-questions should be answered. DSSAT CSM-CERES-Beet model required data of an entire sugarbeet-growth period in 2017. The applicability of the model in the study area in the Netherlands will be validated. After the running process, the evaluation of the accuracy of the DSSAT model will be analyzed. Besides this crop model simulation, the performance of modeling radar observables by using the random forest regression model will be investigated as well, followed by discussions about feature importance. Therefore, the following sub-questions will be

answered in this master thesis:

1. Evaluating DSSAT CSM-CERES-Beet model performance.
    - What is the CSM-CERES-Beet model accuracy of simulating the sugarbeet growing process in the study area in the Netherlands?
2. Evaluating the correlation performance.
    - How can the accuracy of the random forest model be increased?
    - How well does crop biophysical variables correlate with the backscattering coefficients(VH and VV)?
    - How well does crop biophysical variables correlate with the VV polarized interferometric coherence?
    - How does feature selection affect the prediction accuracy of SAR signals?

1.3. OVERALL STRUCTURE

Chapter 2 goes through the existing studies about the DSSAT crop model and the utility of Sentinel-1 SAR data for crop growth detection. Chapter 3 presents the methodology used in this study. Chapter 4 discusses the results and answers the main research questions stated in the previous section. The conclusions and concerns about further studies of this research are given in Chapter 5.

2

Literature Review

This study uses DSSAT(Decision Support System for Agrotechnology Transfer) crop model to simulate dynamic crop growth for sugarbeet. DSSAT is a software developed by integrating the knowledge about genotype specific parameters of variable crop types to simulate options for crop management over time and spatial series[4]. And the Sentinel-1 backscattering coefficient(VH and VV) and coherence VV are SAR signals need to be predicted.

2.1. DSSAT

2.1.1. DSSAT Structure

There is a large volume of published studies describing that the available ground data is not sufficient to satisfy the increasing demand for agricultural decisions. The DSSAT crop model was derived to integrate knowledge about soil, weather and management for making better system behavior prediction under different conditions [5].

Figure 2.1 provides a diagram of the DSSAT general architecture for versions up to 3.5. In these versions, one main potential limitation is that different crop models have their own soil models, resulting in inconsistency and less efficient incorporation of different sets of programming code. Thus, DSSAT has been re-designed these years. Now, the DSSAT is a collection of independent crop simulation models for over 42 crops (as of Version 4.8) and tools to promote efficient operation of the models. The tools comprise compatible database programs describing weather, soil, management and experimental data, software, and application programs.

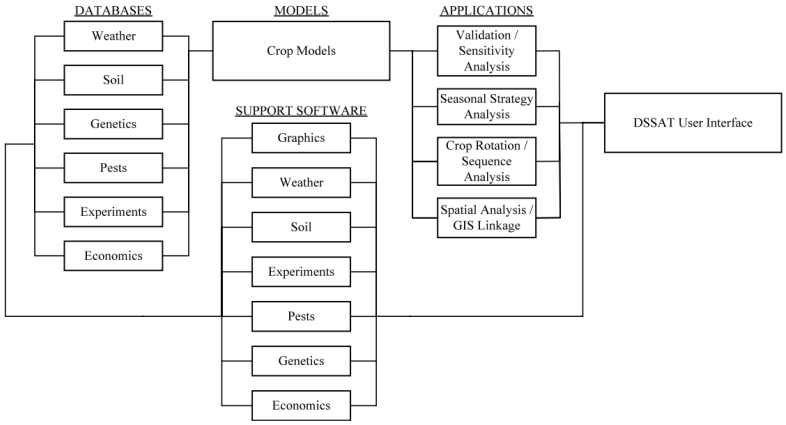


Figure 2.1: Diagram of database, support software and applications for DSSATv3.5.[4]

2.1.2. DSSAT CROPPING SYSTEM MODEL(CSM)

At the heart of the revised DSSAT version is a cropping system model (CSM), which is restructured as a modular format.

DSSAT now integrates all crop models into CSM as modules using a common soil module and a common weather module based on one single set of codes. This reversion of CSM is introduced in Irmak, A. et al.[6].

The CSM simulates expected growth and development of a crop normally in daily or hourly time steps, as well as the soil moisture, nutrient dynamics and management practices, so processes that occur under the cropping system such as organic matter consumption, evaporation and runoff are also taken into consideration. The main components of CSM are shown in Fig 2.2 and include:

1. Main program:
   Controls the timing of simulations.

2. Land Unit module:
   Controls all the simulation processing work and data transfer between primary modules.

3. Primary modules:
   Simulates the diverse processes independently. Primary modules are comprised of weather, management, soil-plant-atmosphere, soil, and plant submodules.

The required CSM input dataset contains daily weather data, soil data, and management data about detailed characteristics of variable genotypes (e.g.row spacing, seeding population, irrigation application). The minimum datasets for DSSAT-CSM operation are listed in Table 2.1.

2.2. SENTINEL-1 SAR DATA

Several studies have examined the capabilities of using microwave sensors, such as Advanced SCATterometers (ASCAT), ESA's Soil Moisture and Ocean Salinity (SMOS)

Input dataset	components
Weather	Daily total incoming solar radiation ($MJ/m^s - day$), Maximum and minimum daily air temperature ($^\circ C$), and Daily precipitation (mm).
Soil	Upper and lower horizon depths (cm), sand, silt, and clay percentage , bulk density(kg/m^3), organic carbon density($kgCarbon/m^2$), pH, and root growth information.
Management	Cultivar type, planting date, density($plants/m^2$) and depth(cm), row spacing(cm) and direction, irrigation and fertilizer dates, methods and amount, harvest dates and methods

Table 2.1: Contents of minimum datasets for DSSAT-CSM operation.[7]

mission and NASA's Soil Moisture Active Passive (SMAP), for monitoring vegetation signals[8]. They are unaffected by weather conditions and provide coverage over large areas. However, the spatial resolution of these observations is not high enough for many applications. For example, in this research, the vegetation growth should be monitored within many separate small fields. If we use the SMOS satellite with a coarse spatial resolution of 40 km, the sensor can not account for the spatial dynamics of each field. The launch of SAR satellites break through the limits of temporal and spatial resolutions, and is making unprecedented opportunities for monitoring and optimizing agricultural management, especially the Sentinel-1 satellites. Sentinel-1 can provide systematic observations with a quite short revisit time, and the promise of continuity which allows developing monitoring tools and services.

The Sentinel-1 Mission (Sentinel-1A and 1B) from the Europe's Copernicus programme was launched in 2014 and 2016, respectively. The satellites carry C-band Synthetic Aperture Radar(SAR) at 5.405 GHz. The default acquisition mode over(non-polar) land is the Interferometric Wide swath (IW) mode serving both co- and cross-polarized(VV and VH) data over a 250 km swath at a 20 m spatial resolution. Each satellite has a temporal revisit time of 12 days, and the revisit frequency is 1-4 days in Europe when integrating ascending (ASC) and descending (DESC) pass directions from both Sentinel-1 satellites.

2.2.1. MICROWAVE INDICES FROM SENTINEL-1

Both C-band co-polarized and cross-polarized backscattering coefficients σ^0 yield valuable information about crop structure and type changes, as well as moisture differences. Additionally, the crop biomass, vegetation water content (VWC) and LAI are associated with SAR σ^0. Many studies[8][9] suggest that the co-polarized(HH and VV) data was usually used for soil monitoring, while the cross-polarized (VH and HV) backscattering coefficients have a high correlation with vegetation conditions.

Many previous papers[8][10] indicate that the amount of energy backscattered over

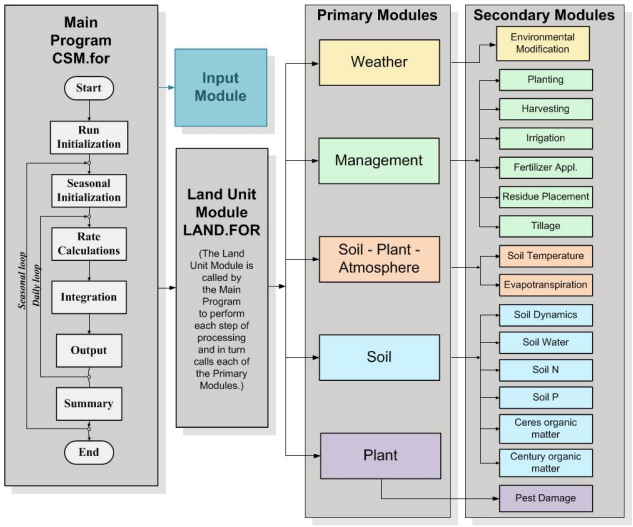


Figure 2.2: DSSAT Cropping System Model schematic.[7]

a vegetated region does not only contain direct scattering from the vegetation itself, but also the sum of the attenuated backscatter producing from the underlying surface and the vegetation-soil interaction(Fig 2.3). The figure suggests that the scattering or attenuation of radar signals will differ in accordance with the dielectric properties and physical geometry of the vegetation. Therefore, analysing the radar backscatter under natural vegetation-covered conditions can be complicated.

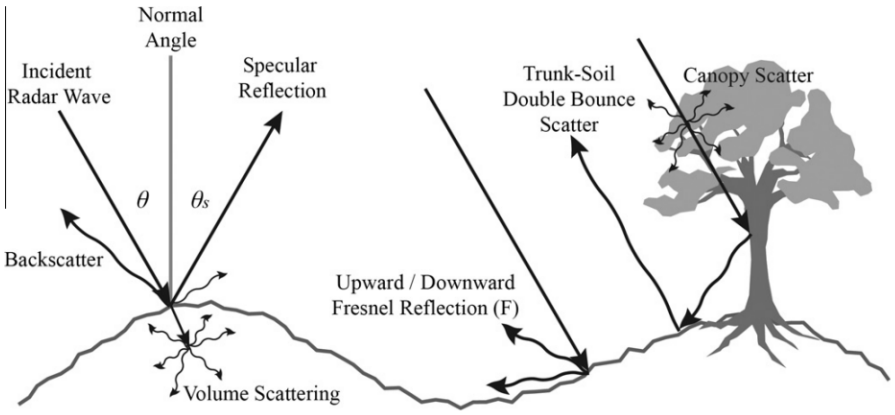


Figure 2.3: Conceptual illustration of the incident radar signal across a vegetated surface[10].

The temporal behaviors of σ^0 over crop cultivated area can reflect important information, such as crop dynamics and management, as well as environmental

conditions. Thus it is necessary to interpret them. Considering the impact of polarization, backscatter responses from direct ground and canopy have a significant contribution to VV polarization, while VH scattering is dominated by the trunk-ground double bounce and by volume scattering[9].

Figure 2.4 shows the C-band Sentinel-1 backscatter data of sugarbeet as an example. The time series of backscatter signals exhibit an explicit seasonal cycle and the temporal variations in both VV and VH channels are quite similar. After sowing in March and April, a downward trend of backscatter VV is derived by stem elongation periods, during which time the direct soil contribution is attenuated by the developing sugarbeet. At this time, an increase in volume scattering is counteracted by a decline of the stem-ground double bounce contribution, leading to the decrease in VH. From mid-May to mid-June, the backscatter values start to increase, as a result of the significant increase in volume scattering associated with sugarbeet leaf development. Afterwards, relatively constant backscatter values are attained due to the completion of above-ground vegetation coverage. During other time intervals without planted sugarbeet, the backscattering coefficients are mainly in response to the soil contribution.

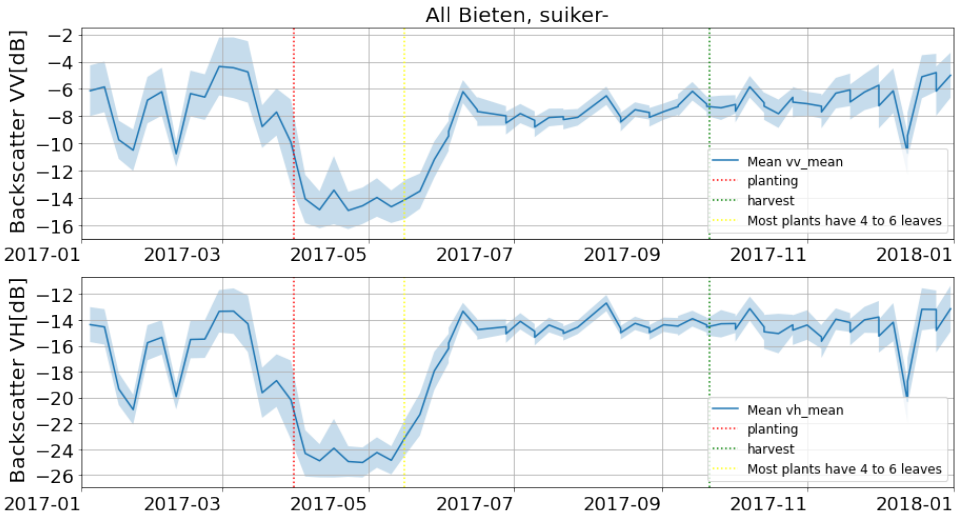


Figure 2.4: Time series of Sentinel-1 orbit 88 backscatter data for all sugar beet parcels in the Noord-Brabant; (top) VV; (bottom) VH.

Aside from backscattering coefficients derived from single images, interferometric processing of pairs of images provide additional observables. In particular, the interferometric coherence reveals the magnitude of similarity between the two associated radar acquisitions[11]. The coherence values always fluctuate from zero to one. Loss of coherence is known as decorrelation. The related factors that cause decorrelation can be categorized as thermal decorrelation, spatial decorrelation, and temporal decorrelation[12]. Thermal decorrelation is used to describe system noises. Spatial decorrelation refers to the effect of viewing geometry of the radar system.

Apart from the system noises and motion of platforms, the loss of coherence can be caused by the temporal target changes. On the issue of agricultural crop monitoring, crop growth detection is based on the outcome, which integrates the consistent high coherence from the soil layer and the temporal decorrelation owing to the crop layer. The correlations between coherence and crop height as well as canopy cover have been developed by several studies.([13], [14]). Consequently, coherence is a suitable statistic that can be used to explore the volume scattering differences over crop fields due to distinct altered crop phenological stages.

Therefore, in this research, coherence is used as an additional information source, complimentary to the backscatter-based crop monitoring technique.

2.2.2. AGRICULTURAL SANDBOXNL

We have established that Sentinel-1 temporal backscatter and interferometric coherence data at C-band seem to be well-suited for surface changes. However, the lack of straightforward access to interpreting SAR data is an obstacle that hinders further studies. To break down this barrier, a parcel-level database called SandboxNL for the Netherlands has been built[15]. The division of crop parcels is based on Basisregistratie Gewaspercelen (BRP), which contains the parcel locations and the crop type linked to them[16].

SandboxNL includes continuous annual data for each province, starting from 2017. The database contains information of Sentinel-1A/B SAR data as well as Sentinel 2 data. For Sentinel-1A/B SAR data, the SandboxNL is composed of parcel-level spatially averaged backscattering coefficients(VV, VH), the parcel-averaged cross-polarization ratio (VH/VV), as well as their corresponding standard deviation[15]. Each parcel has its unique OBJECTID, and some attribute features, such as pixel counts, azimuth angle, etc. Moreover, all the data are separately stored for each relative orbit. Recently, as the value of interferometric coherence data has been widely discovered, SandboxNL also provides interferometric coherence dataset.

All of the data above is stored together with static information in pickle files. Figure 2.5 shows the list of variables and descriptions in the SandboxNL.

In the presence of SandboxNL, the time series patterns of Sentinel-1 data are increasingly intended to be encompassed in agricultural applications at field scale.

No.	Variable	Description
1	Time	Acquisition time of the SAR image
2	Longitude	Longitude of the parcel centroid
3	Latitude	Latitude of the parcel centroid
4	VV_mean	Average VV polarization backscatter intensity over parcel
5	VV_std	Standard deviation VV polarization
6	VH_mean	Average VH polarization backscatter intensity over parcel
7	VH_std	Standard deviation VH polarization backscatter over parcel
8	CR_mean	Average cross-pol ratio (VH/VV) over parcel
9	CR_std	Standard deviation cross-pol ratio (VH/VV) over parcel
10	LIA	Local incidence angle (adjusted for local topography)
11	EA	Approximate viewing incidence angle in the GEE S1 GRD product
12	AZA	Azimuth angle
13	OID	Object ID (unique per parcel)
14	MID	Sentinel-1 satellite mission ID (0 = S1A and 1 = S1B)
15	RO	Relative orbit tracks of Sentinel-1 used in database
16	Pix	Pixel count of the selected parcel
17	Flag	Flags assigned to border parcels for each relative orbit

Figure 2.5: List of variables and descriptions in the SandboxNL[15].

3

METHODOLOGY AND DATA

3.1. STUDY AREA AND CROP

3.1.1. STUDY AREA

The study site is located in Noord-Brabant, which is relatively flat and mostly above sea level. A great number of lands in the province have been cultivated into agricultural land and forest. The location of the Noord-Brabant in the Netherlands is shown in Fig 3.2, along with the spatial distribution of the four main grown crops. Crop types and parcel boundaries are based on the Basisregistratie Gewaspercelen (BRP)[17].

Figure 3.1 shows that the soil at the surface in the north and lower northwest of the province is loam and clay, while the rest parts are sand. The general features of land use in Noord-Brabant are associated with its soil type distribution. As shown in Fig 3.2, maize is farmed on the sandy regions while the cultivations of sugar beet, wheat, and potatoes present densely in northwest marine clay area[18].

In Noord-Brabant, the average daytime temperature during winter is around $6°C$, while the average maximum temperature in the summertime can reach $25°C$[19].

3.1.2. CROP SELECTION

The crop type selected for this research is sugarbeet. The selected crop type is supposed to satisfy some conditions. Firstly, in order to get sufficient simulation results in subsequent steps, the number of the agricultural parcels of the crop type in the study area should be at least thousands; secondly, the crop is expected to have a relatively high coherence time series; and, thirdly, the crop type can be modeled in DSSAT.

The coverage of sugarbeet in Noord-Brabant is shown in Fig 3.3. Sugarbeet is a root crop, cultivated as one of the main sources of sugar production worldwide. Sugarbeet thrives in temperate climates with a growing season of about 170–200 days. Planting usually starts in late March or early April for summer crops and harvesting is carried out promptly in late September or early October so that it is completed before the soil freezes[20].

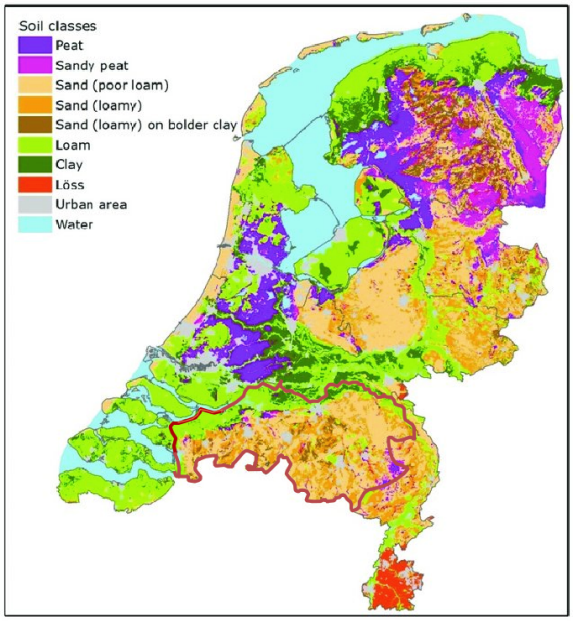


Figure 3.1: Map of soil classes in Noord-Brabant in 2017.

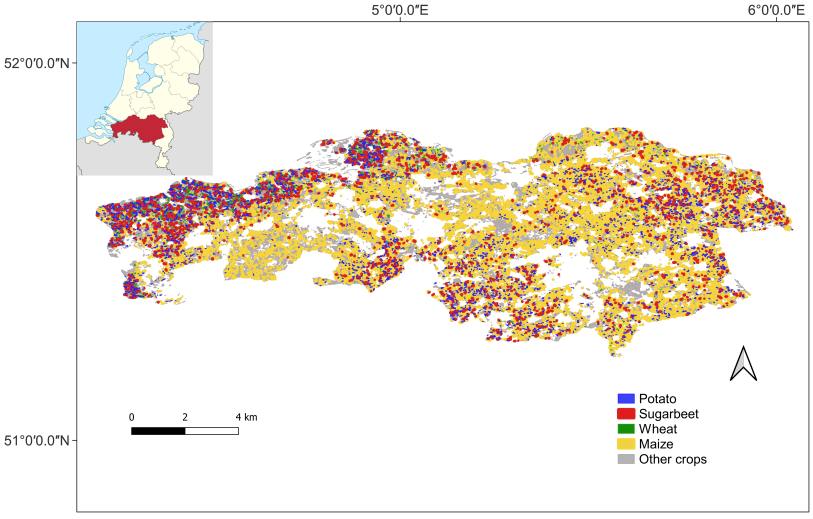


Figure 3.2: Map of the location of the study area and main crop types in Noord-Brabant in 2017.

Figure 3.4 shows the Sentinel-1 parcel-averaged coherence data(VV and VH) from January 2017 to January 2018 for 2437 sugarbeet parcels in the Noord-Brabant. The plotted curves represent the mean coherence values for the corresponding acquisition dates, while the shadows represent the corresponding standard deviation. The general sowing date is at the end of March, and the usual harvest period is from late September. Temporal patterns show that during March and April, the fluctuating sowing conditions lead to high standard deviation over the parcels. After that, the coherence VV keeps at a high level(>=0.5) until June, this period is assumed to link with the stages of sugarbeet leaf development.

Then the coherence is constant below 0.4, indicating all parcels are covered in vegetation leaves. Here, we need to pay attention to a bias resulting from the coherence. The coherence is biased for low coherence values, thus these low values are higher than their actual values[12]. In this fully-covered period, since we know that the actual coherence can be very low, even close to zero, thus the coherence in our study is about 0.2-0.3.

Coherence and standard deviation across parcels generally increase since late September, suggesting the start of harvesting activities.

3.2. DATA

3.2.1. DSSAT INPUT DATA

The CSM-CERES-Beet model considers the sugarbeet as an annual crop and requires standard DSSAT input data, including weather data, soil features data, and crop management data.

METEOROLOGICAL DATA

Daily weather data (solar radiation, minimum, and maximum temperature) for CSM-CERES-Beet model were collected from all 54 KNMI local weather stations in Netherlands[21]. While the daily precipitation data in the Netherlands are measured on +- 300 locations and calculated as gridded files[22]. The averaged temperature values are at a daily step while the precipitation data are summed up to acquire daily volumes.

SOIL DATA

The soil data comes from the SoilGrid portal, a system for mapping the spatial distribution of soil profiles to a global extent. The experimental soil properties containing clay content, bulk density, PH, etc for 6 standard soil depth intervals[23].

MANAGEMENT DATA

The principal field management data consists of sugarbeet cultivar, planting date, planting depth, irrigation applications dates, harvesting date, etc. All the collected management data refers to the general sugarbeet cultivation information in the Netherlands. Field management data for 2017, 2018, and 2019 are provided in Table 3.1.

A cultivar is a genetically distinct variety of crop plants, generally adapted to a specific region[24]. Thus, in DSSAT, cultivar selection is vital to reflect the genetic background and agronomic characteristics of a crop's genetic diversity[25]. Some coefficients are needed to define a cultivar. The cultivar genetic coefficients for the

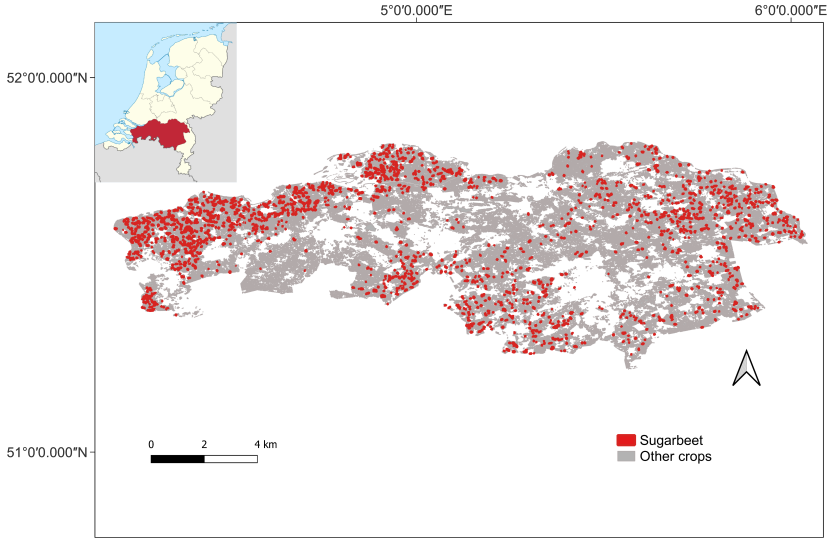


Figure 3.3: All sugar beet parcels in the Noord-Brabant in 2017.

parameter	2017	2018	2019
Planting date	March 31	April 20	April 11
Row spacing(cm)	50	50	50
Planting depth(cm)	3	3	3
Plant population at seeding(plants/m^2)	10	10	10
Harvesting	September 21	October 1	September 9

Table 3.1: Field management for sugarbeet experimental plots in the Netherlands.

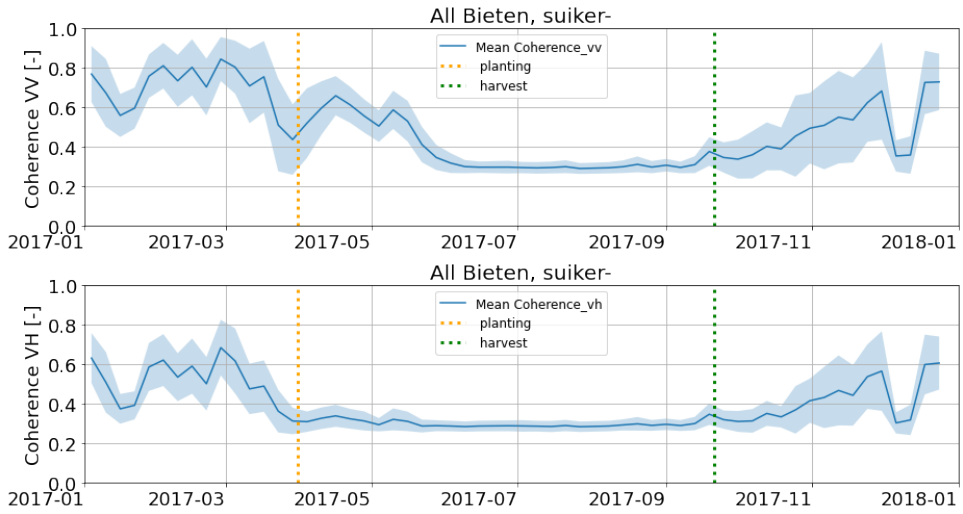


Figure 3.4: Time series of Sentinel-1 orbit number 37 coherence data for all sugar beet parcels in the Noord-Brabant; (top) VV ; (bottom) VH. The orange and green vertical line line indicate the general planting and harvest dates across all sugar beet parcels in the domain, separately.

CSM-CERES-Beet model contain P1, P2, P5, G2, G3, and PHINT[26]. Their definitions and units are listed in Table 3.2.

It is demanding to derive and calibrate a new cultivar for the Netherlands as the magnitude of each genetic coefficient required for the model, some of which are not readily available, is supposed to agree with the gathered sugarbeet growth information. A new German sugar beet cultivar BTS940(Table 3.2) was calibrated by the management data located 30 km from Stuttgart and has been updated in the newly released DSSAT4.8 version. Since the location is adjacent to the study area, a comparison experiment of crop management practices and meteorological conditions between the two study sites in Germany and the Netherlands is set up to examine whether this cultivar can be adopted in this research.

The comparison consists of 2 aspects mainly:

1. Crop management data

   The collected sugarbeet management data for the experiment conducted in Germany is from 2016 to 2018. Table 3.3 illustrates that the entire seasonal growing period and cultivation management of sugarbeet are very similar in the two study sites.

2. Meteorological data

   Temperature and precipitation will exert a significant influence on crop emergence, root yield, and quality, as well as some characteristics of sugarbeet canopies. Article[26] reported the time series of daily minimum temperature and

parameter	Definition	Units	BTS940
P1	Growing Degree Days from the seedling emergence to the end of the juvenile phase (juvenile group of leaves, depending on the cultivar up to 15–20 leaves)	$°C-d$	760.0
P2	Photo period sensitivity	hr^{-1}	0.0
P5	Thermal time from leaf growth to physiological maturity	$°C-d$	700.0
G2	Leaf expansion rate during leaf growth stage	$cm^2cm^{-2}d^{-1}$	420.0
G3	Maximum root growth rate	$gm^{-2}d^{-1}$	27.5
PHINT	Phyllochron interval, the interval in thermal time between successive leaf tip appearances	$°C-d$	43.0

Table 3.2: Sugar beet cultivar(BTS940) genetic coefficients for CSM-CERES-Beet model[26].

	Planting date				Harvest days after planting				Planting depth (cm)	Plant population at seeding (plants/m^2)
	2016	2017	2018	2019	2016	2017	2018	2019		
NL	-	3.31	4.20	4.11	-	174	164	151	3	10.0
DE	4.29	4.4	4.18	-	177	184	169	-	2	10.7

Table 3.3: Comparison of sugarbeet management data between study areas in Netherlands and Germany(NL: Netherlands, DE: Germany.

cumulative rain (2016-2018) of the study area in Germany, thus these two factors are used for comparing meteorological conditions between the two study sites in Netherlands and Germany. If the magnitudes of the meteorological conditions are also similar, then we can adopt the cultivar BTS940 in our research.

As shown in Fig 3.5 and Fig 3.6, the extents and trends of minimum temperature and cumulative rain between the two sites are comparable. Even some small fluctuations are corresponding, such as the unusual drop in temperature, which even below $0^{\circ}C$ after sugarbeet had been planted in 2017, occurs in both sites.

The idea of this cultivar selection is to find the modeled growth profile that is most similar to the Dutch one. In the comparisons above, it is found that the sugarbeet cultivation conditions in these two experimental sites are similar, suggesting that this new German cultivar BTS940 is suitable for this research in Noord-Brabant.

3.2.2. SENTINEL-1 DATA

For sugarbeet, whose plant density and vegetation cover are relatively high, coherence VV is generally more correlated with the growing stages. Therefore, we use three SAR signal channels involved in the sentinel-1 data collection for this research: VV and VH polarized backscatter, and VV polarized coherence. The Netherlands is covered by 6 Sentinel-1 satellite tracks. As shown in Fig 3.7, tracks 15, 161, 139, and 110 are not considered since they only cover part of the Noord-Brabant which is labeled in red. Agricultural SandboxNL is applied to examine the number of missing observations for sugarbeet parcels in tracks 88 and 37, respectively(Table 3.4).

As introduced in section 3.1.2, the total sugarbeet parcels in Noord-Brabant is 2437. In this case, the advantage of selecting orbit 37 is to maintain a sufficient and relatively complete sugarbeet pattern of the whole province.

Based on the discussion above, the mean parcel-leveled backscatter VV and VH and coherence VV of orbit 37 with a temporal resolution of six days between 1 April 2017 to 10 September 2017, corresponding to the growth period of sugarbeet, is collected as the radar dataset for this research.

orbit	Number of parcels backscatter	Number of parcels coherence
88	2401	1626
37	2436	2407

Table 3.4: Number of 2017 sugarbeet parcels in SandboxNL dataset for orbit37 and 88 in the study area.

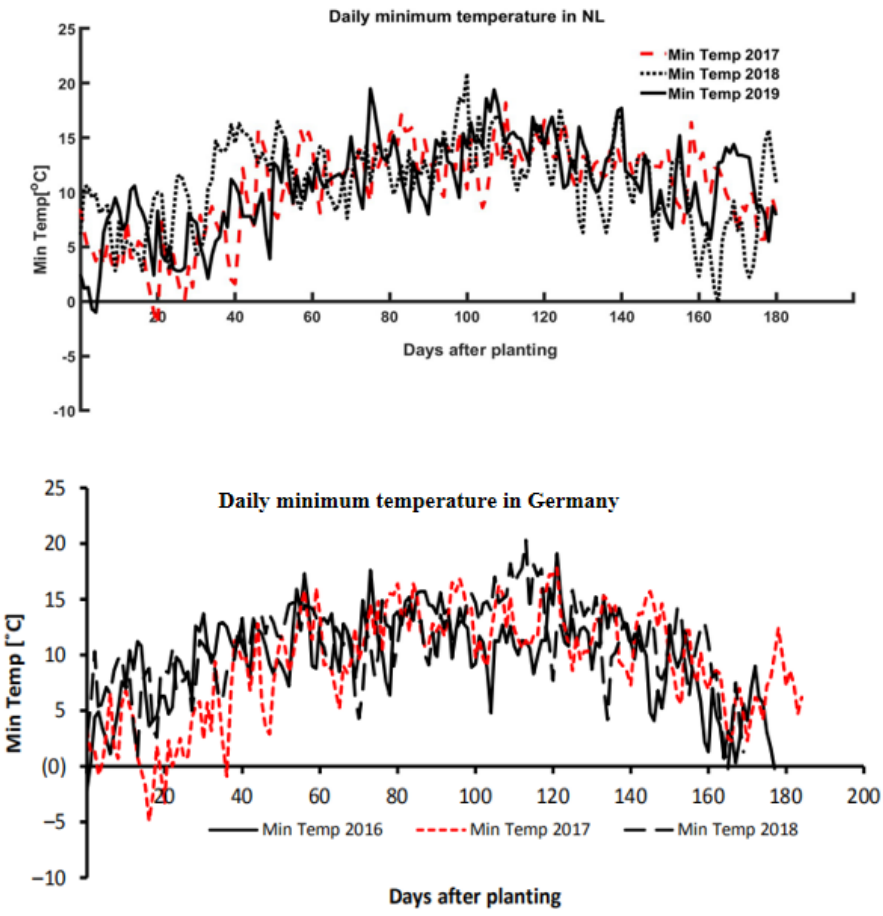


Figure 3.5: Time series of the daily minimum temperature of study areas in the Netherlands and Germany; (top) Netherlands ; (bottom) Germany[26].

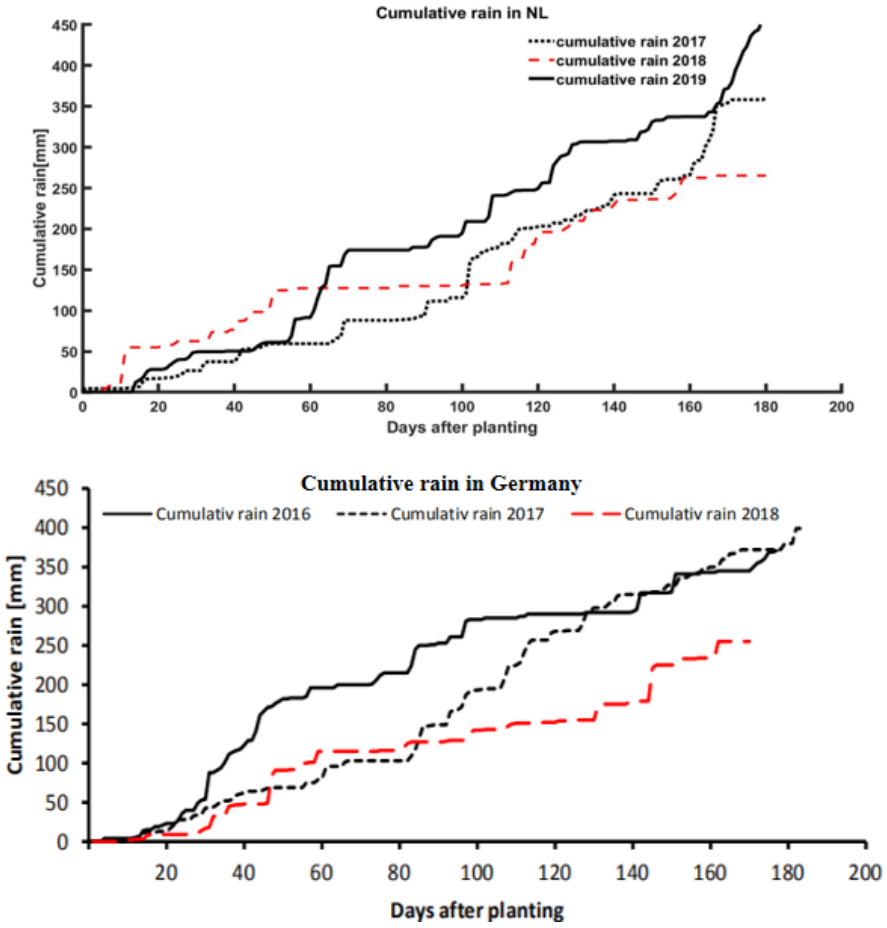


Figure 3.6: Time series of cumulative precipitation data of study areas in the Netherlands and Germany; (top) Netherlands ; (bottom) Germany[26].

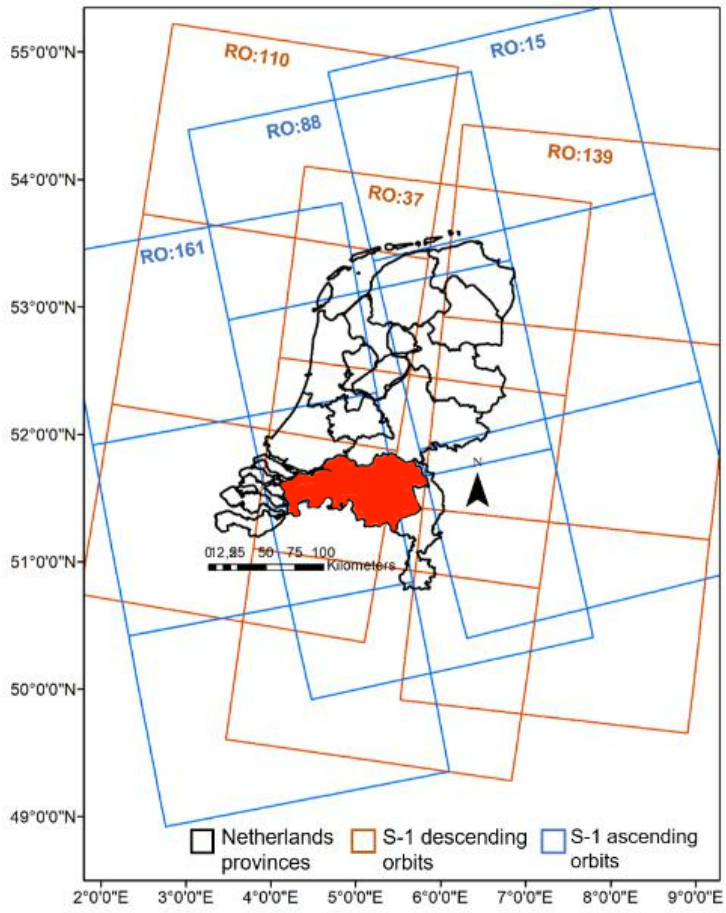


Figure 3.7: Sentinel-1 coverage over the Netherlands based on six relative orbits. The orange polygons show descending orbits(139,110,37) and the blue show ascending orbits(161,88,15), Noord-Brabant province is labeled in red[15].

3.3. METHODOLOGY

This study is about examining whether the Sentinel-1 C-band radar observables and the simulated plant and soil moisture coefficients from the DSSAT CSM-CERES-Beet model can correlate in a close manner in the sugarbeet-covered region. Efforts are made to answer the research questions proposed in Problem Statement . Figure 3.8 shows the workflow of the methodology followed in this study. The methodology used is described in the next subsections.

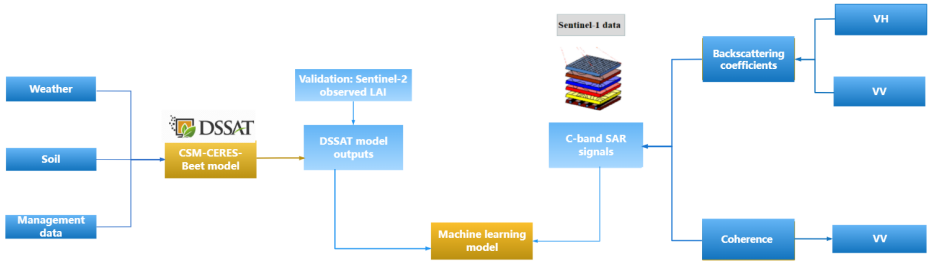


Figure 3.8: Workflow of the methodology, including two models which are labeled in khaki, and multiple required data sets which are labeled in blue.

3.3.1. SUGARBEET GROWTH SIMULATION

Firstly, the sugarbeet simulation is conducted by DSSAT. The simulation is from planting to harvesting in 2017 over a total of 2437 parcels in Noord-Brabant. Then the further evaluation of the model was conducted with the help of calculated temporal LAI data from Sentinel-2 satellite.

DSSAT SIMULATION SETUP

The SM and plant growth variables can be automatically generated after related simulation options are set. Some parameter settings are needed to be taken care to avoid model crashing or the model simulating inadequately[27]. Figure 3.9 shows an overview of inputs and outputs in DSSAT CSM-CERES-Beet model.

In every experiment file, each sugarbeet parcel is defined as a field factor level and assigned a level number, afterwards the treatments are constructed. The treatment schemes in one experiment file can only be two-digit format. Therefore, there are 25 sugar beet experiment files named from SBAL1799 to SBYL1761 over 2437 parcels for 2017 in total. The amount of treatments for SBAL1799 to SBXL1799 experiment files is 99, ranging from 1701 to 1799, while SBYL is formed with the remaining 61 field levels. In this research, the different management approaches are based on that every level has its own corresponding weather and soil management factors.

Here, experiment file SBAL1799 is taken as an example, 99 treatments are required to be set up in the way as following defined:

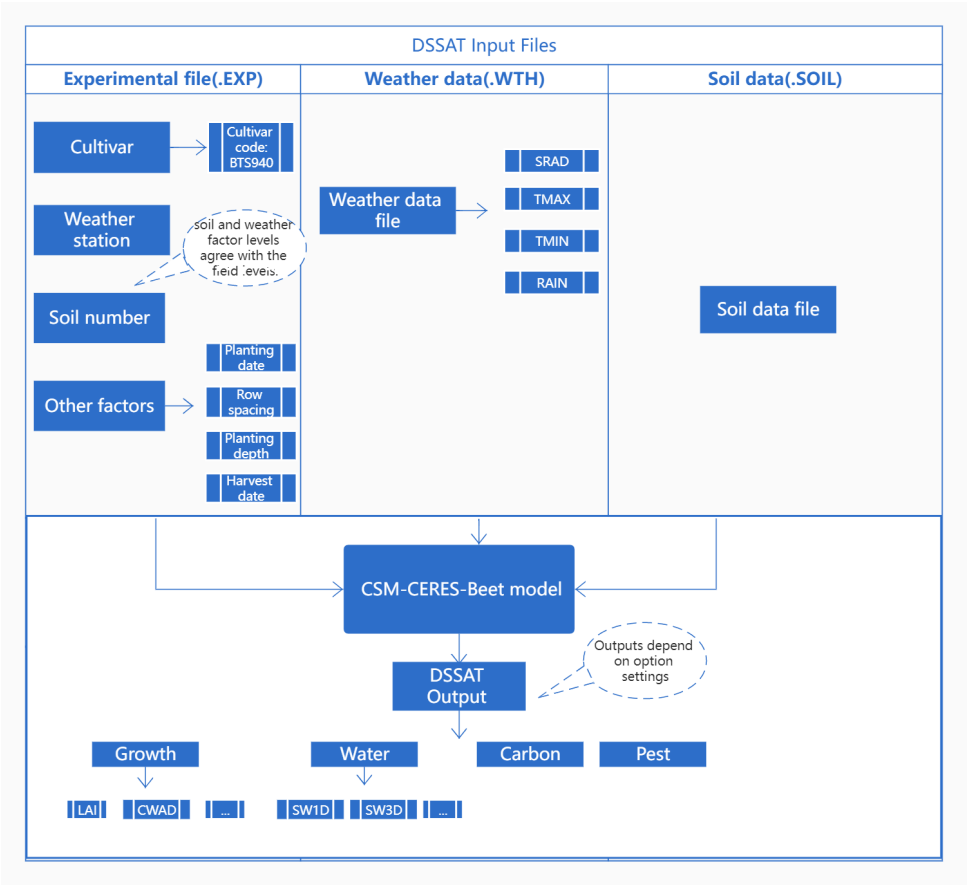


Figure 3.9: Overview of inputs and outputs used by sugarbeet crop models.

- 1 planting factor level.
- 99 field factor levels
- 1 cultivar factor level
- 1 harvest factor level

The developed structure of treatments is shown in Table3.5

level	FieldID	Soil factor level	Weather factor level	Cultivar factor level	Planting factor level	Harvest factor level
1	SBAL1701	1	1	1	1	1
2	SBAL1702	2	2	1	1	1
3	SBAL1703	3	3	1	1	1
4	SBAL1704	4	4	1	1	1
5	SBAL1705	5	5	1	1	1
...	...	...	...	1	1	1
99	SBAL1799	99	99	1	1	1

Table 3.5: The treatment construction in an experiment file in the DSSAT CSM-CERES-Beet model(using an experiment file SBAL1799 as an example).

DSSAT CSM-CERES-BEET MODEL SIMULATION OUTPUTS

The DSSAT CSM-CERES-Beet model produces sugarbeet-specific outputs. Variables correlated with plant growth information and soil water content can be obtained from different output files.

To investigate the soil moisture profile in the root zone of sugarbeet plants, a general understanding of how deep the roots can reach when fully grown is necessary. Normally, the farmers plough into soil for about 22-25(cm) while harvesting the sugarbeet. Accordingly, SW3D(Soil water content of the third layer(15-30cm)) is the most suitable variable for analyzing the soil moisture in the root zone.

DSSAT CSM-CERES-BEET MODEL EVALUATION

In general, temporal LAI is tightly correlated with the evolution of vegetation biomass. Therefore, LAI can be used to evaluate the accuracy of the model simulation. The evaluation dataset comprised estimated LAI from Sentinel-2 Normalized Difference Vegetation Index(NDVI) observations. NDVI is commonly used as a direct vegetation descriptor, since NDVI and LAI display similar seasonal curves with almost equivalent peak times. Model performances were assessed by comparing the simulated LAI data and calculated LAI data from NDVI observations of the sugarbeet.

3.3.2. MACHINE LEARNING MODEL FOR REGRESSION ANALYSIS

The next step is to relate the observed radar signals with the simulated plant and SM variables. Here we use machine learning approach to settle this question. Machine learning approaches can be employed to solve tricky correlations in data sets, especially when the specific shape of the distribution of the data can not be identified properly[28].

For this study, statistical models are not considered. The reason is that for statistical methods, it first needs to select a befitting model referring to some initial assumptions about the correlation between the data sets[28]. We have 4 features in this study, thus it is difficult to propose a model hypothesis by visualizing the data sets. Moreover, what we need is to derive the pattern for radar signals from the vegetation biophysical features instead of getting a specific formula between them. Therefore, the machine learning regression method is suitable to achieve the main goal of this research.

The random forest regression model is a well established approach that has already been used in many studies. Random forest is a supervised learning algorithm that contains multiple independent decision trees. The final regression result is the average prediction of all the trees. This structure can avoid over-fitting due to the subsets and the randomly selected features at each split node. It can handle messy and real datasets and is easy to set up. The random forest model also has the advantage of measuring the feature importance. Thus, it is settled as the machine learning technique for this paper. The regression analysis is going to be developed based on its good efficiency in finding interactions automatically. For the implementation of the random forest we used the scikit-learn Python package, an open-source machine learning library[29].

MODEL INPUT DATA

Using the random forest regression model we aim at establishing a relationship between the SAR observables (the dependent variable) and plant and SM variables (the independent variables), which is one of the main objectives of this work.

Before proceeding with the model, the data needs to be pre-processed. The pre-processing consists of three parts mainly:

Invalid data removal: The time series of DSSAT outputs are continuous from the sugarbeet planting date. However, the temporal resolution of orbit 37 is six days, thus this study uses 28 dates from 1 April 2017 to 16 September 2017 to cover the sugarbeet growth period as much as possible. Moreover, the valid parcels should not only be measured by Sentinel-1 but also can be simulated by the DSSAT model. Before performing the regression model, the removal of invalid data should be done on account of the dates and parcels.

variables flattening: Besides invalid data removal, another essential step of data pre-processing is to flatten all the variables. The data of each variable are assumed to be one column. First, the data are sorted by the IDs of the parcels. Then for each parcel, the data are in time sequence. This way, the data are ready to be split by the model. More importantly, coherence data are generated from interferometric pairs combined by two SAR images with temporal separation,thus it represents a characteristic of difference. Then the prepared X variables, which are going to prescribe coherence, should adapt their structures to develop field temporal difference.

X variables normalization: For random forest regression tasks, the model will be more affected by the high-end values without data normalization[30]. Therefore, the variables normalization is first performed before prediction to avoid misleading the model. Min-max normalization is a linear transformation[31]. The minimum

and maximum values from the original dataset are fetched and each value is recalculated by using the formula below:

$$x' = \frac{x - \min(X)}{\max(X) - \min(X)} \tag{3.1}$$

where X is the original X variables dataset, max(X) and min(X) are the maximum and minimum values of X respectively, and x' and x are the new and old values of each data respectively.

VARIABLES SELECTION

To obtain better prediction results, instead of using all the DSSAT simulation outputs directly, it is necessary to delineate dominant simulated variables prior to training a random forest model[10].

The two used vegetation variables are LAI and CWAD. Significant correspondence exists between these two parameters and canopy structure, biomass, yield and water and carbon balance. LAI is a portion of the leaf surface area per unit of ground surface, defined as

$$\text{LAI} = \frac{\text{leaf area}}{\text{ground area}} \tag{3.2}$$

and has been applied as an efficient indicator of vegetation health and nutrition states[32]. CWAD represents above-ground biomass with regard to intercepted solar radiation and radiation-use-efficiency (RUE)[33], therefore impacting crop growth and yield. The forecast of CWAD is based on intercepted shortwave radiation, temperature, and the amounts of N uptake[34].

The two used soil water content variables are SW1D and SW3D. The shallow surface soil moisture (to 5 cm) pattern can reflect the surface energy balance between soil, vegetation and atmosphere. However, the plant root-zone soil moisture content should also be exploited to estimate the surface evaporation processes and groundwater recharge, since it manages plant transpiration[35].

Finally, these four variables are first defined as the regression model features.

MODEL OPTIMIZATION APPROACH

Different combinations of hyperparameters can control the learning behavior of the model, and afterwards, bring out significantly different results. Therefore, optimizing hyperparameters for random forest models is a crucial step in making more accurate predictions.

Cross validation(CV) should be assimilated to avoid overfitting and test data leakage. The existing approach called k-fold CV splits the training dataset into k smaller folds. $K-1$ folds are taken as training dataset to fit the model, then the remaining fold, considered as test dataset, takes the responsibility for evaluating the model (see Fig 3.10).

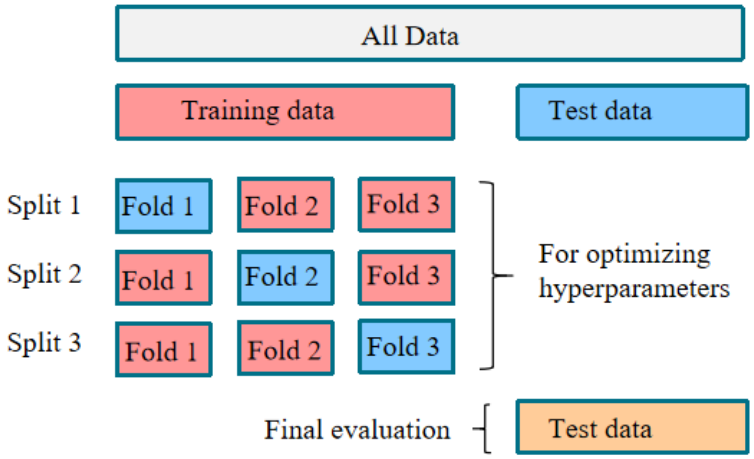


Figure 3.10: The principles of k-fold cross-validation(k=3)[36].

Subsequently, different hyperparameter combinations of the random forest model were evaluated, premised on the dictionary of the candidate hyperparameter values. The following four hyperparameters are adjusted:

1. *N_estimators* .
   The number of decision trees in the forest.
2. *Min_samples_split*.
   The minimum number of samples required for splitting. The split fails if the number of samples in the node is less than min_sample_split .
3. *Min_samples_leaf*.
   The minimum number of samples demanded to be a leaf node. If one of the leaf nodes contains less than min_samples_leaf samples, the corresponding split would be abandoned[37].
4. *Max_depth*.
   The maximum depth of the tree.

After defining the optimal set of hyperparameters, the performance of the random forest model will be maximized .

ACCURACY ESTIMATION APPROACH

Random forest regression model is practised to quantify the connection between predictor variables and response variable. The predictive capability of the regression model can be assessed in various ways without a standard methodology[38]. Here, three metrics are used to evaluate the degree of fitness between the test dataset of the predictor variables and the predicted dataset of the response variable:

1. Mean squared error(MSE) .
   The mean squared error (MSE) between predicted values and actual values in the test dataset. The lower the MSE, the better the regression model fits the dataset. Here is the formula:

$$\text{MSE} = \frac{\sum_i (y_i - f_i)^2}{n} \tag{3.3}$$

   where y_i is the observed value and f_i is the predicted value.

2. R-squared(R^2).
   The proportion of the variance in the response variable that is predictable from the predictor variables[39]. This coefficient normally ranges from 0 to 1. The higher the value, the better the model predictions fit the data. For example, if R^2 = 0.85, this means the predictor variables can explain 85% of the response variable's variance. The formula of the R-squared is defined as

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \overline{y})^2} \tag{3.4}$$

   in which y_i represents the observed value, $\bar{y}$ is the mean of the observed data and f_i is the predicted value.

3. Out-of-Bag(OOB) Score.
   Besides the one main training data set, each tree contains a set of sub-samples that are made by randomly selecting data points from the main set with replacement[40]. This means after each selection, the selected data points are put back into the main set. Thus some data points are not selected in this sampling process, which is formed as out-of-bag sample[41].

   OOB score, calculated as the proportion of accurately predicted rows from the out-of-bag sample, is used particularly for the Random Forest model[42].

FEATURE IMPORTANCE

The trained regression model should not only be accurate but also interpretable. Besides knowing the prediction results, we are also interested in which variables have the most significant impact on the forecast. Identifying which variables are important to the tree decisions can help us in feature selection. Thus, the model can be simplified by concentrating on the most relevant variables[43].

There are multiple methods to determine the feature importance for random forest regression models. Here we listed two kinds of methods[44].

- Random forest built-in method

  Accuracy-based importance One method is based on the mean decrease of accuracy across all trees. As we discussed in section 3.3.2, each tree has its own out-of-bag sample that was not selected during the sampling process. Firstly, we use this kind of sample to calculate the prediction accuracy(the OOB score). Then, we permute the values of the feature in the out-of-bag sample to measure the increase or decrease in prediction accuracy. Finally, the mean change of the prediction accuracy across all trees is calculated[45].

Gini-based importance Another method is based on the mean decrease of impurity across all trees. Figure 3.11 shows the structure of a set of decision trees in a random forest. At each split in each tree, the split criterion is regarding the calculation of how much impurity of the node each feature can reduce. The feature with the highest reduction value is selected as the splitting feature[44]. And this reduction is accumulated across all the trees in the forest independently for each variable. The higher the value, the more important the corresponding predictor is.

- Permutation-based method
  This method permutes the values of a feature then calculate the change of the prediction accuracy. The larger the change, the more importance the feature.

The accuracy-based method is not available in scikit-learn package. However, the permutation-based method uses a similar way to measure the feature importance as the accuracy-based method. Therefore, we generally compared the measured feature importance results between the Gini-based method and the permutation-based method.

By applying these two methods, Fig 3.12 shows an example of the measured feature importance of VH backscatter. The results of these two methods show a vague parallel. Due to the theory behind the Gini-based method, the Gini-based importance is already derived during the training process. Thus, the Gini-based method performs in a less computationally expensive way[44]. Moreover, the appearance of the correlated features is also a potential issue when measuring the feature importance. The study of Nicodemus and Shugart[46] showed that the permutation-based method is less reliable when measuring the importance of correlated features. Cause if the information is not only carried by one feature but also by its correlated features, then the accuracy of the model will not change significantly when one of these features is permuted.

This study uses the Gini-based importance method to measure the feature importance.

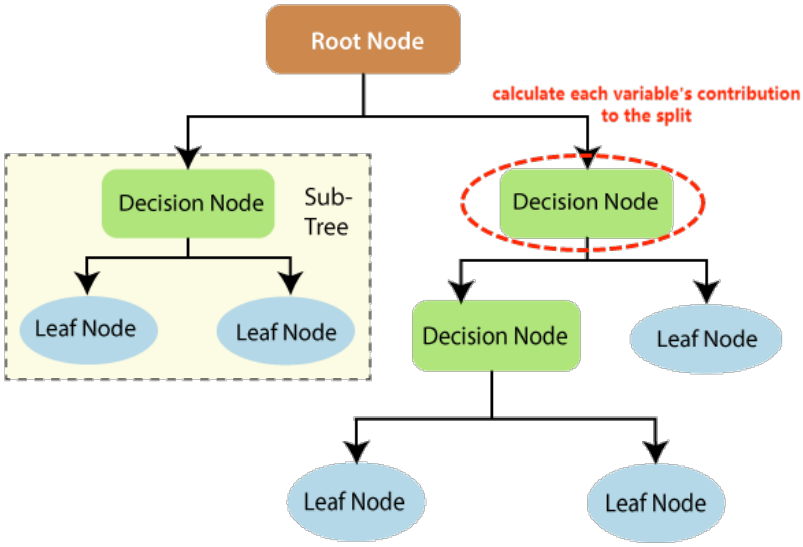


Figure 3.11: The structure of the decision tree in the random forest.

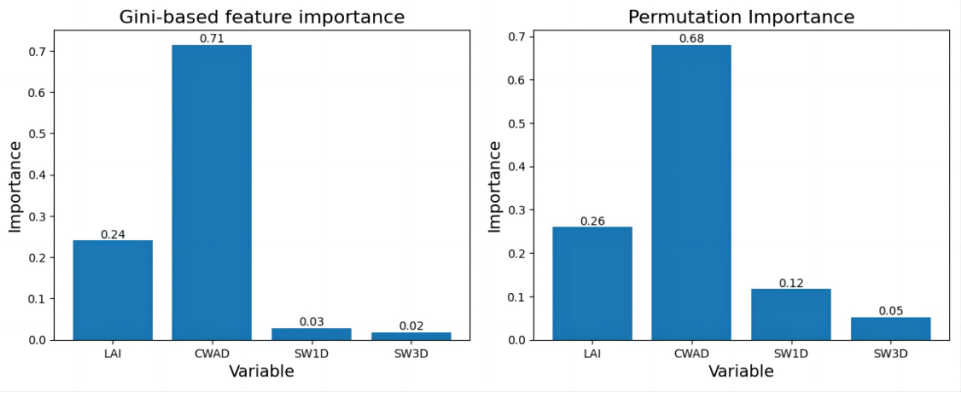


Figure 3.12: The comparison of measured feature importance results between the Gini-based method and permutation-based method for VH backscatter.

4

Results and Discussion

In this chapter, both the results from the DSSAT model as well as the random forest regression analysis will be discussed. In section 4.1 a description of the outputs acquired with the DSSAT crop model is presented. In section 4.2, some characteristic properties of the regression results are discussed.

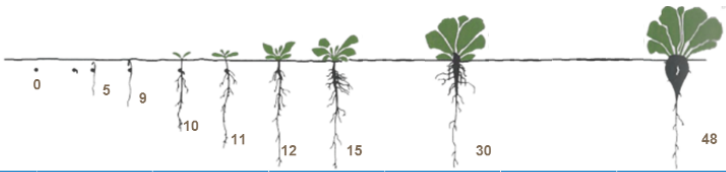
4.1. Evaluation of the DSSAT model

This section discusses the DSSAT CSM-CERES-Beet model outputs. To evaluate the performance of the model, the simulated LAI and the measured LAI are compared.

Figure 4.2 shows the LAI, CWAD, SW1D, and SW3D predicted by the DSSAT CSM-CERES-Beet model for the fields considered in the period of sugarbeet growth in 2017. The simulations of LAI and CWAD start from the planting date, while the soil moisture content variables are from the beginning of the year. The modeled development of the crop is largely controlled by weather variables. The Fig 4.3 shows the mean precipitation and the mean maximum temperature across the study area. And the Fig 4.1 shows a detailed description of the sugarbeet growth stages. Both of these two plots can assist in the interpretation of DSSAT outputs.

For LAI, an increase is observed from mid-May, reaching its maximum value around mid-July. At first, the increase is slow with two or three leaves emerging every week since the low surrounding temperature[47]. In this early growth stage, the photosynthetic efficiency is low. Then after about three weeks, at the beginning of June, the growth rate rapidly increases due to more true leaves having been developed. The peak of LAI represents an effect for densely vegetated areas at which point it reached saturation.

CWAD shows a simultaneous increase with LAI but continues to increase until the end of August due to the continuous dry biomass production. At first, it shows a rapid increase due to most of the produced dry matter from photosynthesis being distributed to fulfill the demand of developing new leaves and canopy. Then the growth rate decreases from mid-June due to the canopy closure[47]. From now, more and more dry matter is going to be stored in the root. We can also use the LAI plot to identify this



Growth Stage	Description of stage	Growth Stage	Description of stage	Growth Stage	Description of stage	Growth Stage	Description of stage
Germination		Leaf Development (Youth stage)		Principal Growth Stage 3		Development of harvestable vegetative plant parts	
00	Dry Seed	10	First leaf visible (pinhead-size): cotyledons horizontally unfolded	31	Beginning of crop cover: leaves cover 10% of ground	49	Beet root has reached harvestable size
01	Beginning of imbibition: seeds begins to take up water	11	First pair of leaves visible, not yet unfolded (pea-size)	33	Leaves cover 30% of ground		
03	Seed imbition complete (pellet cracked)	12	2 leaves (first pair of leaves) unfolded	39	Beet root has reached harvestable size		
05	Radicle emerged from seed (pellet)	14	4 leaves (2nd pair of leaves) unfolded				
07	Shoot emerged from seed (pellet)	15	5 leaves unfolded				
09	Emergence: shoot emerges through soil surface	19	9 and more leaves unfolded				

Figure 4.1: The detailed description of the development of sugar beet[47].

phenomenon, which is corresponding to LAI = 3. After staying stable for about three weeks, this above-ground biomass starts to decrease due to the wilting of old leaves.

For SW1D and SW3D, the values are constantly high before the planting date. At this time, the fluctuations in both SW1D and SW3D variables are dominated by the swift change of bare soil properties while sugarbeet fraction cover is still low. Then, the time series shows a decreasing trend until July because of the warm and dry weather conditions(Fig 4.3). Later, during the summer season, more rain events lead to overall increases in soil moisture content. In addition, the surface soil moisture curve shows more fluctuations, which are closely linked to rainfall events, than that of the root zone layer. Both SW1D and SW3D values were found to be highly sensitive to weather conditions. The soil moisture parameters are sensitive to a systematic mechanism: canopy interception and soil infiltration[35], which is considered in DSSAT, thus the particular behavior of soil moisture is captured by the model.

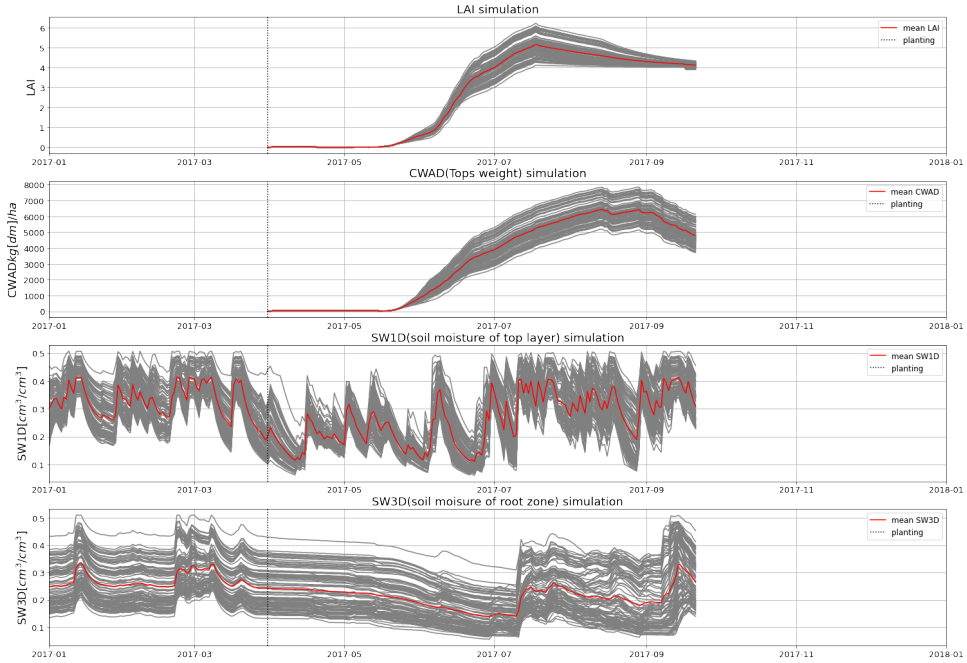


Figure 4.2: Model-simulated values of (a) leaf area index (LAI), (b) CWAD(Tops weight), (c) SW1D(soil water content of top layer), and (d) SW3D(soil water content of the third layer) for sugarbeet in Noord-Brabant 2017. Notes: Single simulated values are plotted in grey lines, and the averaged values across all parcels are in red lines.

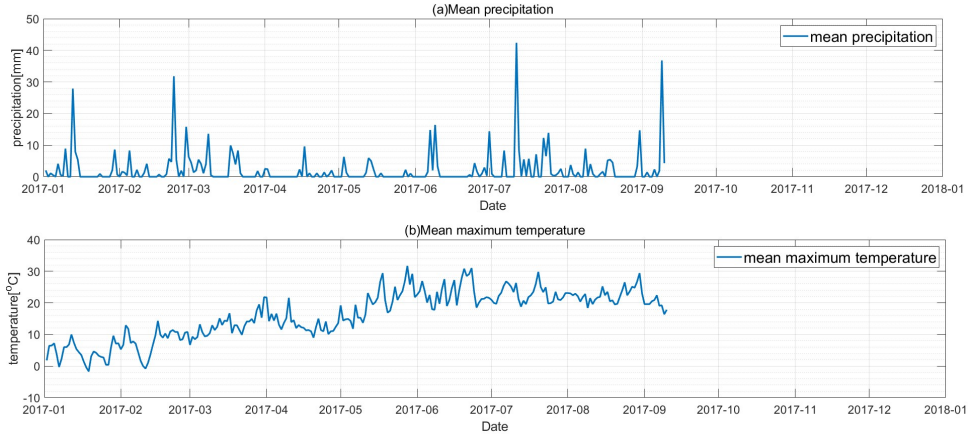


Figure 4.3: (a)Daily mean precipitation data (b)daily mean maximum temperature data across sugarbeet parcels in Noord-Brabant from the beginning of 2017 till the sugarbeet harvest.

Figure 4.4 shows the comparison between the simulated LAI values and the

Sentinel-2 observed LAI values. The general trend of both curves is similar. These two curves reached their peak LAI values almost at the same time and same magnitude, which can fit the sugarbeet phenology stages well. Therefore, DSSAT performs realistic simulations of the sugarbeet growth in the study area.

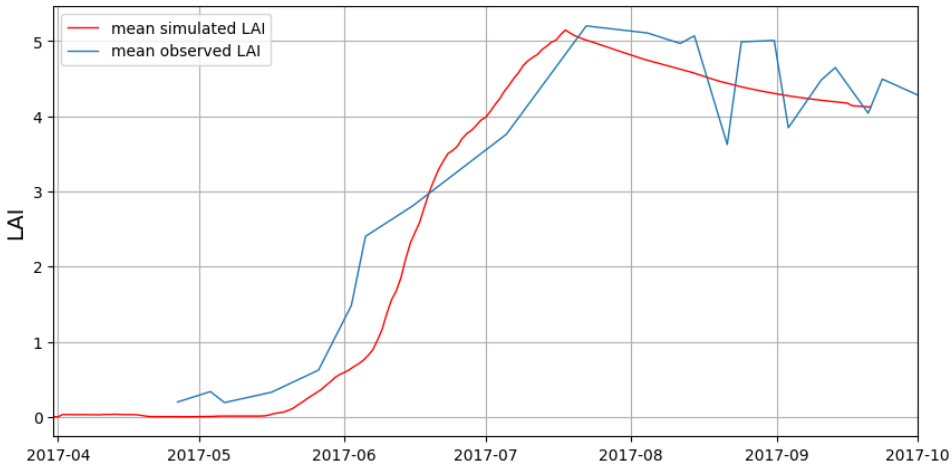


Figure 4.4: LAI comparison. The red line represents LAI values from the DSSAT simulation, and the blue line is from Sentinel-2 LAI calculation .

4.2. EVALUATION OF THE MACHINE LEARNING MODEL

In this section, the correlation between SAR signals and vegetation biophysical coefficients are going to be discussed.

4.2.1. THE FEATURE IMPORTANCE

Here, Fig 4.5 provides the partial dependence plots of the variable importance for backscatter VV and VH and coherence VV.

In Fig 4.5, both polarizations of the C-band σ^0 exhibit a strong correlation with CWAD. Moreover, LAI comes to the second most important factor, especially for σ^0_{VH}. This distribution of variable importance can be explained by the fact that σ^0_{VH} is more related to vegetation elements than σ^0_{VV}. However, lower variable importances with σ_0 are reported at soil moisture features, regardless of polarization and soil layer.

Figure 4.6 indicates that for VV coherence, LAI made the most contribution to the final prediction, while all the rest variables have a similar level of importance.

The main difference between the variable importance of backscatter and coherence is that LAI is much more important for coherence, whereas CWAD dominates the backscatter. Figure 4.2 shows that the temporal trend of LAI is more correlated with the crop growth steps than CWAD. And the coherence data is highly sensitive to the vegetation phenology. As a result, LAI becomes the top important variable when predicting coherence. However, the LAI and CWAD of crops are correlated, thus there is indeed a lot of information in common between the two predictors. The correlation

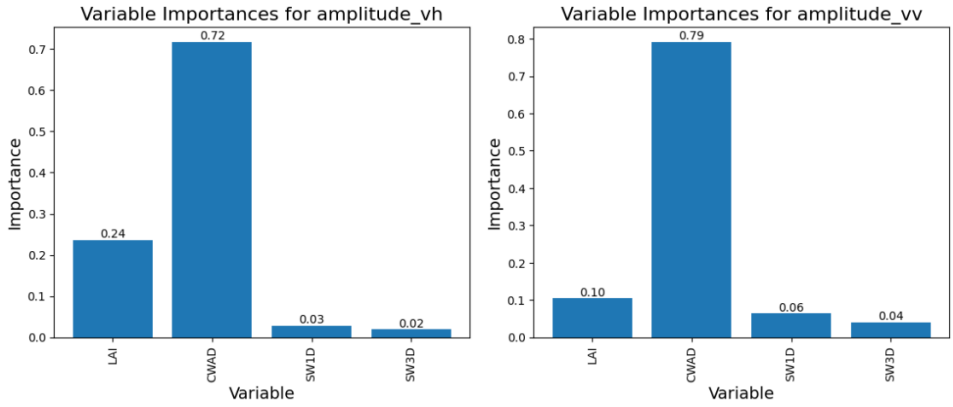


Figure 4.5: Visualization of the variable importance for backscatter VH and VV of the random forest regression model.

between them will be discussed in the later section 4.2.3.

4.2.2. ACCURACY OF THE MACHINE LEARNING MODEL

Figure 4.7 and 4.8 provide a comparison of the radar observables predicted by the random forest model and the values retrieved from Sentinel-1 acquisitions. To help the visual interpretation, the time series of backscattering coefficients and coherence are interpolated to the dates for better continuity.

The variable importance in a random forest regression model is measured by how much the variable decreases the error, which is valued as the variance of the difference between the measured and predicted values. Therefore, based on the variable importance test, besides the original 'full' model which contains all four features, a 'reduced' model which contains only LAI and CWAD, the two most important features, was also developed to examine the changes in the model accuracy metrics(Table 4.1).

In general, the seasonal evolution of the backscattering coefficients data is well predicted. Obviously, the correlations are scattered, more poorly in some time stages. April is inside the period of sugarbeet sowing, thus the standard deviation of the original backscatter data for both VV and VH polarizations is high. This large variance makes it harder to predict accurately.

The green curves in the figures represent the difference values, which are the absolute difference value between the means of predicted and actual values. The difference peaks at around mid-May which is induced by a sudden appearance of leaves. In addition, during April and May, the sugarbeet is not fully covering the soil, so the vegetation and soil contributions are combined, which are the factors that creates uncertainties. There are also some fluctuations during July, which can be explained by some abrupt summer precipitation(Fig 4.3).

When compared (a)(b) plots in Fig 4.7, σ^0_{VH} generally shows a better correlation than σ^0_{VV}. In fact, sugarbeet with taller plants, along with randomly oriented stems, causes higher volume scattering profiles and lower attenuation of the signal from the ground.

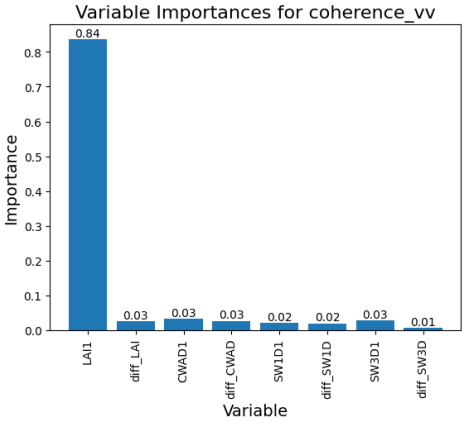


Figure 4.6: Visualization of the variable importance for coherence VV of the random forest regression model.

And the density of sugarbeet stems is high, which also leads to a weak surface scattering from the soil. Consequently, backscatter VH is more correlated with the selected features over the sugarbeet-covered area.

Then an inspection of (c)(d) plots in Fig 4.7 reveals that for the regression model using LAI and CWAD out of all four vegetation-related features, backscattering coefficients' features present different correlations. This shift can be reflected by a slightly larger range in difference values between mean prediction data and actual test data. As listed in Table 4.1, the variation between 'full' and 'reduced' regression models is also statistically significant.

The reason for this is that the model variance is driven by variable selection. As calculated previously, CWAD and LAI with higher variable importance lead to bigger information gains. But SW1D and SW3D are also information carriers, especially when the vegetation cover is not that dense. Thus the decorrelation, which is attributable to the impact of the soil moisture, can be observed in April and May when the soil is almost bare after first sowing. Moreover, the CERES-Beet model defined that the germination process which happened in this period is a function of soil moisture content[48]. Therefore, without the information from soil moisture features, the predicted courses even perform in a flat way without fluctuations. As a result, the correlation between observations and prediction deteriorates. Then when sugarbeet grows, a lower sensitivity of the σ^0 to soil contribution resulted, thus the prediction shows a better-fit behavior.

In a process analogous to the σ^0 analysis, the regression results for coherence VV channel have been built. The Fig 4.8 shows the prediction results. As for the difference curves analysis, some fluctuations that occur during April and May coincide with the increasing temporal decorrelation due to the induced movement caused by quick-growing sugarbeet. The loss of model performance is associated with sugarbeet growth in this period.

Interestingly, the 'reduced' model for coherence VV also performs well. Indeed, it

provides comparable accuracy to the 'full' model. Thus if only LAI and CWAD features are considered, results are similar to those obtained from the 'full' model. This means that the main influential issue to the coherence VV signal is related to the development of the sugarbeet plants in this field and the soil contribution becomes marginal. Therefore, the variation of sugarbeet growth phenology conditions would affect the performance of the coherence VV substantially.

Table 4.1 shows the statistical accuracy metrics of the predictions. Both the 'full' and 'reduced' models of σ^0_{VH} prediction yielded the maximum agreement index(R^2 and OOB_score) about $0.84-0.87$ with the RMSE of $2-2.5dB$. The RMSE is relatively low when compared to the variable scale. Relatively low correlation coefficients were observed at the σ^0_{VV} forecasts. The R^2 and OOB_score of the σ^0_{VV} forecasts were about $0.65-0.70$ with an RMSE of $2.3-2.7dB$. These varied results suggest that σ^0_{VH} has a higher sensitivity than σ^0_{VV} to vegetation-covered surfaces. Moreover, the vegetation biophysical variables also showed significant forecast skill for the coherence VV prediction. The coherence prediction also had high agreement indexes and low RMSE. And the difference between its 'full' and 'reduced' models is the lowest among these three SAR signals.

Furthermore, according to Table 4.1, the accuracy metrics values decrease by approximately 3.5% and 7.2% for backscatter VH and VV channels, respectively. This suggests that the VV channel has been affected more. After running the model several times, this phenomenon can be primarily explained by the fact that the total variable importance of LAI and CWAD for the backscatter VH channel is stable at around 0.94-0.96, which is larger than that of the VV channel(0.87-0.89). VV polarization achieves a higher sensitivity to soil contribution (soil moisture in this research). Therefore, both backscattering coefficients are obviously affected by the feature selection modification, and the impact on VV is even greater.

4.2.3. THE ISSUE ABOUT THE CORRELATED PREDICTORS

The feature selection can be harder when the predictors are highly correlated. Archer and Kimes[49] have found that both the Gini-based and permutation-based methods become less able to determine the most important features when the correlation between the features increases. Our predictors are four biophysical variables of sugarbeet and they can be interrelated. For example, tops weight(CWAD) is about the above-ground biomass, associated with the sugarbeet cover. Therefore, LAI and CWAD can be largely relevant to each other. This section provides the evaluation of the correlation between the features and their influence on the final prediction results.

We use Agglomerative Hierarchical Clustering on Spearman rank correlation to group our features into clusters based on hierarchy[50].

The Spearman rank correlation represents the rank correlation between two variables[51]. Eq 4.1[52] shows the full version of the Spearman correlation formula. The coefficient ρ ranges from -1 to 1, in which +1 and -1 represent the perfect positive and negative correlation between the variables, respectively, and 0 means variables are uncorrelated.

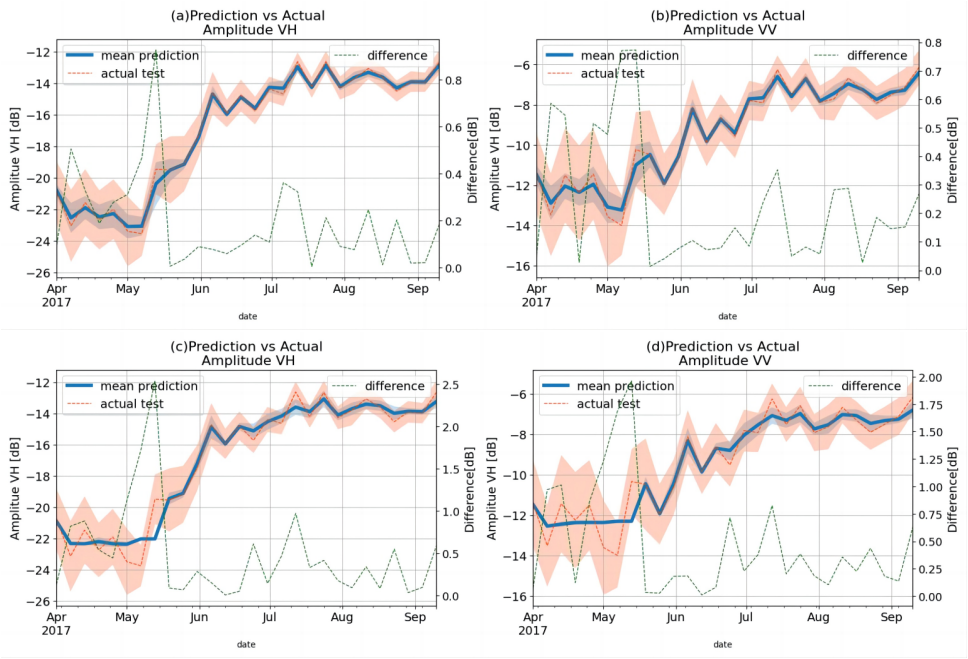


Figure 4.7: Comparison of SAR backscattering coefficients between random forest simulated and Sentinel-1 observations.(a) and (b) are the results by using all four variables while the bottom plot (c) and (d) by only using LAI and CWAD. Mean prediction and actual values are represented by solid blue lines and orange dashed lines, respectively. Standard deviations are shown by the filled areas surrounding the curves. The time series of differences between prediction and test values are plotted in green dashed lines. The later coherence plots can also refer to this interpretation.

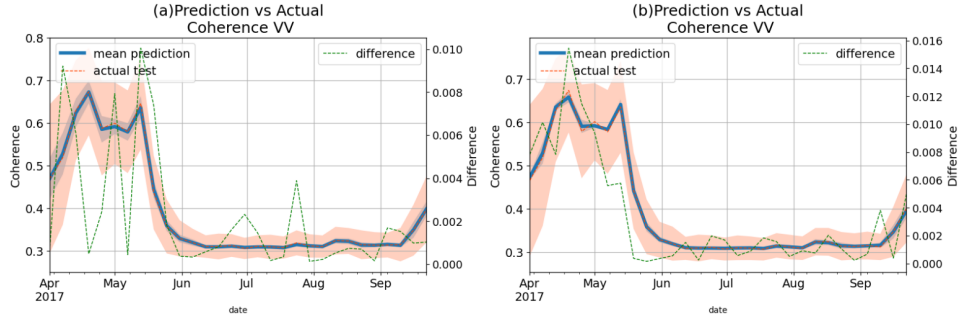


Figure 4.8: Comparison of SAR coherence VV signal between random forest simulated and Sentinel-1 observations. (a) is the results by using all four variables while the left plot (b) Only uses LAI and CWAD variables.

$$\begin{aligned}\rho &= \frac{S_{xy}}{S_x S_y} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}(R(x_i)-R(\bar{x}))\times(R(y_i)-R(\bar{y}))}{\sqrt{(\frac{1}{n}\sum_{i=1}^{n}(R(x_i)-R(\bar{x}))^2)\times(\frac{1}{n}\sum_{i=1}^{n}(R(y_i)-R(\bar{y}))^2)}}\end{aligned} \tag{4.1}$$

With:

- ρ: the Spearman correlation coefficient
- x_i, y_i: the original raw data of the two variables, respectively;
- $R(x_i)$, $R(y_i)$: the ranks of the original raw data;
- $R(\bar{x})$, $R(\bar{y})$: the mean ranks of each variable;

In the clustering process, we first develop a cluster for each feature, then consecutively merge the most similar clusters until only one cluster is left. The method used to measure the similarity between the clusters is Ward's Method. Instead of calculating the direct distance between the clusters, Ward's Method is about minimizing the variance when combining new clusters[53]. This kind of variance is quantified by a metric called E (sum of squares)[53]. Eq 4.2[50] shows a complete statistical description of E.

$$\begin{aligned}\triangle(A,B) &= \sum_{i\in A\cup B} \parallel \vec{x}_i - \vec{m_{A\cup B}} \parallel^2 - \sum_{i\in A} \parallel \vec{x}_i - \vec{m}_A \parallel^2 - \sum_{i\in B} \parallel \vec{x}_i - \vec{m}_B \parallel^2 \\ &= \frac{n_A n_B}{n_A + n_B} \parallel \vec{m}_A - \vec{m}_B \parallel^2\end{aligned} \tag{4.2}$$

With:

- A, B: two clusters that are going to merge;
- m_j: the center of the cluster j;
- x_i: every data point in the defined data set;
- n_j: the number of data points in the cluster;

We applied clustering to the features of our study. Figure 4.9 shows the result of Agglomerative Hierarchical Clustering which is represented by a dendrogram. The y-axis indicates the variance cost when merging the two clusters. It can be observed that the correlation between LAI and CWAD is the highest. SW1D is closer to the combined cluster of LAI and CWAD than SW3D. SW3D shows the weakest connection to the other features.

Figure 4.10 is a heatmap showing the correlated features. The data value in each cell is the calculated Spearman correlation coefficient ρ. It can be observed that for CWAD and LAI, the value is 0.92, very close to 1, indicating that the correlation between CWAD and LAI is very high, and the positive value suggests that the larger the CWAD the higher the LAI. The coefficient of SW3D with CWAD and LAI is -0.0054 and -0.046, respectively.

The values are close to zero showing that the correlations between SW3D and these two features are very low, and the negative values suggest that the SW3D tends to decrease when CWAD and LAI increase. Moreover, there is a positive correlation with a more moderate degree between SW1D and the other three features.

The relationship between features can be explained by the calculation mechanism of the CERES-Beet model. CWAD is related to the above-ground dry matter yield. Firstly, the daily total amount of sugarbeet dry matter is converted from the intercepted photosynthetically active radiation (PAR), which is calculated as a function of LAI[48]. Then depending on the growing stages of the sugarbeet, the generated dry matter will be allocated to different parts of the plant. The above-ground dry matter demand is measured by the potential leaf area growth, which is also highly associated with LAI. Based on these simulation structures, LAI and CWAD are the most relevant among all features. Then the soil moisture content is a part of the factors that can influence the efficiency of dry matter production in photosynthesis and also the subsequent allocation issues. Moreover, the surface soil moisture is more relevant with CWAD and LAI since the surface layer contains more coarse roots which can improve the efficiency of plant growth using soil water[54].

Since the estimation of correlation between features shows that CWAD and LAI are highly correlated, we ran the random forest regression model again with the removal of features. Based on the measured feature importance results in section4.2.1, we removed CWAD and LAI for backscatter and coherence regression models, respectively. In order to validate the changes in the prediction accuracy when the most important feature is dropped but the second most important variable is still there, and is highly correlated with the most important one.

Table 4.2 shows the statistical accuracy metrics of the regression model with correlated feature removal. It can be clearly observed that when the model contained only one predictor of CWAD or LAI, the accuracy of the predictions did not decrease significantly. This indicates these two predictors do carry a lot of similar information.

SAR signals	X variables	MSE	R2_score	OOB_score
Amplitude VH	All variables	1.966	0.876	0.875
	LAI&CWAD	2.459	0.845	0.844
Amplitude VV	All variables	2.305	0.698	0.694
	LAI&CWAD	2.661	0.648	0.640
Coherence VV	All variables	0.0054	0.734	0.739
	LAI&CWAD	0.0052	0.734	0.737

Table 4.1: The accuracy of the random forest regression model with different combinations of SAR signals and sugarbeet growth variables.

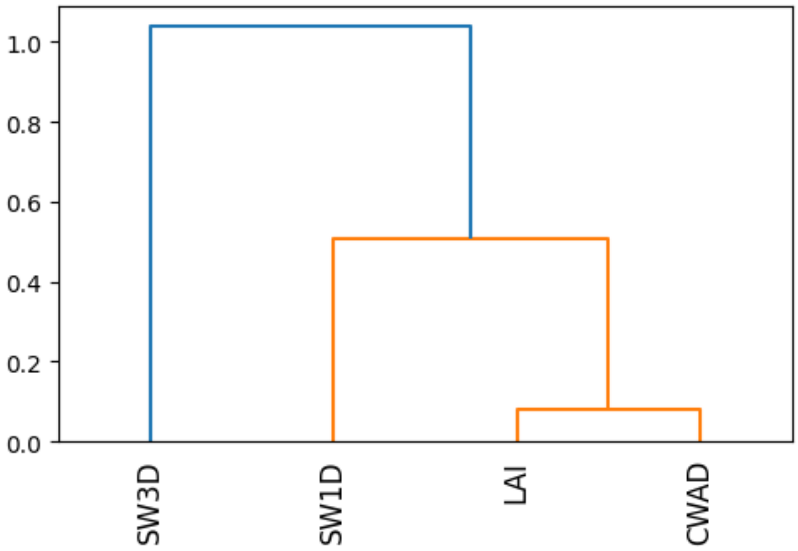


Figure 4.9: The dendrogram by performing the hierarchical clustering method.

SAR signals	X variables	MSE	R2_score	OOB_score
Amplitude VH	All variables	1.966	0.876	0.875
	without CWAD	2.077	0.867	0.871
Amplitude VV	All variables	2.305	0.698	0.694
	without CWAD	2.422	0.682	0.688
Coherence VV	All variables	0.0054	0.734	0.739
	without LAI	0.0052	0.734	0.738

Table 4.2: The accuracy of the random forest regression model with highly correlated features removal.

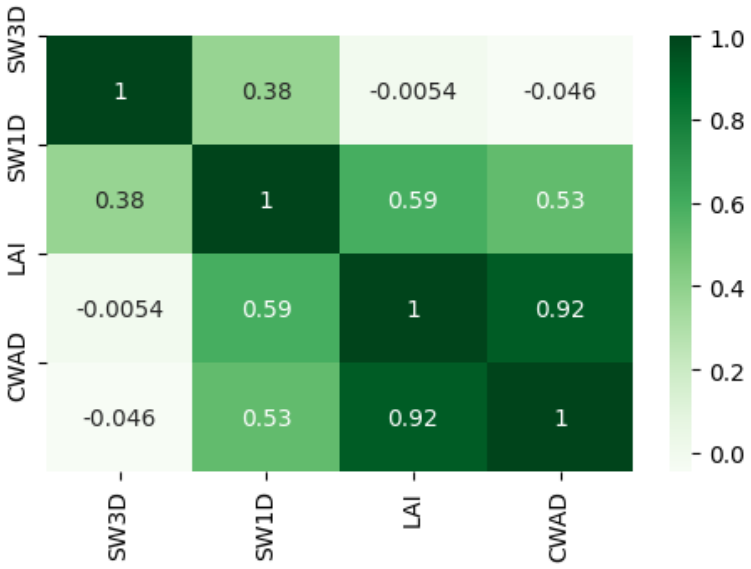


Figure 4.10: The heatmap of the correlation between features, and the data value in each cell is the result of Spearman correlations .

5

Conclusions and Recommendations

5.1. Conclusions

In this research, the relationships between Sentinel-1 C-band SAR signals and several DSSAT simulated sugarbeet growth and soil moisture descriptors have been explored through regression analysis. LAI, CWAD, SW1D, and SW3D are picked as the features to help predict the backscattering coefficients(VH and VV) and coherence VV over the sugarbeet-covered region over time. The changes in predicted SAR patterns due to various feature combinations are also analyzed. Answers are organized in corresponding to the research questions proposed in section1.2.

Evaluating DSSAT CSM-CERES-Beet model performance

- **What is the CSM-CERES-Beet model accuracy of simulating the sugarbeet growing process in the Netherlands?** The DSSAT crop model aims to capture how plant biophysical variables vary with the growth stages. We evaluated the simulated LAI values analytically with the estimated LAI values from NDVI observation. As we expected, both simulated and observed LAI values are between 0-5 over the sugarbeet-covered area. And the measured LAI temporal patterns resemble the modeled LAI patterns simulated by DSSAT. The performance of the CSM-CERES-Beet model is sufficiently robust.

Evaluating the correlation performance.

- **How to increase the accuracy of the predictions made by random forest model?** Hyperparameters are key settings controlling the performance of the random forest model fitness. The hyperparameters including the number of trees, the minimum number of samples to split the nodes, the minimum number of samples to be a leaf node, and the maximum depth of the tree can be optimized to fine-tune the random forest algorithm. An example of optimal combinations is listed in Table 5.1,

SAR signals	X variables	n _estimator	min_samples _split	min_samples _leaf	max _depth
σ^0_{VH}	All variables	137	20	24	93
	LAI&CWAD	162	12	17	24
σ^0_{VV}	All variables	189	7	25	71
	LAI&CWAD	181	3	28	69
Coherence VV	All variables	174	32	5	16
	LAI&CWAD	163	30	19	12

Table 5.1: The optimal hyperparameters for a random forest regression model with different combinations of SAR signals and sugarbeet growth variables.

- **How well does crop biophysical variables correlate with the backscattering coefficient(VH and VV)?** The performance is evaluated by statistical parameters. Backscattering coefficient profiles are in good agreement with the simulated vegetation and soil features. The high R^2 values around 0.85-0.87 and 0.65-0.70 for VH and VV channels, respectively, indicate that the predictors actually predict the observed values. Both two polarization channels of the backscatter keep a clear sensitivity to crop growth.

- **How well does crop biophysical variables correlate with coherence VV?** A good model fit was achieved for coherence VV, with coefficients of determination (R2) reaching around 0.73, and a low RMSE here suggests that the residuals are tight around 0. Coherence is sensitive to canopy development and soil preparation.

- **How does feature selection affect the prediction accuracy of SAR signals?** The top two features with the highest importance for the backscattering coefficients are LAI and CWAD, while for coherence LAI comes to the top. However, the 'reduced' model with only the most important features did not gain a better prediction accuracy. By contrast, backscatter data is widely linked to changes in soil moisture during the first growth stages when vegetation coverage remains moderate. The 'reduced' edition model indicates that it is not adequate to express the interaction of SAR signals with a complicated vegetation-over-soil field by only using crop growth variables. Although LAI and CWAD are considered as the two main prediction drivers for σ^0 patterns, the weaker contributions from soil moisture variables are by no means negligible.

5.2. RECOMMENDATIONS

This section provides recommendations and is divided into two sections. Firstly, the discussion about the practical use of C-band radar observables regarding the regression results(section 5.2.1), followed by suggestions for future studies(section 5.2.2).

5.2.1. PRACTICAL USE OF THE RESULTS

The regression results show that both the backscatter and coherence data are closely correlated with the periodic crop growth and soil moisture features. Moreover, the backscatter and coherence data contain complementary information and therefore, the use of both data sources is beneficial in the observation of field dynamics.

This approach facilitates the potential adoption of assimilating C-band radar signals into the improvement of crop model simulations. We can use the radar observables to improve the initial crop model settings and identify some boundary conditions. When the crop model completes the simulations, radar observables can also be employed to estimate the performance of the model and provide some advice on calibrating.

Moreover, this study demonstrates the high sensitivity of radar observables to crop biophysical variables, especially for CWAD. Thus, radar signals can provide opportunities to track and predict these kinds of closely correlated variables directly. Then the farmers and governments can get timely information about the growing stages of crops, and as a result, could help with the fertilization and irrigation decisions. Furthermore, the health conditions of crops can be assessed according to the appearance of anomalies in radar observations. For example, the yield of sugarbeet can be highly influenced by the Cercospora leaf spot disease. And the infection with this disease can be reflected in the daily losses of leaf area and biomass. Radar signals can report such unusual reduction and provide early detection of the disease, then some chemical measures can be taken suitably.

.

5.2.2. SUGGESTIONS FOR FURTHER STUDIES

MORE MACHINE LEARNING METHOD

For this study, the correlations between the vegetation biophysical variables and C-band radar signals are analyzed by random forest and therefore the performances of other machine learning regression methods are not considered. Therefore, the potential of other methods can be further exploited.

Random forest is one of the ensemble learning models whose final prediction is obtained by training multiple machine learning models and using some logic to combine their prediction results[55]. Based on this general design strategy, this kind of model can reduce the influence of outliers more than a single model. Here, we prefer to select from other ensemble learning methods besides the random forest.

There are mainly three classes of ensemble learning methods:

1. Bagging
   Develop a set of same machine learning models, then train each model with a varied sub-sample of the one main training dataset[55]. For regression problems, the final prediction result is the averaged results across all the individual models. Random forest used in this study is a typical bagging method.

2. Boosting
   Figure 5.1 shows the design idea of the boosting method. It uses the same training data set, but the sample points have different weights in the new ensemble members to optimize the performance of prior added ensemble members. Then

the final prediction result is a weighted average of predictions across all the models.

3. Stacking
   Use a uniform training dataset and a set of different machine learning models. Then an additional machine learning model is developed to learn the best way of combining the predictions from various models.

Figure 5.2 shows these three main approaches and their related algorithms. Based on the discussion above, we can try some techniques derived from Boosting and Stacking approaches in the future.

ADDITIONAL CALCULATION OF COHERENCE BIAS

As mentioned in section 3.1.2, the coherence has a bias for low values. In this study, we did a preliminary estimate of the bias based on agricultural experience. However, if one needs to capture the subtle coherence changes associated with plant growth, this bias requires careful measurement. Figure 5.3 shows an example of the biased and unbiased temporal coherence data of sunflower. We can observe that when the value of coherence drops to 0.4, the bias will start to appear and gradually increase as the coherence value decreases.

ADDITIONAL DATASET OF POLARIZATION RATIO VH/VV

Many studies[9][56] indicate that the backscatter polarization ratio VH/VV is a good indicator of the fresh biomass and vegetation water content(VWC), and is in good agreement with the crop phenology.

Moreover, Veloso et al.[56] demonstrate that the ratio VH/VV is able to reduce the double-bounce effect from the soil. And it has been observed that heavy precipitation events have little effect on the change in the ratio. This means that the ratio is less sensitive to soil moisture changes. When compared to VH and VV backscatter, this characteristic makes the ratio a more reliable metric when encountering some crop monitoring cases where soil contributions need to be marginalized.

Consequently, the ratio VH/VV could assist in the future crop biophysical parameters retrieval work.

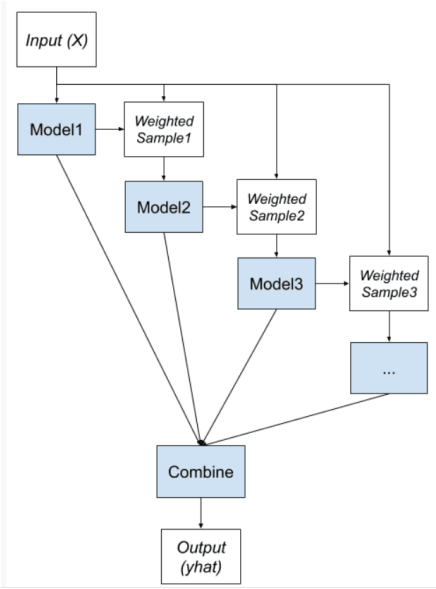


Figure 5.1: The structure of the boosting ensemble learning method[55].

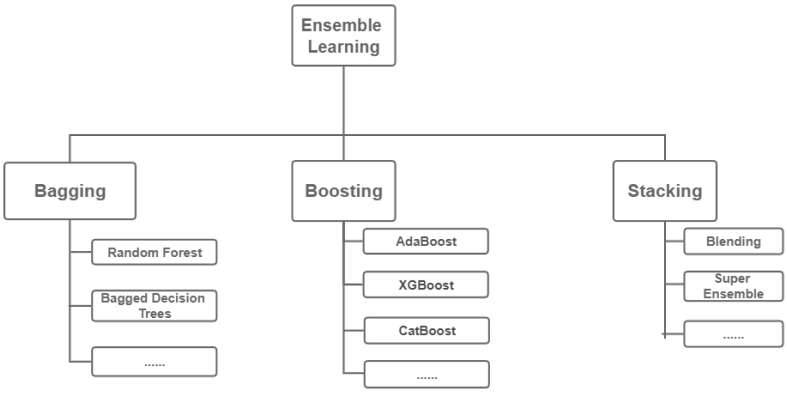


Figure 5.2: The three main ensemble learning types and their related algorithms.

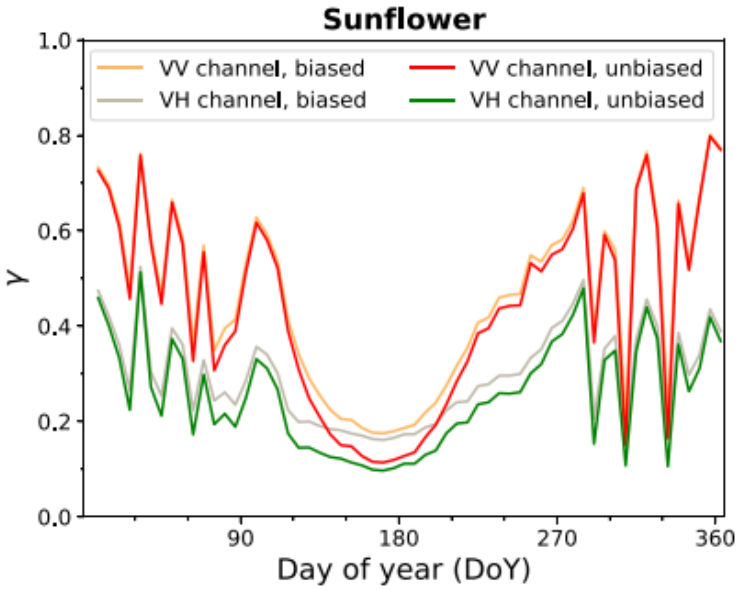


Figure 5.3: Example of the time series for biased and unbiased coherence at both polarimetric channel for sunflower[12].

BIBLIOGRAPHY

[1] N. Ouaadi, L. Jarlan, J. Ezzahar, *et al.*, "Monitoring of wheat crops using the backscattering coefficient and the interferometric coherence derived from sentinel-1 in semi-arid areas," *Remote Sensing of Environment*, vol. 251, p. 112 050, 2020, ISSN: 0034-4257. DOI: https://doi.org/10.1016/j.rse.2020.112050. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S003442572030420X.

[2] G. Hoogenboom, "Decision support system for agrotechnology transfer (dssat) version 4.8," [Online]. Available: https://dssat.net/.

[3] A. Pandit, S. A. Sawant, J. Mohite, and S. Pappula, "Sentinel-1 derived coherence time-series for crop monitoring in indian agriculture region," *Geocarto International*, 2021.

[4] J. Jones, G. Hoogenboom, C. Porter, *et al.*, "The dssat cropping system model," *European Journal of Agronomy*, vol. 18, no. 3, pp. 235–265, 2003, Modelling Cropping Systems: Science, Software and Applications, ISSN: 1161-0301. DOI: https://doi.org/10.1016/S1161-0301(02)00107-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1161030102001077.

[5] G. Uehara and G. Y. Tsuji, "The ibsnat project," 1993.

[6] A. Irmak, J. Jones, W. Batchelor, and J. Paz, "Estimating spatially variable soil properties for application of crop models in precision farming," –, vol. 440570, pp. 1343–1353, Jan. 2001. DOI: 10.13031/2013.6424.

[7] G. Hoogenboom, "Dssat model components," [Online]. Available: https://dssat.net/models-overview/components/.

[8] M. Vreugdenhil, W. Wagner, B. Bauer-Marschallinger, *et al.*, "Sensitivity of sentinel-1 backscatter to vegetation dynamics: An austrian case study," *Remote. Sens.*, vol. 10, p. 1396, 2018.

[9] S. Khabbazan, P. Vermunt, S. Steele-Dunne, *et al.*, "Crop monitoring using sentinel-1 data: A case study from the netherlands," *Remote Sensing*, vol. 11, no. 16, 2019, ISSN: 2072-4292. DOI: 10.3390/rs11161887. [Online]. Available: https://www.mdpi.com/2072-4292/11/16/1887.

[10] K. C. Kornelsen and P. Coulibaly, "Advances in soil moisture retrieval from synthetic aperture radar and hydrological applications," *Journal of Hydrology*, vol. 476, pp. 460–489, 2013, ISSN: 0022-1694. DOI: https://doi.org/10.1016/j.jhydrol.2012.10.044. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022169412009444.

[11] P. G. Minotti, M. Rajngewerc, V. Alí Santoro, and R. Grimson, "Evaluation of sar c-band interferometric coherence time-series for coastal wetland hydropattern mapping," *Journal of South American Earth Sciences*, vol. 106, p. 102 976, 2021, ISSN: 0895-9811. DOI: https://doi.org/10.1016/j.jsames.2020.102976. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0895981120305198.

[12] A. Villarroya-Carpio, J. M. Lopez-Sanchez, and M. E. Engdahl, "Sentinel-1 interferometric coherence as a vegetation index for agriculture," *Remote Sensing of Environment*, vol. 280, p. 113 208, 2022, ISSN: 0034-4257. DOI: https://doi.org/10.1016/j.rse.2022.113208. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425722003169.

[13] X. Blaes and P. Defourny, "Retrieving crop parameters based on tandem ers 1/2 interferometric coherence images," *Remote Sensing of Environment*, vol. 88, no. 4, pp. 374–385, 2003, ISSN: 0034-4257. DOI: https://doi.org/10.1016/j.rse.2003.08.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S003442570300213X.

[14] M. Engdahl, M. Borgeaud, and M. Rast, "The use of ers-1/2 tandem interferometric coherence in the estimation of agricultural crop heights," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 39, pp. 1799–1806, Sep. 2001. DOI: 10.1109/36.942558.

[15] V. Kumar, "Agricultural sandboxnl: A national-scale database of parcel-level processed sentinel-1 sar data," *Scientific Data*, vol. 9, Jul. 2022. DOI: 10.1038/s41597-022-01474-4.

[16] "Dataset: Basisregistratie gewaspercelen (brp)," [Online]. Available: https://www.pdok.nl/introductie/-/article/basisregistratie-gewaspercelen-brp-.

[17] Introductie—PDOK., "Introductie—pdok.," Jan. (2022). [Online]. Available: https://www.pdok.nl/over-pdok.

[18] B. InZicht, "Facts, figures and maps about brabant in a well-arranged manner," Oct. (2017). [Online]. Available: www.brabantinzicht.nl.

[19] klimaatinfo, "The climate of north brabant," [Online]. Available: https://klimaatinfo.nl/klimaat/nederland/noord-brabant/.

[20] T. h.-b. Yamane, "Sugar beet. encyclopedia britannica," Jan. (2022). [Online]. Available: https://www.britannica.com/plant/sugar-beet.

[21] R. N. M. Institute, "Knmi weather stations-day observations," [Online]. Available: https://daggegevens.knmi.nl/.

[22] R. N. M. Institute, "Precipitation - daily precipitation sum in the netherlands," [Online]. Available: https://dataplatform.knmi.nl/dataset/rd1-5.

[23] I. W. S. Information, "Soilgrids — global gridded soil information," [Online]. Available: https://www.isric.org/explore/soilgrids.

[24] "Cultivar," in *Encyclopedia of Genetics, Genomics, Proteomics and Informatics*. Dordrecht: Springer Netherlands, 2008, pp. 449–449, ISBN: 978-1-4020-6754-9. DOI: 10.1007/978-1-4020-6754-9_3921. [Online]. Available: https://doi.org/10.1007/978-1-4020-6754-9_3921.

[25] A. Kobayashi, K. Hori, T. Yamamoto, and M. Yano, "Koshihikari: A premium short-grain rice cultivar – its expansion and breeding in japan," *Rice*, vol. 11, 2018.

[26] E. Memic, S. Graeff, O. Hensel, and W. Batchelor, "Extending the csm-ceres-beet model to simulate impact of observed leaf disease damage on sugar beet yield," *Agronomy*, vol. 10, p. 1930, Dec. 2020. DOI: 10.3390/agronomy10121930.

[27] A. Singels, "Dssat v4.5 - canegro sugarcane plant module: Scientific documentation," Mar. 2008.

[28] W. Teboul, "Why use machine learning instead of traditional statistics?," 2018. [Online]. Available: https://towardsdatascience.com/why-use-machine-learning-instead-of-traditional-statistics-334c2213700a.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[30] J. Fernandez, "How data normalization affects your random forest algorithm," [Online]. Available: https://towardsdatascience.com/how-data-normalization-affects-your-random-forest-algorithm-fbc6753b4ddf.

[31] Varsha, "What is data normalization?," 2021. [Online]. Available: https://www.geeksforgeeks.org/what-is-data-normalization/.

[32] A. Ali, M. Imran, A. Ali, and M. A. Khan, "Evaluating sentinel-2 red edge through hyperspectral profiles for monitoring lai chlorophyll content of kinnow mandarin orchards," *Remote Sensing Applications: Society and Environment*, vol. 26, p. 100 719, 2022, ISSN: 2352-9385. DOI: https://doi.org/10.1016/j.rsase.2022.100719. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352938522000271.

[33] T. Kawakata and M. Yajima, "A simple model estimating leaf and top dry weight for rice plants based on accumulated air temperature.," *Journal of Agricultural Meteorology*, vol. 50, pp. 115–120, Sep. 1994. DOI: 10.2480/agrmet.50.115.

[34] H. Kamali, S. Zand-Parsa, M. Zare, A. Sapaskhah, and A. Kamgar-Haghighi, "Development of a simulation model for sugar beet growth under water and nitrogen deficiency," *Irrigation Science*, vol. 40, May 2022. DOI: 10.1007/s00271-022-00769-z.

[35] J.-C. Calvet and J. Noilhan, "From near-surface to root-zone soil moisture using year-round data," *Journal of Hydrometeorology*, vol. 1, no. 5, pp. 393–411, 2000. DOI: 10.1175/1525-7541(2000)001<0393:FNSTRZ>2.0.CO;2. [Online]. Available: https://journals.ametsoc.org/view/journals/hydr/1/5/1525-7541_2000_001_0393_fnstrz_2_0_co_2.xml.

[36] scikitlearn, "Cross-validation: Evaluating estimator performance," [Online]. Available: https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation.

5

[37] scikitlearn, "Sklearn.ensemble.randomforestregressor," [Online]. Available: https://scikit-learn.org/0.16/modules/generated/sklearn.ensemble.RandomForestRegressor.html.

[38] J. Kaweewong, S. Tawornpruek, S. Yampracha, R. Yost, S. Kongton, and t. Kongkeaw, "Cassava nitrogen requirements in thailand and crop simulation model predictions," May 2013.

[39] t. f. e. Wikipedia, "Coefficient of determination," [Online]. Available: https://en.wikipedia.org/wiki/Coefficient_of_determination.

[40] J. Catanzarite, "What is an "out-of-bag" sample in a random forest model?," 2019. [Online]. Available: https://medium.com/learning-from-data/what-is-an-out-of-bag-sample-in-a-random-forest-model-8a826020e6f6.

[41] F. Wikipedia, "Out-of-bag error," [Online]. Available: https://en.wikipedia.org/wiki/Out-of-bag_error.

[42] Radhika, "Out-of-bag (oob) score in the random forest algorithm," [Online]. Available: https://www.analyticsvidhya.com/blog/2020/12/out-of-bag-oob-score-in-the-random-forest-algorithm/#h2_1.

[43] Eryk, "Explaining feature importance by example of a random forest," 2019. [Online]. Available: https://towardsdatascience.com/explaining-feature-importance-by-example-of-a-random-forest-d9166011959e.

[44] P. Płoński, "Random forest feature importance computed in 3 ways with python," 2020. [Online]. Available: https://mljar.com/blog/feature-importance-in-random-forest/.

[45] J. Hoare, "How is variable importance calculated for a random forest?," [Online]. Available: https://www.displayr.com/how-is-variable-importance-calculated-for-a-random-forest/.

[46] K. Nicodemus and Y. Shugart, "Impact of linkage disequilibrium and effect size on the ability of machine learning methods to detect epistasis in case-control studies," vol. 31, no. 6, pp. 611–611, 2007.

[47] "Crop nutrition sugar beet," [Online]. Available: https://www.yara.co.uk/crop-nutrition/sugar-beet/growth-and-development-of-sugar-beet/.

[48] M. J. Anar, Z. Lin, G. Hoogenboom, *et al.*, "Modeling growth, development and yield of sugarbeet using dssat," *Agricultural Systems*, vol. 169, pp. 58–70, 2019, ISSN: 0308-521X. DOI: https://doi.org/10.1016/j.agsy.2018.11.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0308521X18303597.

[49] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," *Computational statistics & data analysis*, vol. 52, no. 4, pp. 2249–2260, 2008.

[50] K. L. Yuxuan Hu and A. Meng, "Agglomertive hierarchical clustering using ward linkage," 2018. [Online]. Available: https://jbhender.github.io/Stats506/F18/GP/Group10.html.

[51] "Spearman's rank correlation coefficient," [Online]. Available: https://en.wikipedia.org/wiki/Spearman%5C%27s_rank_correlation_coefficient.

[52] "Spearman rank correlation (spearman's rho): Definition and how to calculate it," [Online]. Available: https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/spearman-rank-correlation-definition-calculate/.

[53] "Ward's method (minimum variance method)," [Online]. Available: https://www.statisticshowto.com/wards-method/.

[54] W. Li and M. Migliavacca, "Widespread increasing vegetation sensitivity to soil moisture," *Nature Communications*, vol. 13, 2022, ISSN: 2041-1723. DOI: https://doi.org/10.1038/s41467-022-31667-9.

[55] V. Lyashenko, "How to use random forest for regression: Notebook, examples and documentation," [Online]. Available: https://cnvrg.io/random-forest-regression/.

[56] A. Veloso, S. Mermoz, A. Bouvet, *et al.*, "Understanding the temporal behavior of crops using sentinel-1 and sentinel-2-like data for agricultural applications," *Remote Sensing of Environment*, vol. 199, pp. 415–426, 2017, ISSN: 0034-4257. DOI: https://doi.org/10.1016/j.rse.2017.07.015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425717303309.

A

APPENDIX

A.1. ADDITIONAL FIGURES

A.1.1. ADDITIONAL FEATURE IMPORTANCE FIGURES

This section provides some additional measured feature importance figures.

1. Measured feature importance of two methods
   Figure. A.1 shows the measured feature importance by using both Gini-based and permutation-based methods for VV backscatter and VV coherence.

2. The measured feature importance for VV backscatter and VV coherence when two features are highly relevant and one of them is removed from the model. Figure. A.2 shows the measured results.

A.1.2. ADDITIONAL PREDICTION RESULTS

In this section, additional figures of the prediction results of section 4.2.3 are provided. CWAD was removed for the regression model of both polarization channels of backscatter and LAI for the coherence VV model. The figure. A.3 visualizes how the prediction accuracy changed.

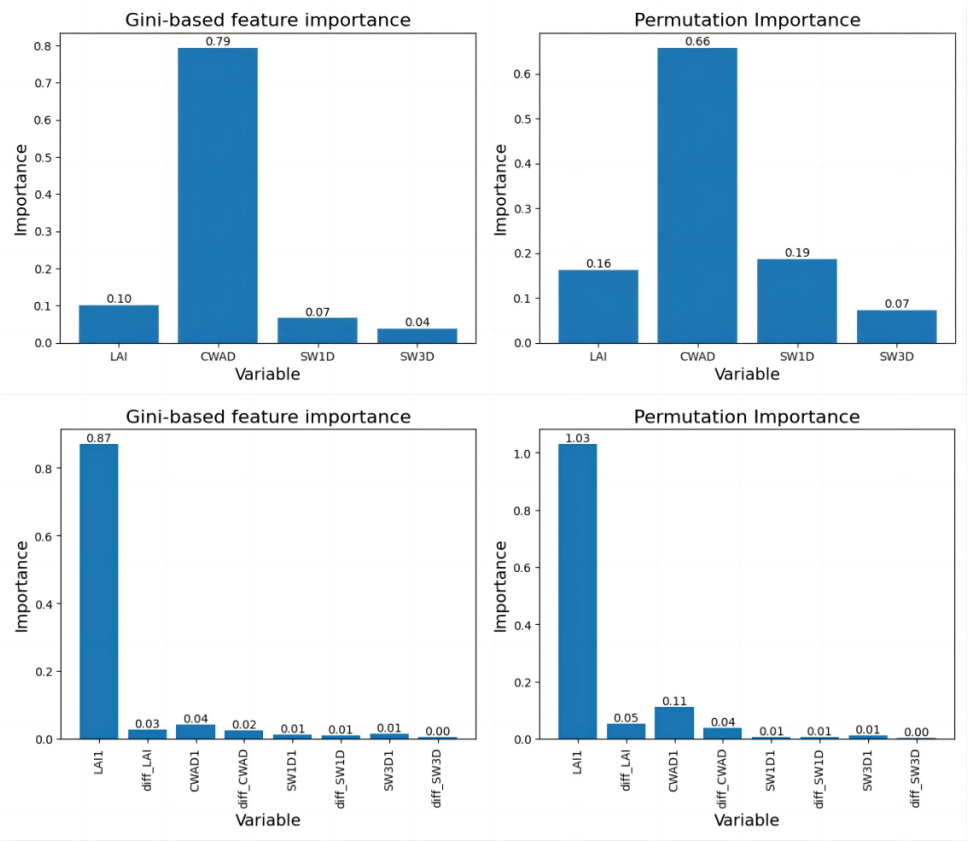


Figure A.1: The comparison of measured feature importance results between the Gini-based method and permutation-based method for VV backscatter and VV coherence.

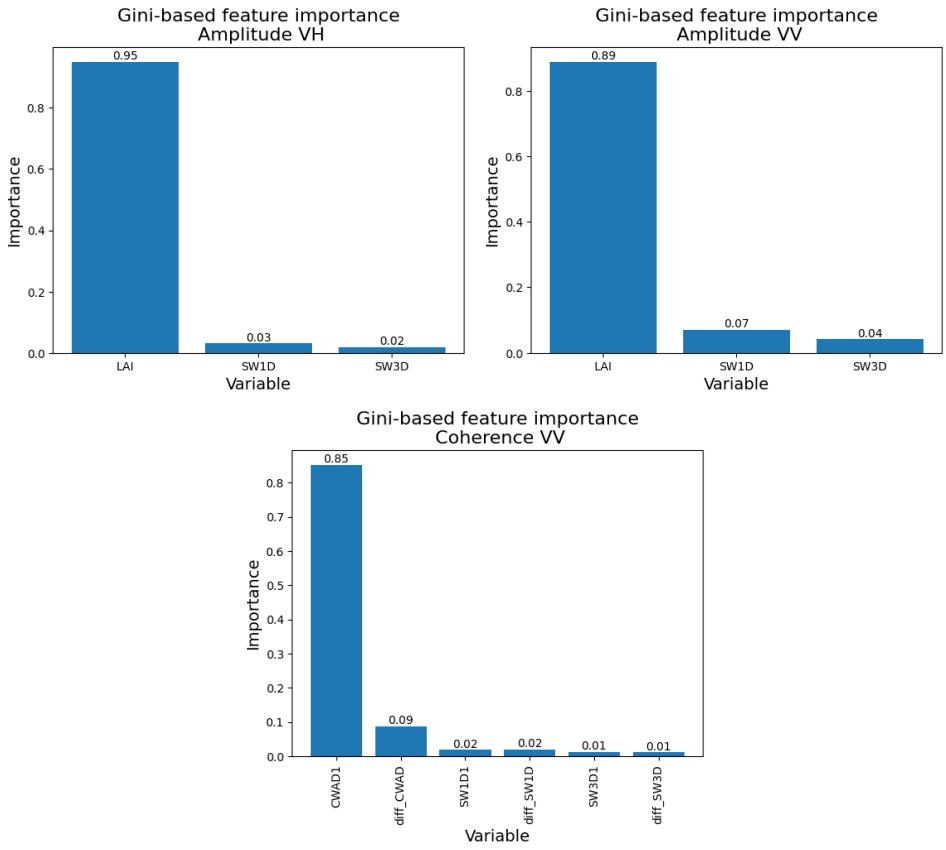


Figure A.2: The measured feature importance results with features that the correlation between them is not strong. CWAD is removed for amplitude VH and VV, and LAI related features are removed for coherence VV

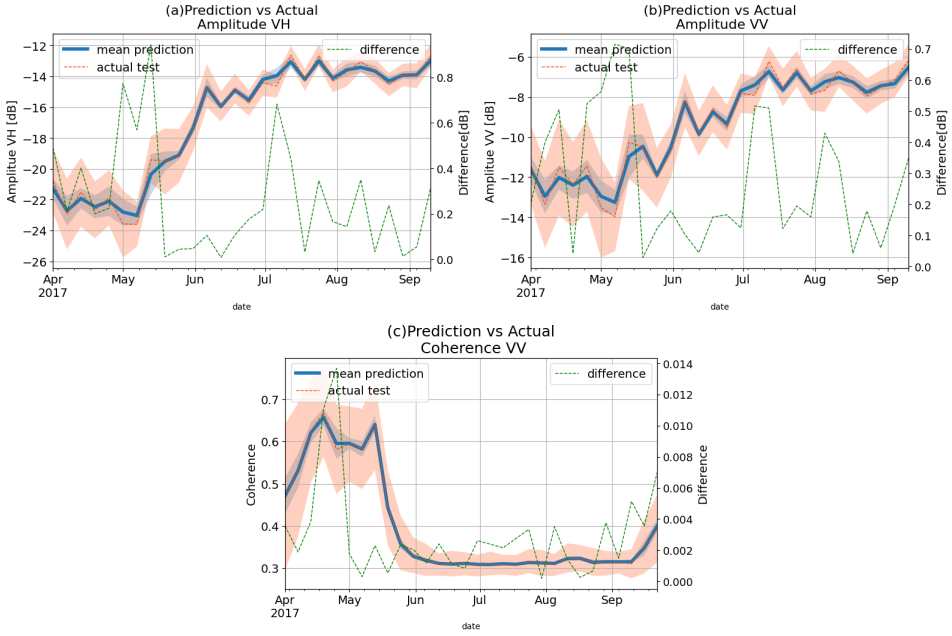


Figure A.3: Comparison of radar observables between random forest predictions and Sentinel-1 observations by applying the random forest model with the most important feature removal. (a) and (b) are the results of σ^0 VH and VV. (c) is the result of coherence VV. Mean prediction and actual values are represented by solid blue lines and orange dashed lines, respectively. Standard deviations are shown by the filled areas surrounding the curves. The time series of differences between prediction and test values are plotted in green dashed lines.