Soybean Phenotyping for Rapid Variety Breeding in Zambia

Emmanuel Muteba Ngonga

Master of Science by Research in Biology

University of York

Biology

November 9, 2020

“*We make our world significant by the courage of our questions,*
and the depth of our answers.”

Carl Sagan.

Abstract

In this study, maximum potential yield and maturity periods of three soybean varieties: Dina, SC-Safari, and SC-Spike were studied using remote sensing. To optimise this process, Landsat-8, PlanetScope, and Sentinel-2 satellites were combined into a virtual constellation that tracked the chlorophyll levels of the soybean fields with high spatial, spectral and temporal resolutions. A Random Forest algorithm was used to classify pixels thereby masking clouds and cloud's shadows from the images. NDVI (Normalised Difference Vegetation Index), and EVI (Enhanced Vegetation Index) formed the basis for determining chlorophyll levels of the soybean's canopies. Average values of NDVI and EVI per field were plotted against time using Gaussian process modelling and cubic spline interpolation to obtain their time-series profiles for 2016/2017, 2017/2018 and 2018/2019 farming seasons. Analysis done using linear, logarithm and power functions modelled the relationship of the maximum average EVI and NDVI to yield for each of the varieties. The analysis found that linear and logarithm functions explain this trend more accurately than power functions. Maximum average NDVI showed a higher correlation to yield in all three regressions than maximum average EVI. Maximum average NDVI versus yield R^2 values ran between 0.65 and 0.85, while maximum average EVI versus yield R^2 values ran between 0.28 and 0.83. Different regression equations were observed per variety. Extrapolation of their equation trend lines was used to rank the varieties according to maximum potential yield. SC-Spike showed the highest maximum yield potential in terms of metric tons per hectare after the crops were threshed. Followed by SC-Safari, Dina showed the least maximum potential yield. Maturity period was determined by measuring the time taken for the EVI and NDVI values at germination to recur during senescence. SC-Safari showed the earliest maturity period, followed by SC-Spike. Dina showed the most extended maturity period.

Contents

List of Tables

List of Figures

Acknowledgments

The research presented in this thesis could not have been possible if not for the assistance, patience, and unwavering support of many individuals. I take this opportunity to acknowledge the most outstanding among them.

I extend my deepest gratitude first and foremost to my research supervisor, Professor Katherine Denby, for mentoring me throughout my research. She helped me to overcome this exciting challenge, and for that, I am sincerely grateful. In the same vein, I am sincerely thankful to Professor Lisa Emberson for her wise counsel in my Thesis Advisory Panel meetings. I would also like to take the pleasure of thanking Doctor Godfree Chigesa for helping me with the logistics during my placement at IITA, and Doctor Joseph Fennel for his support in developing the Python codes and QGIS processes that made the data analysis in this study possible.

The research that led to this document could not have been possible without the farmers who so graciously provided me with the yield data from their fields. Mr Bob Chewetu at the MRI-Syngenta farm, Mrs Josephine Phiri at Gaulunia farm, Mr Sean Cooke at Chartonel farm and Mr Festo Mwemba at Zamseed. I am eternally grateful to you all.

I extend my appreciation to members of the Denby group in the Center for Novel Agriculture Products (CNAP) lab at the University of York who served as a voice of quiet wisdom in many matters ranging from the most fundamental aspects of this research to the way of living and getting around in York. They made my stay in York a memorable one. I also extend my appreciation to the members of the soybean and maize breeding program at IITA who accepted me into their home and made my placement a fabulous experience. It would be remiss of me not to acknowledge the members of the Dubai 2025 group who so tirelessly provided great moral support during this research. You are great friends and I am extremely lucky to know you all.

Last but not least, I would like to express unending gratitude to my brothers and sisters. They have stood by me through thin and thick during my research. Thank you very much for such unwavering love. You are the best family in the world.

This research was funded by the Development in Africa with Radio Astronomy, DARA Big Data program. It was conducted between the University of York, University of Manchester and the International Institute for Tropical Agriculture, IITA.

Declaration

I, Emmanuel Muteba Ngonga, declare that this thesis titled, 'Soybean Phenotyping for rapid variety breeding in Zambia' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at The University of York.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my work.
- I have acknowledged all primary sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed.

Signed: ______________________

Date: 28th March 2020

Chapter 1

Introduction

1.1 General introduction

Agriculture plays a significant role in supporting livelihoods, economic growth, and food security over most of Sub-Saharan Africa (SSA). Soybean breeding and farming are critical to this endeavour [1]. Sub-Saharan Africa is geographically the area of the continent of Africa that lies south of the Sahara desert. More than 70% of this region's burgeoning population lives in rural areas where the livelihood of over 85% of the people depends on rain-fed agriculture [2]. The type of agriculture carried out in this region is mostly unmechanised and is becoming more and more dependent on having drought and disease resilient, high yielding crops because of the effects of climate change [3]. Food production needs to increase in Africa to provide for the dietary needs of both its population and the rest of the global population, especially as the world's population grows, and more and more people get lifted out of poverty [4, 5]. SSA has realised exponential growth in soybean production in recent decades, from 13,000 metric tons in the 1970s to 2,300 000 metric tons in the early 2000s [6].

Opportunities for scaling up of soybean farming in the African tropics and sub-tropics as a major cash crop are necessary and promising, especially since this vast continent possesses 60% of the world's uncultivated arable land [7]. Investments in and evaluation of impacts of soybean improvement research across SSA are vital to making this a reality. Currently, less than 1% of the global soybean production is done in Africa [8]. This region's population is one of the most vulnerable to the impacts of global climate change because of overwhelming reliance on types of agriculture that are highly sensitive to weather and climate variables such as rainfall, temperature, and light, meanwhile the region is experiencing a substantial change in these parameters in recent years. To assist the soybean farmers in SSA counter these environmental challenges, organisations such as the International Institute for Tropical Agriculture (IITA), Syngenta, SeedCo, Zamseed, and many others, are breeding drought and disease resilient high yielding soybean varieties.

This study was conducted using Earth observation satellites to monitor soybean fields in Lusaka and Chongwe districts in Zambia (13.1339^{o}S, 27.8493^{o}E), Southern Africa using the process of remote sensing. Zambia has a tropical and sub-tropical climate depending on the altitude of a place. Lusaka and Chongwe districts neighbour each other and due to their high altitude, feature a humid subtropical climate. Most of the rainfall in these districts falls in summer between the months of November and March, which is when soybean is mostly grown. In recent years these districts have been experiencing warmer summers and diminished, unpredictable rainfall patterns in a shortened rainy season [2]. These changes are expected to continue and even worsen. The country lies in tropical latitudes, and according to current climate change projections, the tropics and SSA in particular will see roughly 2.0 to 4.5^{o}C of temperature rise by 2100. This is expected to be larger than the average global warming and could have a tremendously adverse effect on

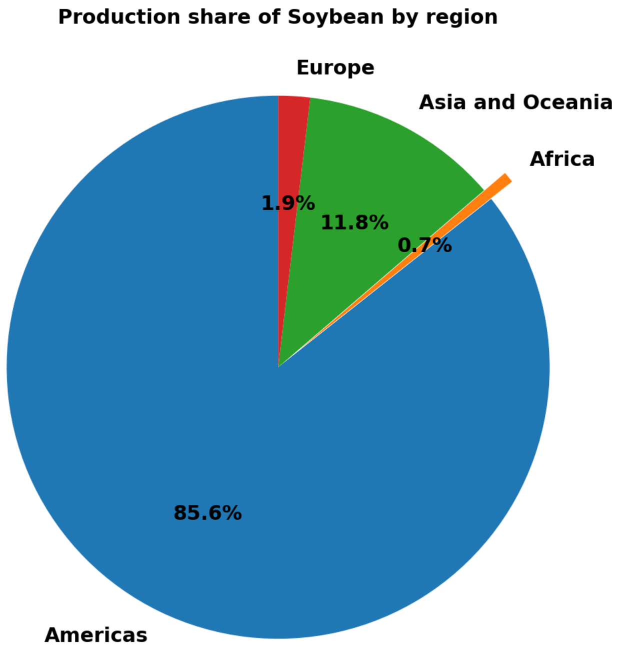


Figure 1.1: Soybean production share of the different regions of the world. Although Africa has the largest amount of arable land in the world, it contributes less than 1% of the total soybean production. FAO 2019.

the agriculture industry in the region [9]. To assist the soybean farmers in SSA counter these environmental challenges, organisations such as the International Institute for Tropical Agriculture (IITA), Syngenta, SeedCo, Zamseed, and many others, are breeding drought and disease resilient high yielding soybean varieties.

Remote sensing in the form of multispectral satellite imagery has great potential to revolutionise the way farmers, and extension officers monitor crops, it is making the process of field observation, increasingly more accessible, non-destructive, and accurate [10]. It is particularly well suited for a place like Africa where fields can be very far apart. Large swathes of land can be observed and the relative health of the crops in them determined with cutting edge precision without having to visit the places in person. Remote sensing was used in this study to observe the fields in which soybean crops were growing, paying particular attention to the variety, chlorophyll levels (as a means to monitor canopy photosynthesis rates), resulting yields, and maturity periods.

1.2 Soybean in Zambia

Soybean (*Glycine max* (L.) Merrill) is a leguminous, self-pollinating crop native to China that grows in tropical, subtropical, and temperate climates. It grows well in many soil types, especially those with high clay content and does less well on weak sands. Its ability to embrace diverse climatic conditions and soil types has made it a globally cultivated crop, something that's not possible with many cash crops. It was first introduced to SSA by Chinese traders along the east coast in the nineteenth century and was cultivated commercially as early as 1903 in South Africa and by 1910, it had reached Zambia [11]. Its production plays an increasingly dynamic role in stimulating economic growth in Zambia's agrarian society, and the arable land in the country is vast enough to accommodate future expansion. The Agriculture sector's contribution to the Gross Domestic Product (GDP) of the country has been steadily declining over the last few years. This trend has been closely related to the climatic events that the country has been experiencing and the high dependence of smallholder farmers on rain-fed production [12]. Zambia is self-sufficient in soybean production with 85% of this coming from commercial farmers whose yield rates of about 2.5tons/ha are comparable to the global average since they can afford farm inputs and irrigation systems. The smallholder farmers, on the other hand, have meagre yield rates of about 1.5 tons/ha, which drops further during drought years [13]. Soybean is cultivated in nearly all parts of Zambia, with the Eastern, Central and Northern Provinces taking the lead. It is important to encourage the cultivation of such a versatile crop among small scale farmers and to help them make their endeavours profitable in a developing country like Zambia as the growth of the agricultural sector is also the clearest avenue to achieve sustainable economic growth. It is also the best means of stopping the country from over-reliance on copper revenues so that the economy can become more diverse and therefore help alleviate poverty [14]. When growth is encouraged in sectors where most poor people earn their living, like agriculture, in this case, the likelihood that poverty will be alleviated increases.

Soybean is arguably the world's undisputed champion of crop versatility. Its seeds have the highest protein and oil content of any crop. An average soybean seed may contain 40% protein and 20% edible oil. Soybean now accounts for 53% of the global production share of all oilseed crops [15]. Growing it can improve the nitrogen levels in the soil, thereby making the soil more fertile. These attributes have resulted in its widespread use in food, animal feed, nutraceuticals and industrial raw materials. Approximately 87% of soybean seeds produced globally are crushed into soy meal and soy oil. The remaining 13% is used for direct human consumption and as nutraceuticals [16]. As a nutraceutical, it has been used extensively to prevent severe malnutrition among infants, children, and pregnant and lactating women during times of famine in many countries [17]. Demand for this crop is expected to continue to surge across the world, underpinned by a growing population and the rising demand for high protein and high-fat meals for both humans and livestock and the ever-increasing need for industrial raw materials that comes with the world's economic growth [18, 19]. Soybean will, therefore, continue over the coming decades to offer a huge opportunity for smallholder farmers in Zambia and the rest of SSA to improve their cash base [16, 20]. In Zambia, as of 2019, animal feed accounted for 89% of the soybean consumption, with the majority of this going to the poultry sector. The remaining 11% went to human consumption where it was usually consumed in the form of soya chunks, oil, and other products such as porridge. Both the animal and human consumption sectors are growing at significant rates yearly in Zambia [14].

Soybean is susceptible to the photoperiod of the area in which it grows [26]. Since it is a legume, soybean has the ability to live in a symbiotic relationship with a type of bacteria called rhizobia. This mutually beneficial relationship works in such a way that the bacteria fix nitrogen from the air into ammonia which is used by the plant as a source for nitrogen, in turn, the crop provides carbohydrates to the bacteria which live in a part of the roots called nodules that are formed as a result of this symbiotic arrangement [21]. This mechanism results in a biological nitrogen-fixing process that is beneficial not only to soybean but to other crops that are grown in the same field. Therefore susceptibility to nodulation with Rhizobium strains present in SSA soils is one of the most desired traits that is being bred for by the seed-producing organisations in the region. Inoculation of the seeds with compatible and appropriate rhizobia before sowing is necessary for places with a low population of native rhizobial strains and is one of the leading solutions for increasing soybean yields since it is much cheaper than adding artificial fertilisers. This ability to biologically fix nitrogen into the soil has driven soybean's use in crop rotation or mixed cropping with maize, the main staple food for 50% of the population in Africa to increase its yield. The majority of smallholder farmers in Zambia are unable to afford the high priced mineral artificial fertilisers. Hence, soybean provides a good alternative for nitrogen fixation [6, 22]. In addition, Striga hermonthica, a parasitic weed that infests over 60% of farmland in

SSA on which maize is monocropped is suppressed by having a soybean-maize intercropping or crop rotation system. These and other factors are assisting farmers in SSA to improve their gross income from growing maize and soybean. Thereby improving the resilience of their livelihoods and the agricultural, economic and environmental sustainability of the region [23].

Photosynthesis is the process by which electromagnetic energy is converted to chemical energy in chlorophyll-containing organisms like soybean. There is generally a strong relationship between nitrogen per unit leaf area and the photosynthetic ability of a soybean crop's leaves [24, 25]. Soybean needs sustained high photosynthetic rates and the accumulation of large amounts of nitrogen in its seeds to achieve high yields. The nitrogen formed in the nodules plays a vital role in forming photosynthetic pigment-proteins, including chlorophyll a and b. The nitrogen exists in leaves primarily as Ribulose Biphosphate Carboxylase/Oxygenase (RuBisCO) which is a significant protein in the stroma of chloroplasts. The abundance of sunlight hours in Zambia provides a conducive environment for the photosynthetic capability of all soybean varieties.

Soybean is heavily influenced by the photoperiod of the area in which it grows, it is categorised as a short-day plant [26]. This means that it will only flower once the daylight hours fall below a certain critical amount, i.e. soybean is susceptible to the photoperiod of the area in which it grows. During the vegetative period, when the daylight hours become sufficient for the flowering to begin, a chemical in the leaves called phytochrome which responds to day-lengths, sends a signal to the meristem (a tissue that consists of undifferentiated cells) in the nodes where flowers appear. This encourages the initiation of flowering, hence marking the start of the reproductive phase. Since soybean is a crop from temperate latitudes, it may often flower too early when cultivated in the short days of the tropical and subtropical regions. The result is insufficient biomass accumulation during the vegetative phase prior to the reproductive phase and since it's yield is highly dependent on its photosynthetic capacity this contributes to the low yields experienced by this crop in this part of the world [26]. Breeders of soybean in SSA are therefore developing strains that are less sensitive to this short day photoperiod.

Soybean varieties are divided into two types of cultivars; determinates and indeterminates [27]. This classification distinguishes them according to the growth style of the variety. Determinates stop producing new leaves once flowering begins, that is to say, their vegetative and reproductive phases do not overlap each other. This makes them more susceptible to low yields when grown in an area with short days. On the other hand, indeterminates will continue to produce leaves at least three weeks into flowering, i.e. their vegetative and reproductive phases overlap thus making them a lot less susceptible to having low yield in areas with short days. Indeterminate varieties are suited to regions that have readily available rain or good irrigation schemes. The Lowveld (150 to 600 metres above sea level) in Sub-Saharan Africa provide a conducive environment for such varieties. Both types can do well in the Middle veld (between 900 and 1200 metres above sea level) and Highveld (between 1500 and 2100 metres above sea level) [28].

After the 2017/2018 farming season, Zambia emerged as the third-largest producer of soybean in Africa surpassed only by Nigeria and South Africa at number one and number two respectively while Uganda was fourth. Soybean yields have remained stagnant at 1.5 tons/ha for most varieties for the past two decades in SSA which presents one of the most significant challenges for breeders of this crop in the region. Given such a scenario there are two main possible ways to meet the burgeoning demand for soybean, one is to increase the number of hectares that go into its cultivation, and the other is to increase the yield of the crops by breeding varieties that thrive in this region. Ultimately the best way is in using both of these processes together so that better varieties can be grown on more land while making the production system more efficient and productive by embracing modern technologies and farming techniques that assist in increasing yield. As of 2019, South Africa is the only SSA country that has approved genetically modified (GM) soybean varieties for commercial production. This means that the rest of the region has to rely on conventional (classical) breeding processes to produce varieties that are better suited for the SSA conditions.

1.2.1 Soybean breeding

Since 1974, IITA has played a leading role in breeding soybean varieties that are well suited for the tropical environment, they have been joined in recent years by other organisations such as Zamseed, SeedCo and Syngenta. These organisations desire to develop high-yielding varieties that can withstand the many stresses of the tropics while promoting agronomic technologies and soybean processing and utilisation suitable for smallholder farmers in SSA [29].

These researchers are developing varieties that are resistant to soybean rust and are high yielding. Other traits they are trying to breed for include low pod shattering, soil nutrient deficiency tolerance, resistance to frog-eye, leaf spot, bacterial pustule and bacterial blight [30]. Multiple-year and multiple-location testing is fundamental to identifying and selecting stable soybean varieties that are well adapted to the diverse agro-environmental conditions and management techniques in SSA [31]. The breeding process occurs over a large area and involves many different varieties of soybean. Observing the soybean crops still in the breeding process and those that have been released to farmers can be transformed by using remote sensing technology. Remote sensing provides the possibility for early, efficient, objective and non-destructive ways for monitoring large numbers of fields and varieties of the sort that IITA and the other organisations are developing [32].

The conventional (classical) breeding process involves intentionally crossing closely or distantly related varieties of soybean in order to produce new varieties that have the desired properties. For example, an early maturing variety may be crossed with a high yielding but late-maturing one, to introduce early maturity without losing the high-yield quality [33]. In a process called backcrossing, progeny from the first cross are then crossed with the high-yielding variety to ensure that the new progeny are more like the high-yielding parents. The progeny so developed are then tested for yield and early maturity, and the most high-yielding early maturing individuals are selected and developed further. The soybean that passes the desired criterion is then left to self pollinate to produce inbred varieties for breeding, and the crops are observed for several farming seasons and in as many different farming conditions as possible before they are introduced to farmers [18]. Evaluation of yield and maturity period of the new breed-lines is fundamental to having these cultivars certified for release to the general public [34, 35]. The large number of breeding plots that result from providing different farming conditions makes it extremely tedious and logistically challenging to observe the varieties on a daily basis. Using Satellite Images for phenotyping can make this a lot easier, more accurate and less time-consuming. It is becoming ever more evident that the designing of great soil and crop management techniques that exploit the climatic and genetic yield potential of soybean will be a great deciding step of the amount of economic and social development Zambia and many other Sub-Saharan African countries can attain. For the breeding program to meet the future challenges posed by climate change, it has to become more rapid than its current pace of five to six years for a new variety to be released to the public. New techniques have to be developed that tell the characteristics of a new strain within one or two years of growing the experimental crop.

Extensive research on plant physiology has been carried out to select and breed for genotypes with higher photosynthetic rates. Such research has made it widely accepted that increasing leaf photosynthetic rates is one of the most effective ways of increasing yields in grain crops [36].

This is especially true in cultivars such as wheat, maize, and soybean [36]. It is also understood that different varieties of a crop that have the same photosynthetic rate may have different yield values because of their different genetic yield potential [37]. Genetic yield potential is the yield of a cultivar when grown in an environment to which it is well adapted, with enough nutrients and water and without any limiting fators such as pests, diseases, weeds, lodging, and other stresses [37]. In theory a crop canopy with high Photosynthetically active biomass levels will have a higher yield than one with lower levels provided the crops are of the same variety and grow in the same environmental conditions [37]. Furthermore, this relationship between the chlorophyll levels in a canopy and the resulting yield of the crop may follow a mathematical relationship that depends on the variety of the crop being observed. Therefore, a spectral index like Normalised Difference Vegetation Index (NDVI) and Enhanced Vegetation Index(EVI) that quantifies the chlorophyll levels of a soybean canopy as captured by multispectral imaging, in essence, quantifies the rate of photosynthesis going on in the canopy at the time the image is captured. These spectral indexes and the resulting features of their time-series profiles might in some way be correlated to the yield that a canopy produces provided the crop is not damaged by some environmental stresses. Such a correlation can prove useful in the classical soybean breeding program as each variety of soybean may have its relationship between yield and the EVI and NDVI of its canopy and thus shows its maximum possible yield and its maturity period. Having a framework that determines these parameters in a shorter period can help improve the speed of the soybean breeding process.

1.3 Remote Sensing

Remote sensing is the science of obtaining and processing information from a distant object or phenomenon. Generally, the objects or events of interest are identified, measured and analysed without direct contact with the sensor. In Optical remote sensing, sensors are fundamentally electro-optical transducers that take up photons as a source of energy and eject a digital bit-stream (electrical signal) that represents the energy of the incoming photons. Other energy forms (physical carriers) that are used in remote sensing include gravity, sound, and magnetism. Sensors can acquire data remotely while being on board different types of platforms such as space satellites, Earthbound antennae, aeroplanes, UAVs and handheld devices [38]. The medium through which these physical carries pass may be a vacuum and any or all of the states of matter. Remote sensing has been extensively used to study crop traits such as chlorophyll levels of canopies, leaf area index(LAI), crop heights, crop sensitivity to drought, nitrogen concentration, disease susceptibility, etc., and how these traits are related to crop yield [39, 40, 41]. In this study we will use remote sensing to determine the chlorophyll levels. The other varibles will not be considered due to spectral band constaints in one of our satellite constellations (PlanetScope).

There are two types of remote sensing systems, namely passive remote sensing, where the sensing device measures the energy in a reflected medium produced by another source such as the reflected solar radiation from the Earth's surface being detected by a satellite, and active remote sensing, where the sensing device emits a medium onto the object being studied and the energy of the medium that is reflected back is measured. The remote sensing techniques used in this study are based on passive optical remote sensing. All objects have an individual and characteristic manner in which they interact with incident electromagnetic radiation, and thus an individual and characteristic reflected spectral signature which distinguishes them from other objects [42]. In this study, the distant phenotypic features of soybean crops were observed by detecting the characteristic solar electromagnetic radiation which was reflected from them.

Four resolutions are essential in remote sensing using Earth observation satellites. These are spectral, spatial, temporal and radiometric resolution. These resolutions were of primary importance when choosing the satellite constellations to use in this study. The spectral resolution is the ability of a sensor to distinguish minute electromagnetic wavelength intervals from each other [42]. Spectral resolution divides satellite data into two types, multispectral and hyperspectral imagery. Multispectral imagery has anywhere between 3 and 15 spectral bands while hyperspectral imagery has hundreds of spectral bands. Hyperspectral imaging measures continous spectral bands while multispectral imaging measures discrete spectral bands. Spatial resolution is the ground surface area that forms one pixel in a satellite image [43]. Radiometric resolution or dynamic range of a sensor is its ability to distinguish between the differences in the energy of the incoming radiation. Radiometric resolution is determined by the bit depth of the respective pixels of a satellite image. An image with a 12-bit radiometric resolution is more sensitive to the energy levels than one with an 8-bit radiometric resolution. Temporal resolution is a measure of the revisit cycle, i.e. the frequency with which a satellite constellation observes the same part of the Earth's surface. This is determined by the constellation's design and orbit pattern [42, 43]. Combining different satellite constellations in an observation can help provide better temporal resolution, but it also brings about the challenge of harmonising the data from the different satellites. With a good enough spatial, spectral and radiometric resolution, objects can be distinguished even from other objects of their kind, especially when we use a spectral index that makes them stand out.

To capture the soybean canopy images accurately, we needed robust, high performing satellites which capture high-quality data. We then developed an image analysis pipeline that cleaned the images, cropped the desired field, computed the biomass levels in the form of average EVI and NDVI and produced their time series profiles for that particular farming season. The satellites used in this study were chosen for their heliosynchronous orbits, this means that their geocentric orbits combined altitude and orbital velocity in such a way that they passed over any given point on the planet's surface at the same local solar time of that point [44]. Three satellite constellations were used in this study to improve the temporal resolution of observing the soybean fields. PlanetScope, Sentinel-2 and Landsat-8 satellites were used to form a single constellation with a high temporal resolution. Sections 1.3.2, 1.3.1, and 1.3.3 describe these individual constellations.

1.3.1 Sentinel 2

The Sentinel-2 constellation is comprised of two identical satellites; Sentinel-2A and Sentinel-2B which are Earth observation satellites developed by the ESA (European Space Agency) under their Corpenicus program to provide continuity of SPOT and LANDSAT-type image data, they were designed to provide image data for agriculture, forestry, food security programs, and humanitarian relief operations. These observatories are in low earth orbit 180^o apart facilitating a revisiting time of 5 days under the same viewing angles for low latitude regions like Zambia. At high latitudes, the two satellite's image swaths overlap, allowing for some regions to be observed twice or more every five days. The only downside is that the viewing angles are different for most such observations[45]. The twin satellites have a coverage limit between latitudes 56^o south and 84^o north to maximise their land surface coverage while efficiently saving power. Sentinel-2 images are multispectral with 13 spectral bands spanning through the visible, near-infrared and short wave infrared part of the electromagnetic spectrum [46]. The image swaths are a massive 290 km field of view and since each image is a square it observes an area on the Earth that is about 84,100 km^2. Despite the large swath, the images are of high spatial resolution; 10m for the visible and near-infrared bands, 20m for the red vegetation edge, SWIR and 60m for water vapour and SWIR cirrus bands. The fact that Sentinel-2A was launched in June 2015 and was only joined by Sentinel-2B in March 2017 means that the revisit time of the constellation in 2015, 2016 and parts of 2017 is once every ten days for low latitude regions.

The narrowness of the near-infrared band (band 8a) at 865nm is designed to avoid contamination from water vapour while still being able to capture enough near-infrared for vegetation and iron oxide sensing. This is similar with what was implemented in Landsat-8 to correct for the water vapour contamination observed in previous versions of the Landsat constellation. Table 1.1 shows the 13 spectral bands captured by the sensors, their bandwidth, and central wavelengths.The inclusion of the spectral band in the blue domain at 443nm allows for precise aerosol correction of acquired data. Sentinel-2 images can be downloaded from many different sources. For this experiment, they were downloaded from USGS's Earth explorer website, they came as level 1 products and were bottom of atmosphere corrected using QGIS (Quantum Geographic Information System) to do away with atmospheric effects.

Table 1.1: Sentinel-2 spectral band wavelengths and resolutions

Sentinel-2 Spectral bands	**Sentinel-2A**		**Sentinel-2B**		Spatial Resolution
	Central wavelength (nm)	Bandwidth (nm)	Central wavelength (nm)	Bandwidth (nm)	
Band 1 - Coastal aerosol	442.7	21	442.2	21	60
Band 2 - Blue	492.4	66	492.1	66	10
Band 3 - Green	559.8	36	559.0	36	10
Band 4 -Red	664.6	31	664.9	31	10
Band 5 - Vegetation red edge	704.1	15	703.8	16	20
Band 6 - Vegetation red edge	740.5	15	739.1	15	20
Band 7 - Vegetation red edge	782.8	20	779.7	20	20
Band 8 - NIR	832.8	106	832.9	106	10
Band 8A - Narrow NIR	864.7	21	864.0	22	20
Band 9 - Water vapour	945.1	20	943.2	21	60
Band 10 - SWIR cirrus	1373.5	31	1376.9	30	60
Band 11 - SWIR	1613.7	91	1610.4	94	20
Band 12 - SWIR	2202.4	175	2185.7	185	20

1.3.2 PlanetScope

The PlanetScope satellite constellation consists of about 120 individual satellites called Doves. They are operated by a private Earth-imaging company called Planet Labs based in San Francisco USA. Their on-orbit capacity is constantly improving in capability and quantity with frequent launches and technological advancements. PlanetScope refers to their first three generations of satellites; PlanetScope 0, PlanetScope 1 and PlanetScope 2. This constellation images a greater part of the Earth's surface every day, resulting in almost daily coverage of any area of interest at 3-5 meter spatial resolution. The Doves are in two types of orbits depending on how they are deployed, some are launched from the International Space Station (ISS) and are in a 51.6^o orbital inclination (inclination is measured relative to the equator in the same way longitudes are measured) similar to that of the ISS while those deployed as rocket payloadS are in $\backsim 98^o$ orbit. Both types of Doves have a camera dynamic-range of 12 bits [44]. Table 1.2 shows the properties of the two types of satellites operating in this constellation. Their high temporal resolution makes them a prime constellation for observing cloud ridden times of the farming season. Days from 15^{th} to 30^{th} December were the most cloudy for most of the farming seasons that were observed in this study and so many more PlanetScope images that were not completely overcast where downloaded for this period to increase the chance of capturing a field when few or no clouds and cloud shadows were over it. For this research readily available bottom of atmosphere and radiometrically corrected image rasters of this constellation were downloaded from `www.planet.com`. Their research account allowed us to download images covering 10,000 km^2 per month.

Table 1.2: PlanetScope satellite constellation sensor characteristics

Mission Characteristic	**ISS Orbit**	**Sun Synchronous Orbit**
Orbit Altitude (reference)	400 km (51.6^o inclination)	475 km (∼98^o inclination)
Max/Min Latitude Coverage	±52^o	±81.5^o (depending on season)
Equator Crossing Time	Variable	9:30 - 11:30 am (local solar time)
Spectral Bands	Blue 455 – 515 nm	Blue 455 – 515 nm
	Green 500 – 590 nm	Green 500 – 590 nm
	Red 590 – 670 nm	Red 590 – 670 nm
	NIR 780 – 860 nm	NIR 780 – 860 nm

1.3.3 Landsat-8

Landsat-8 is the latest in a continuous series of land remote sensing satellites by the global research program known as National Aeronautics and Space Administration's (NASA's) Science Mission Directorate (SMD), a long-term program that studies changes in Earth's global environment. The program began in 1972 and has operated several satellites starting with Landsat-1 up to currently, Landsat-8. The United States Geological Survey (USGS) is responsible for the operation and maintenance of this observatory. They capture, process and distribute the satellite images and maintain the data archives of the Landsat series satellites. The data archive is comprised of both Landsat-8's and the previous Landsat satellites images. Landsat-8 offers many improvements over its predecessors it has a temporal resolution of 16 days with an equatorial crossing at 10:11 a.m. (±15 min) Mean Local Time, a radiometric resolution of 12 bits and a spatial resolution of 30 meters for all bands except the 15 meters Panchromatic band and two 100 meters Thermal Infrared bands, table 1.3 shows the respective spectral bands of a Landsat-8 image and their characteristics. Each scene of the satellite covers a 190 by 180 km surface area[47]. Landsat-8 image products are available at no cost to the user at `http://www.arthexplorer.usgs.gov/`. All the Landsat-8 images used in this study were already radiometrically and top and bottom of atmosphere corrected at the time of download. Table 1.3 shows the characteristics of the spectral bands captured by this constellation.

Table 1.3: Landsat 8 Spectral band wavelengths and resolutions

Landsat 8/OLI-TIR Bands	**Wavelength (nm)**	**Resolution (m)**
Band 1- Coastal aerosol	430-45	30
Band 2- Blue	452-412	30
Band 3- Green	533-590	30
Band 4- Red	636-673	30
Band 5-Near Infrared(NIR)	851-879	30
Band 6-SWIR-1	1566-1651	30
Band 7-SWIR-2	2107-2294	30
Band 8-Panchromatic	503-676	15
Band-9-Cirrus	1363-1384	30
Band 10- Thermal Infrared (TIR) 1	1060-1119	100
Band 11- Thermal Infrared (TIR) 2	1150-1251	100

1.3.4 Spectral Reflectances of Green Plant's

Robert N Cowell was the first scientist to attempt to use remote sensing techniques to monitor crop health. He used aerial infrared photographs to track plant diseases in the field and was one of the first people to realise that the spectral reflectance of plants depends heavily on the state of their biomass [48, 49].

Over the years, it has become apparent to plant physiologists that light is a critical environmental factor in a plant's development and is very crucial to the expression of its genes. Furthermore, a crop's capacity to maximise its photosynthetic productivity depends on its ability to sense and intercept the light incident on it and evaluate and respond to its quality, quantity, and direction. Plant photosynthesis involves two types of photoreaction centres (protein complexes) found in the thylakoid membrane of chloroplasts; Photo System I (PSI) and Photosystem II (PSII). PSIs are the primary source of electrons for the reduction of carbon dioxide while Photo System II centres generate the oxidation power required for the abstraction of electrons from water. The light-dependent reactions of photosynthesis begin in PSII when a chlorophyll a molecule within the reaction centre of PSII absorbs a photon. Photosynthetic rate of soybeans (on a leaf area basis, PA) estimated from the incorporation of ^{14}CO2 is highly correlated with chlorophyll content of the photosynthetically active biomass of the soybean crops. High photosynthetic rates in soybean are in turn correlated with higher yield values [50]. This is because a high photosynthetic rate allows the plant to create enough starch for its energy allowing it to grow optimally and thus produce adequate yield.

The amount of solar radiation reflected by a plant canopy in the near-infrared determines the optical properties of the leaves in a canopy and the air–cell wall–protoplasm–chloroplast interfaces in the leaf's cellular structure[51]. In stressed vegetation, leaf chlorophyll content may decrease, thereby changing the proportion of light absorbed and reflected. Plant senescence is yet another major cause of chlorophyll depletion and is easily detectable with a vegetation index such as NDVI and EVI. In plants, molecules like chlorophyll a and b absorb only certain bands in the light spectrum such as red and blue for photosynthesis while re-emitting near-infrared and green. The quantity of absorbed and re-emitted radiation can serve as an indicator of the overall state of the photosynthetic rate of the canopy of a crop.

1.3.5 Spectral Indexes

In passive remote sensing, a spectral index is a combination of spectral reflectances from two or more wavelengths using a mathematical function. The result is an image with a particular feature like plants or buildings standing out more than it does in the individual spectral bands. The most common form of spectral indexes in plant remote sensing, also called vegetation indexes, are Normalised Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI). These are the best way to evaluate the vegetative vigour and growth dynamics of any plant canopy using surface reflectance imaging [52]. Many other indexes exist that make other features stand out.

Analysing vegetation with a single spectral band such as red or near-infrared is also feasible, but using a single spectral band does not account for seasonal sun angle differences that result from the revolution of the Earth around the sun and is very vulnerable to atmospheric attenuation unless the observatory is not spaceborne [53]. NDVI and EVI Vegetation indexes are widely understood to correlate with plant parameters such as pigment status and grain yield. Other indexes include OSAVI (Optimised Soil-Adjusted Vegetation Index), RVI (Ration Vegetation Index) which show the chlorophyll concentration of the plant, PVI (Perpendicular Vegetation Index) and DVI (Difference Vegetation Index) which can be used to estimate leaf nitrogen content in hyperspectral images [54]. In order to simplify our research and due to spectral band contraints in the PlanetScope constellation, only NDVI and EVI will be used to monitor soybean fields in this research.

In this study, we used three indexes in preprocessing and analysing the rasters of the soybean fields. The fact that spectral indexes are a normalised ratio of electromagnetic intensities that are dimensionless means that they can be used with images of different spatial resolutions to help improve the overall temporal resolution with which fields are monitored.

Radiance and Reflectance

Radiance refers to the flux energy that is emitted, reflected or transmitted by an object and received by a sensor per unit solid angle per unit projected area. Satellite images usually have their pixel values recorded in radiance. Reflectance is the measure of the proportion of light incident on a surface which is reflected off it. It is usually computed by dividing the radiance values by a scaling factor; for example, radiance values can be divided by 10000 to gain reflectance values that range between 0 and 1. Most spectral indexes are defined in terms of reflectance and not radiance.

Normalised Difference Vegetation Index (NDVI)

NDVI is used to determine the chlorophyll levels in a plant canopy due to the fact that chlorophyll absorbs most of the red radiation while reflecting almost all the near-infrared radiation falling on the chloroplasts in the biomass. Therefore, the more chlorophyll a plant canopy has, the more red light it absorbs and the more near-infrared it reflects. This makes for a very accurate means of observing plant biomass remotely using reflected Sunlight which is composed of 49% near-infrared radiation ($\lambda = 700 \rightarrow 1300nm$). NDVI analysis can, therefore, be used to detect stress that depletes chlorophyll concentration in a crop before it is visible with the naked eye [38]. Using satellites, we can determine the chlorophyll concentration of the canopy of the crops, thereby providing a good way to measure their above ground biomass levels. In remote sensing, NDVI stands out as the most popular means of analysing crop growth and photosynthetic rates to estimate overall crop health and predict yield [55]. NDVI is calculated using the formula:

$$NDVI = \frac{b_{nir} - b_{red}}{b_{nir} + b_{red}} \tag{1.1}$$

where b_{red} and b_{nir} stand for the atmosphere corrected spectral reflectance measurements acquired in the red and near-infrared bands, respectively. Different values of NDVI represent different ground features; water ranges from -0.4 to -0.1, fallow land ranges between 0.05 and 0.2, germinating crops have only a small canopy over the land so that their range is between 0.2 and 0.35. The vegetative and reproductive stages of a plant both fall in the range of 0.2 and 1.0. The NDVI values fall back to the range of 0.2 then the crops are almost ready for harvest signifying chlorophyll depletion from crop senescence.

Enhanced Vegetation Index (EVI)

NDVI analysis in vegetation studies has some limitations related to soil background brightness. Atmospheric conditions, low solar zenith angles, aerosol concentrations, sensor viewing geometry, and off-nadir viewing are some of the other factors that can cause noise in NDVI analysis [56]. Different NDVI values can be observed for canopies with different soil and moisture conditions even when the canopies are at the same growth stage. Another important factor that is very detrimental to NDVI analysis, especially when using different satellites to observe the same field, is its dependency on the time of day at which the satellite images are taken. This is because NDVI does not correct for changes in solar incidence angles [57].

To counter these effects, Huete A.R. introduced a soil correction factor when he derived the Soil Adjusted Vegetation Index (SAVI) [58]. Over the years this has been built upon to create a much more robust vegetation index called Enhanced Vegetation index developed for the MODIS

(Moderate-resolution Spectroradiometer) satellite but has been used extensively with Landsat-8 and Sentinel-2 in many studies with good results [59]. Like NDVI, EVI shows the chlorophyll concentration in a canopy and just like NDVI the range of values for EVI is -1 to 1 with healthy vegetation generally falling between values of 0.20 to 1.0. Unlike NDVI however, it accounts for soil and atmosphere influences and thus provides improved sensitivity in high biomass regions. In addition to the red and near-infrared bands used in NDVI analysis, EVI also requires the blue band [60]. EVI does not become saturated as easily as the NDVI does when viewing areas with large chlorophyll concentration. However, EVI is only optimal for certain satellites (especially MODIS) for which it was developed. This has secured NDVI as the most widely used vegetation spectral index for scientific enquiry into above ground plant biomass using satellite imaging. EVI is calculated using the following formula,

$$EVI = 2.5 * \left(\frac{b_{nir} - b_{red}}{b_{nir} + 6 * b_{red} + 7.5 * b_{blue} + 1} \right) \tag{1.2}$$

where b_{red}, b_{nir} and b_{blue} stand for the atmosphere corrected spectral reflectance measurements acquired in the red, near-infrared and blue bands, respectively. 2.5 is a gain factor, 6 and 7.5 are the coefficients of the aerosol resistance term, which uses the blue band to correct for aerosol influences in the red band and 1 is the soil-adjustment factor.

Normalised Difference Water Index (NDWI)

One major problem incountered when using satellites to study crops is clouds and cloud shadows. In order for our analysis to be more accurate we need to do away with cloud and shadow infested pixels in our rasters. To do this we developed a machine learning algorithm that detects such pixels and masks them from our rasters. This algorithm needs to view clouds and shadows in different spectral bands and spectral indexes. Especially those that highlight the features we need to mask. For this reason we will provide it with NDWI, EVI, and NDVI, spectral indexes as well as red, blue and green spectral bands. This will give it many different ways to distinguish the clouds and cloud shadows from the desired pixels. In this study, NDWI is used solely for this purpose.

Two types of Normalised Difference Water Index (NDWI) exist. One measures water levels in plant canopies while the other measures water content in water bodies. In this study we used the NDWI that measures water content in water bodies. This is because PlanetScope images do not come with the Short Wave Infrared band which is crucial to computing NDWI for water content in plant canopies.

Water bodies have strong absorbability and low radiation in the range from visible to infrared wavelengths. NDWI uses the green and near-infrared bands of remote sensing images to make water bodies stand out from other features based on this phenomenon. It ranges between -1 and 1 and pixel values of water bodies are larger than 0.5. Vegetation has much smaller values, which makes distinguishing vegetation from water bodies easier. Built-up features have positive values between 0 and 0.2. Water body NDWI is computed using the following equation;

$$NDWI = \frac{b_{green} - b_{nir}}{b_{green} + b_{nir}} \tag{1.3}$$

Where b_{green} and b_{nir} are the atmosphere corrected spectral reflectance measurements acquired in the green and near-infrared bands, respectively.

1.4 Relationship of soybean canopy EVI and NDVI to yield

Many studies have looked into the relationship between soybean yield and absorption and reflectance of photosynthetic flux density by analysing the correlation of yield to the Chlorophyll levels of canopies using EVI and NDVI. In a 1999 study B. L. Ma et al. used UAVs to acquire vegetation index values and study the relationship between chlorophyll levels in soybean canopies and their yields, the study concluded that "measurement of soybean canopy reflectance at the R5 stage differentiated high from low yielding genotypes; thus, there are potential uses of canopy reflectance as a reliable, quick, repeatable indicator for ranking genotypes in screening, moreover, estimating grain yield". On a physiological level this makes sense as the soybean starts seeding at the R5 stage. Therefore their photosynthetic capacity determines their yield capacity. They used a power function to fit the correlation between soybean yield and NDVI because it gave excellent R^2 values and fitted the data distributions in the plots very well [48]. In 2018 Feng Gao et al. used Landsat-8, Sentinel-2 and MODIS to form a single constellation. They found that the optimal time window for crop yield prediction using NDVI and EVI-2 spectral indexes was by getting their values on days that ranged from 192–236 of the farming seasons according to the soybean varieties they studied. They used a linear best-fit to plot the correlation between soybean yield with both NDVI and EVI-2. Xiaoyan Zhang et al. in their 2019 study used UAV acquired Hyperspectral images and showed that "linear regression is the more suitable for plot-yield prediction in comparison with exponential and logarithm regression".

In this study, we built upon some of the concepts found in these studies and improved upon them. Unlike these studies, we monitored not only the chlorophyll levels of the soybean canopies but also other aspects of the chlorophyll level time-series profiles as measured using NDVI and EVI to come up with mathematical models that distinguish between high and low yielding, and late and early maturing varieties of soybean using remote sensing. To measure maturity periods the time between emergence and maturity was measured by observing the time it takes for the NDVI or EVI at gemination to recur at maturity. The characteristics of the time series profiles that were studied and correlated with yield include the following

- The maximum average EVI/NDVI of the vegetation index profiles.
- Gradient of the curve at ascent of the vegetation indexes
- Gradient of the curve at descent of the vegetation index
- The full width of the curve at half the maximum EVI/NDVI of the time series profile.

The virtual constellation formed in this study had a higher temporal and spatial resolution than that used by Feng Gao et al. because PlanetScope provided an almost daily (3m x 3m per pixel) imaging rate, especially in 2018 and 2019. It, however, suffered from a lower spatial resolution compared to the UAV images used by B. L. Ma et al., however, they did not image their fields over an extended time period as done in this study.

1.5 Aims and objectives

This study aims to develop a framework that analyses soybean fields using remote sensing techniques. This framework will determine the relationship between the chlorophyll levels of the soybean canopies and their resulting yield and maturity periods. It will also asses which of the two vegetation indexes is most suited for this analysis. To achieve these goals, the following are the objectives;

1. Developing an image analysis pipeline that isolates a field from a satellite image processes the image and computes its average EVI and NDVI.

2. Establishment of a process that automatically tracks the NDVI and EVI of these fields over time from crop emergence to maturity and harvest.

3. Collection of yield data of soybean fields in the areas of interest, i.e. Lusaka and Chongwe districts in Zambia.

4. Analysis of feature of NDVI and EVI time-series profiles and how they correlate with crop yield and discrimination between high and low maximum potential yields of varieties under study.

5. Determination of whether the relationship between yield and satellite-derived canopy reflectance measurements of soybean can be used to predict traits of new breeding lines.

6. Ranking of the soybean varieties under study according to yield potential, stress resilience and maturity period as observed remotely compared to experimental data.

Chapter 2

Optimisation of the Image Analysis Pipeline

2.1 Introduction

An image analysis pipeline was developed to load, process and analyse satellite images of soybean fields around Lusaka and Chongwe districts in Southern Zambia. It was optimised to do so at high spatial, temporal and spectral resolutions. These high resolutions were useful in separating the fields from the rest of the surroundings accurately in order to capture the physiological condition of the crops in them throughout their growth period with accurate detail using the multi-spectral images.

For this pipeline to achieve a high temporal resolution, a virtual constellation was created by using multiple satellite constellations together to image the soybean fields. PlanetScope (Dove), Sentinel-2 and Landsat-8 satellites were used in consort to improve the revisit time to as many as 100 days out of the 185 days that were designated as the observation period for each farming season observed in this study. Most soybean varieties have maturity periods ranging between 90 and 150 days so that the 185 days observation period was enough to track the crops from a few days before germination up to a few days after maturity. To harmonise the data obtained from these different satellites, their pixel values were radiometrically and atmospherically corrected using QGIS (Sentinel-2) or downloaded in that form (Landsat-8 and PlanetScope) and were normalised using python to run between 0.0 and 1.0, i.e., they were converted from Digital Numbers (DNs) to reflectance values. This was done by converting the DNs to radiances using QGIS or downloading images that were already converted. The radiance pixel values were then divided by 10,000 to obtain reflectance values. Due to the immense size of each of the Satellite images and their large numbers needed to analyse a single field, the image analysis pipeline needed a way to analyse large amounts of data quickly and accurately. Therefore, it was automated using computer programming and machine learning techniques to preprocess and analyse the data faster at a fixed high standard, and to avoid human bias.

The programing language that was chosen for this task is Python. It is very well suited for high volume, velocity and veracity data management and manipulation [61]. Although its performance in computationally intensive tasks is slower compared to lower-level programming languages, the development of NumPy for numeric vector and matrix operations, pandas for Dataframe manipulation and Scikit-learn built upon Gfortran and C for general-purpose machine learning algorithms has catapulted Python to be the most popular programing language for data science in the world. Its use in this study made debugging and overall programming much easier due to its easy to read syntax and the many readily available Python-support sites on the web [62].

The pipeline followed the following course; (i) image acquisition, (ii) Field clipping, (iii) Cloud masking and vegetation index computation, (iv) Generating time series plots of the average vegetation index for each field, and (v) Saving observed values of the features of the vegetation index time series plots for them to be analysed for correlation with the yield ,and to maturity period of fields with the same soybean variety. The vegetation indexes that were observed in this study are; Normalised Difference Vegetation Index (NDVI), and Enhanced Vegetation Index (EVI). The NDWI (Normalised Difference Water Index) was only used as an additional spectral index for training the machine learning algorithm that distinguished cloud and shadow infested pixels from desired ones. This satellite image analysis pipeline can used by extension officers and soybean breeders in Sub-Saharan Africa to come up with creative solutions when monitoring crops and solving various complex challenges in their work. It would be particularly useful in situations were they encounter great logistical hurdles such as those experienced due to the COVID-19 global pandemic.

2.2 Image Acquisition

Data required for this research was obtained through online data pools at www.planet.com, for Dove (PlanetScope) satellites, and the United States Geological Survey's www.earthexplorer.usgs.gov website for both Sentinel-2 and Landsat-8 images. These images can be acquired in an automatic way using the various Application Programming Interfaces (APIs) of these websites. However, for simplicity, they were obtained manually in this study by requesting for them on the Graphical User Interfaces (GUIs) of the respective websites. This involved entering the geographical coordinates or name of Lusaka or Chongwe districts according to the district where the fields of interest were into the prompt part of these websites and the timeline of the raster images required and choosing the desired rasters according to cloud cover levels. Images that were entirely obscured by clouds were left out as they cannot be used in vegetation index analysis. Satellite image analysis was done in Python using the Rasterio (Raster Input Output) and GDAL (Geospatial Data Abstraction Library) modules. These provided a secure means of offloading geospatial data into the NumPy and SciPy modules and thereby made the raster analysis process much more manageable. Rasterio is simply GDAL wrapped in Python. The pythonic syntax makes it easier to use at the expense of some speed. GDAL is written in C. This means its faster, but the fact that rasterio has pythonic syntax results in less code that is easier to read and debug [63]. Rasterio and GDAL were used interchangeably in this study according to whether speed or easy readability was desired for any particular task.

Satellite multi-spectral images are different from conventional images in the sense that they come with geographical metadata and have a different dynamic range (radiometric resolution), i.e. most conventional images have pixel Digital Numbers (DNs) with an 8-bit bit-depth so that their DNs range from 0, obtained from $2^0 - 1$, to 255, obtained from $2^8 - 1$. While satellite images usually have DNs of a 12-bit, 14-bit or 16-bit bit-depth so that in the case of a 16-bit dynamic range, their DNs range between 0, $(2^0 - 1)$, and 65,535, $(2^{16} - 1)$. They also come either as one image with multiple spectral bands (more than three bands) superimposed into one as was the case with Dove's PlanetScope images that were used in this study or as individual greyscale images of different spectral bands which was the case for both Sentinel-2 and Landsat-8 images used in this study. The Sentinel-2 images used in this study came as 13 separate JPEG 2000 (JP2) format greyscale images, each of which was a different spectral band and needed atmospheric correction, i.e. they were not surface reflectance images. Their attributes are shown in table 1.1 in chapter 1. For their atmosphere correction, the Semi-Automatic Classifier Plugin (SCP) in Quantum Geographic Information System (QGIS) was used. QGIS is open-source Geographic Information system software with readily available versions in many operating systems including

Linux and Windows, which were the operating systems used in this project. PlanetScope images came as GeoTiff format surface reflectance images that are of the superimposed type, their properties are shown in table 1.2 in chapter 1, their surface reflectance product is provided as a 16-bit GeoTIFF image. Their radiance values were scaled by 10,000, after which their pixel values became reflectances ranging between 0 and 1. Landsat-8's images come as 11 greyscale surface reflectance images of GeoTiff format, each of which was a separate spectral band as shown in table 1.3 in chapter 1. They also needed scaling by 10,000 to get reflectance values, whereas those of Sentinel-2 got scaled during the atmosphere correction process in QGIS. The images of Sentinel-2 and Landsat-8 were downloaded on a mass scale by simply providing the period of observation and coordinates of Lusaka city into Earth explorer while PlanetScope images required proper inspection during the downloading process to account for cloudy images in which the field was still fully or partially visible. This was done to minimise the number of redundant images downloaded, as www.planet.com had a $10,000km^2$ cap on the number of images that could be downloaded from their website per month using the research account that was made for this study.

To optimise the process of vegetation index analysis of the crops in the fields, they were observed from 20^{th} November, which coincided with the beginning of the rainy season and with it the soybean growing season in Lusaka and Chongwe districts, to 20^{th} May the following year which was close to harvest time. For most soybean fields in the districts under study, the crops would still not have been harvested at this time of the year. The period from December to January is usually the peak of the rain season in these two districts and is therefore, primarily cloudy. Which made it challenging to observe the fields during this time as many fields were completely overcast by clouds and their shadows. Planet Lab's Dove constellation of over 120 satellites was able to capture some cloud-free patches of a field of interest even during this time because of its rapid revisit rate and higher spatial resolution. Coupling this with Sentinel-2 and Landsat-8 constellation gave an excellent temporal resolution and provided some data even in this very hard to observe period.

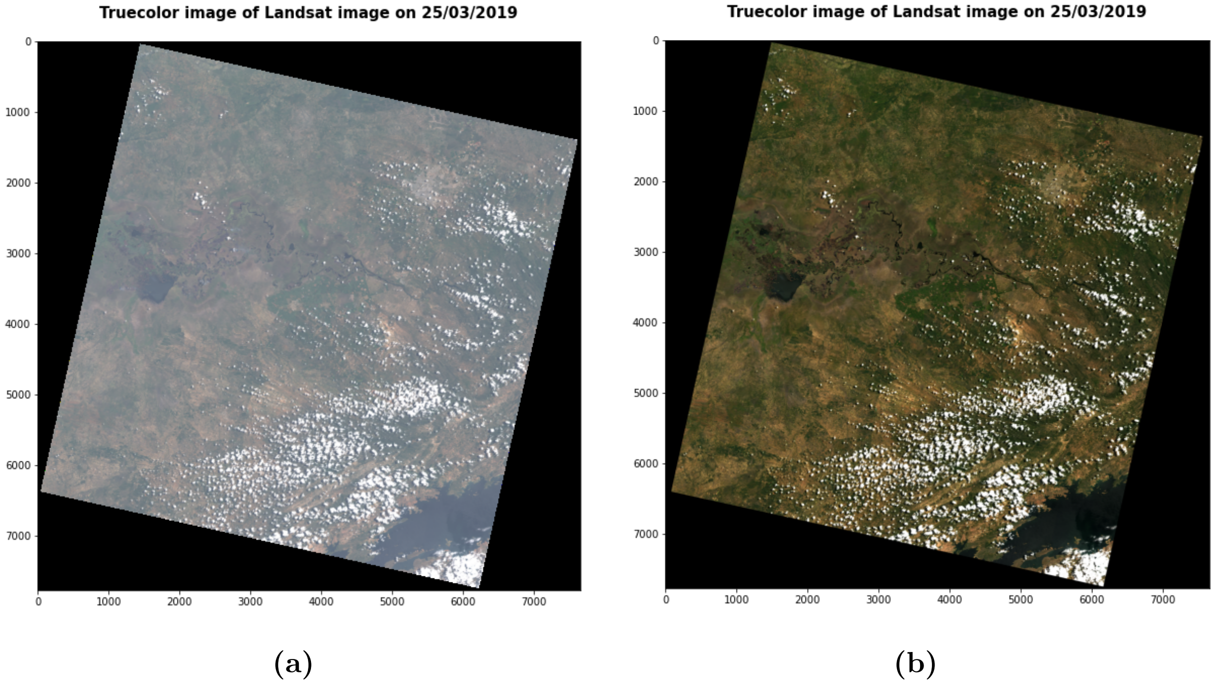

Truecolor image of Landsat image on 25/03/2019
0
1000
2000
3000
4000
5000
6000
7000
(a)
Truecolor image of Landsat image on 25/03/2019
(b)

Figure 2.1: **(a)** top of the atmosphere Landsat-8 color image and **(b)** bottom of atmosphere corrected surface reflectance (scaled by 10,000) Landsat-8 image of the same place as **(a)**. This image encompasses much of Southern province and parts of Lusaka province in Zambia.

Clipping the field automatically from each of these satellite images was one of the primary challenges that needed to be overcome in order to automate the whole pipeline. Sentinel-2 and PlanetScope satellite images used the epsg:32735 cartographic Coordinate Reference System(CRS) while Landsat-8 used the epsg:32635 CRS, this meant that cartographic coordinates acquired from the first two constellations to clip the field from them would not work in the later [64]. So the CRS of the shapefile uses to clip the field needed to be adopting the CRS of the host image. Only orthorectified images were used in this study. Orthorectification is the process of removing the effects of image tilt and terrain effects, thereby creating a planimetrically correct image which has a constant scale, this is vital for calculating accurate average vegetation indexes of an area.

2.3 Field clipping

In a geospatial analysis, image clipping is the process of separating a small part of an image (the area of interest, AOI) from the rest of it so that it can be analysed separately. This is especially useful when trying to separately analyse the vegetation index pixel values of a particular variety of crop to prevent contamination from other plants in the surroundings, and helps make sure that the area scanned is consistently representative of the canopy coverage of the crop we are interested in studying regardless of the spectral band or satellite image being used. There are two ways in which clipping can be done on a satellite image. One such way involves making a shape in a Python code that coincides with the boundary of the field. This can be very tedious and time-consuming. The other method, which is more effective and suited for automation and was therefore used in this study, involves creating what is called a shapefile of the field in a Geographic Information System (GIS) using a raster containing the field we are interested in creating the shapefile for. In this project, shapefiles were made using QGIS. The shapefile was created by drawing lines along the boundary of the field until the field was entirely encased in the resulting polygon. Upon its creation the points on the polygon adopted the CRS coordinates from the raster. Therefore, during the creation of the shapefile, it adopted its CRS from the raster from which it was made. For accuracy, all shapefiles used to clip soybean fields were created using PlanetScope rasters because they had the highest spatial resolution of the three constellations; 3 meters per pixel. This spatial resolution made it easier to see the boundaries of the fields during the shapefile creation process. Hence, the CRS of the shapefiles created in this way was always epsg:32735, the CRS of PlanetScope rasters. The procedure used to create shapefiles in QGIS for this study is presented in subsection 2.4.4 of this chapter.

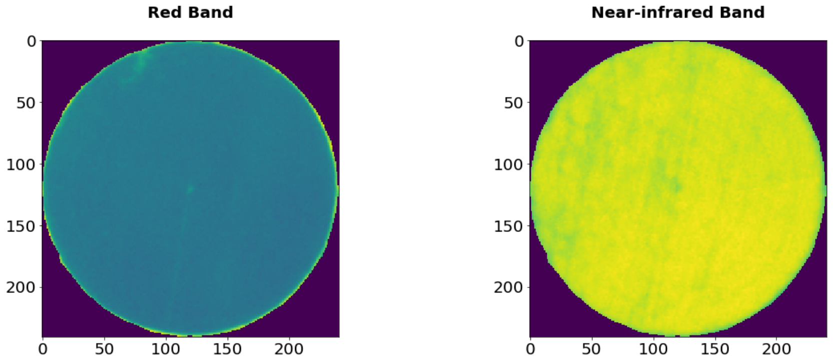


Figure 2.2: Clipped red and near-infrared spectral gray scale bands from a PlanetScope image of a soybean field with a vibrant crop growing in it; noticeable is the brightness of reflected near-infrared and the dimness of the absorbed red light. This is a sign that the ground is covered up by a chlorophyll-rich plant canopy since chlorophyll reflects near-infrared and absorbs red light.

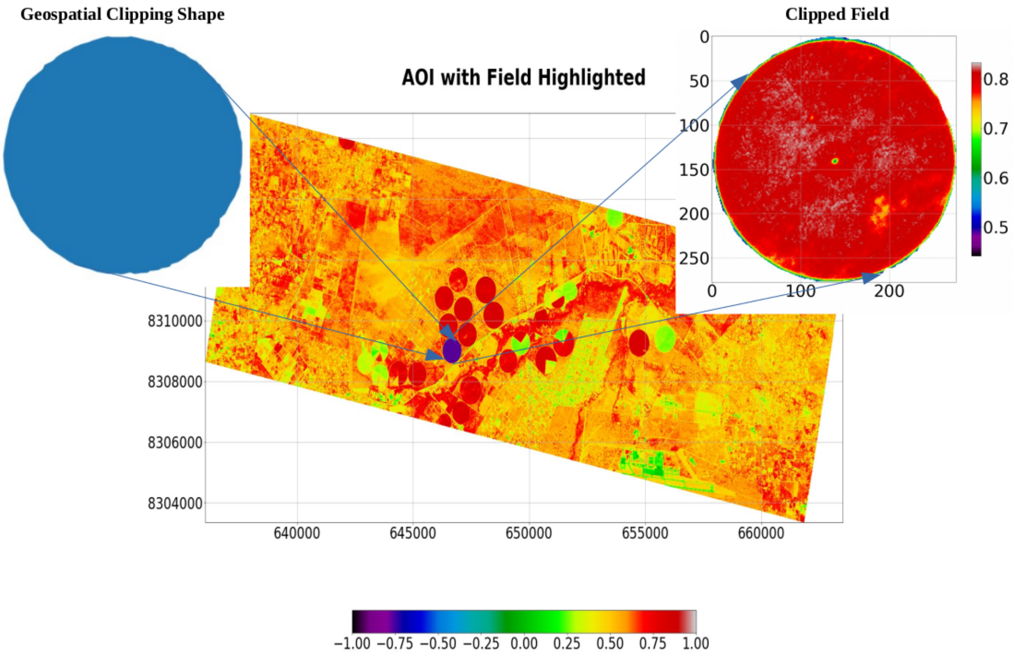


Figure 2.3: Illustration of the clipping of a soybean field of interest from a PlanetScope image using a shapefile. The shape was made in QGIS in the form of a polygon whose points are referenced to the cartographic coordinates of the field as it appears in the raster. Notice that the vegetation index scale of the entire image runs from -1.0 to 1.0 while of the clipped field's scale has been zoomed into to highlight the vegetation index values that are more relevant to it; 0.4 to 1.0 of the vegetation index scale.

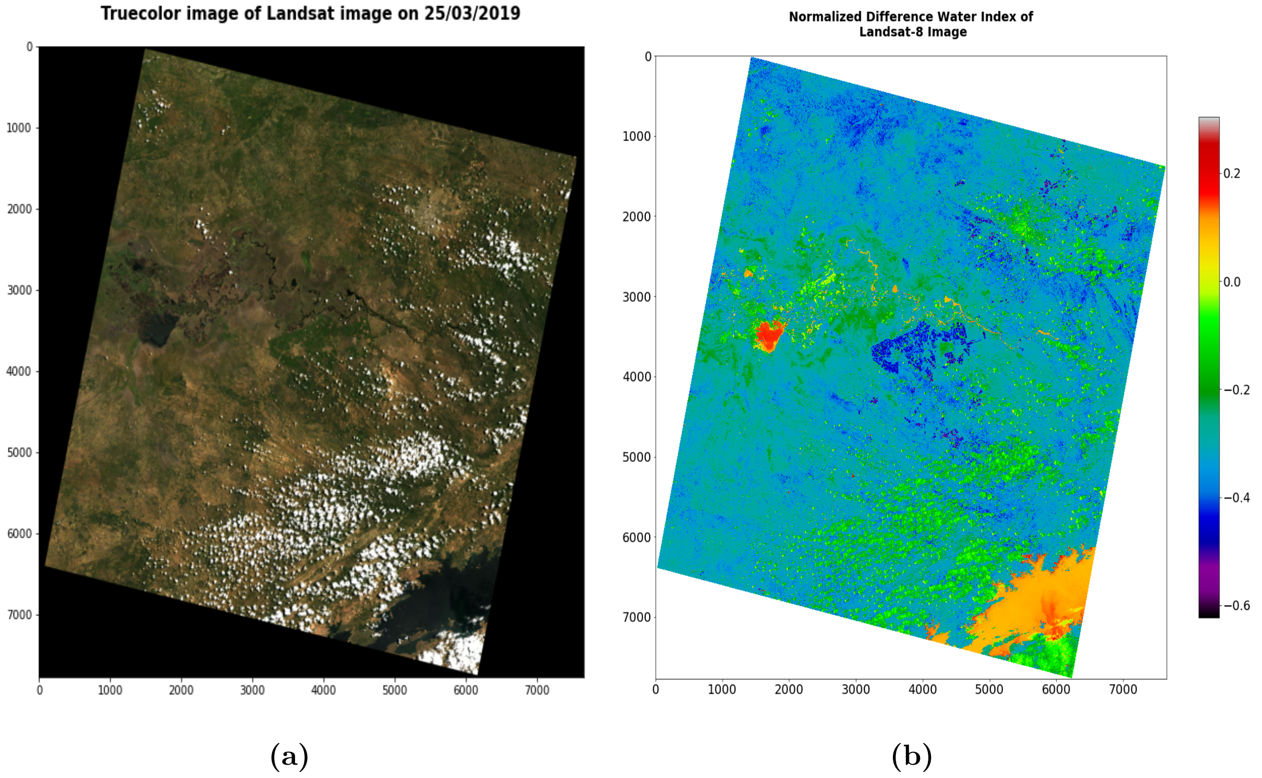
In order to analyse fields in Landsat 8 images, the shapefile's CRS needed to be changed to epsg:32635, which is the CRS of Landsat-8 rasters, this was done using the OSGeo module in Python. Once the shapefile was ready, it was loaded into the Python program developed for this pipeline using the Geopandas module. GeoPandas extends the data types used by the Pandas module allows for spatial operations on geometric types such as shapefiles. It loaded the shapefiles into Python in the form of a data frame. Matplotlib and Descartes modules were used to display the shapefiles in Jupyter notebook. When making the clipping shape in QGIS, great care was taken to not include areas surrounding the field because this had the abject effect of bringing in vegetation index values from plants in the field's surroundings and thereby contaminating our readings. This is because the plants in the surroundings of the field could have been in a different condition from the soybean in the field. They were likely to be of a different species of plant with a varying vegetation index from that of the soybean variety in the field. They could also be at different growth stages with the soybean. At the beginning of the farming season when the crops in the field were still young (most of the ground in the field was not covered), the shrubs next to the field could have been fully grown and chlorophyll-rich, so that getting the maximum vegetation index from a satellite image of such a field gave the NDVI value of the shrubs instead of the crops in the field. To reduce this contamination, the shapefiles were made using an image where the crops in the field were fully grown to make boundary identification easier. The geometry type that was used when making shapefiles in QGIS is a polygon, this allowed for the creation of a shapefile for almost any field including circular ones that may have had some irregularities at their boundaries due to problems with the irrigation pivots [65]. Since the field clipping shapefiles were made using PlanetScope images since they had the best spatial resolution and captured the boundary of a field in the best detail, the boundaries became a little less accurate in Sentinel-2 and even less so in Landsat-8 images owing to the pixel sizes becoming bigger in these satellites; 10m per pixel and 20m per pixel, respectively. This means more of the plants outside the boundary were captured with the field's clipped image in Sentinel-2 and Landsat-8 images

2.4 Cloud masking and Vegetation Index computation

Clouds can entirely or partially obscure a field in a satellite image, and the shadows they cast can significantly contaminate the vegetation index of the crops. To reduce this contamination, it became imperative to remove the clouds and their shadows from partialy obscured field images. To assist with this, satellite images come with quality assurance files that can be used to crop out untrustworthy pixel values, including clouds and their shadows. For example, PlanetScope images come with a file called Unusable Data Masking (UDM), which can be used to mask clouds, shadows and haze. When these UDM files were used to mask unwanted pixels in this study, it was found that the file included light haze in its masking process which led to losing even images that could be salvaged for usable data. Another challenge was the fact that each separate constellation of the satellites used in the virtual constellation has its cleaning/masking procedure and thus would have required much time to understand and implement. For these reasons an independent masking process was developed for this project. This cloud masking process made use of a machine learning algorithm called Random Forest Classifier (RFC) to determine which pixels of the image were occupied by clouds or shadows and which were not. Random Forest Classifier uses an ensemble learning method, i.e. it is made of many decision trees; this makes it a formidable too for use both in classification and regression kind of problems. The Python module Scikit-Learn was used to create, train, and test the classifier used in this study. In a bid to optimise the classification method, a model was created for each satellite so as to avoid any misclassification that could result from using one classifier for all the satellites due to their different pixel sizes. After the pixels were classified by the RFC, the Python code masked out the undesired pixels using the Earthpy and Numpy modules by forming a Boolean array of the image such that the undesired pixels attained Boolean False values, while the desired pixels attained Boolean True values, and the multiplication of the Boolean array with the image array resulted in the pixel values that coincided with the Boolean False values being masked out.

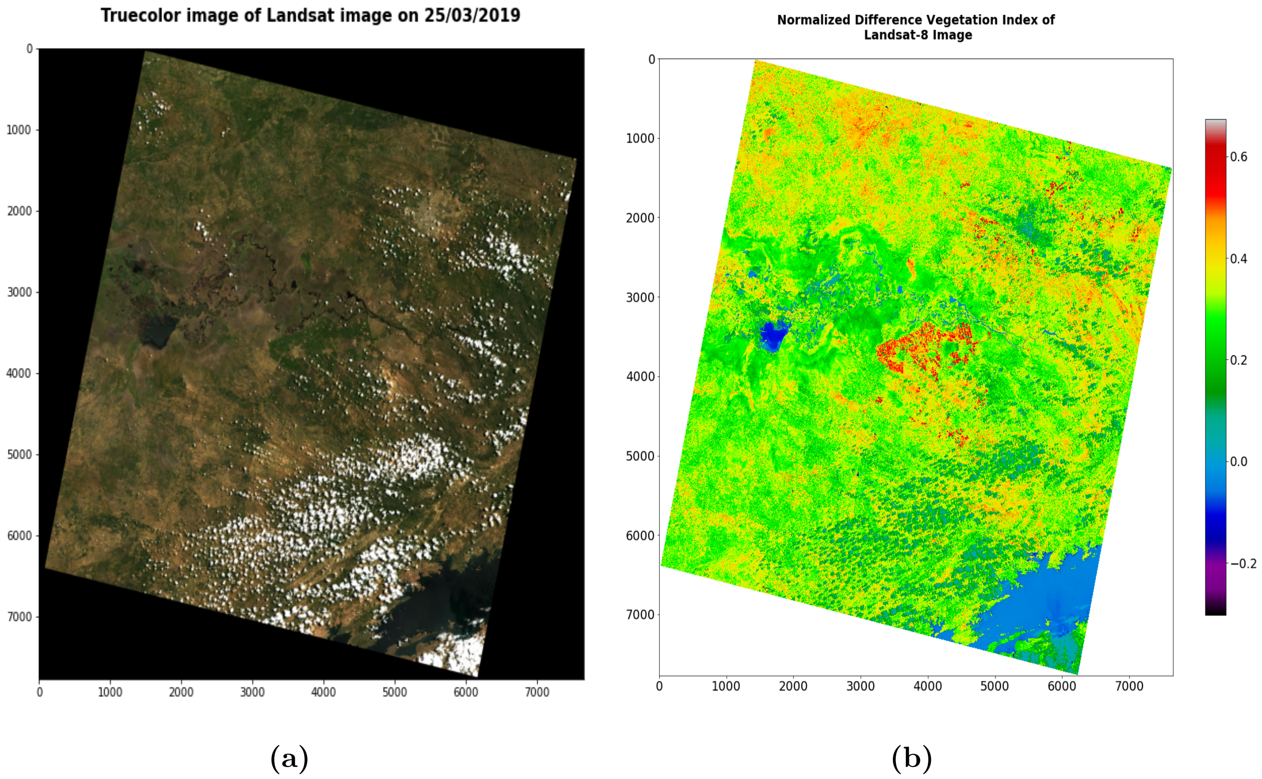
2.4.1 Feature Engineering

In machine learning algorithms such as Random Forest Classifier, a feature is a numeric representation of certain aspects of raw data sets, the algorithms take in these features as inputs for training and modelling. In the case of this study, features are spectral bands and spectral indexes. Features are vital in the creation of a classification model. Feature engineering is the act of extracting new data forms from pre-existing data. Therefore, the computing of spectral indexes from pre-existing spectral bands is a form of feature engineering. The engineered features are transformed into formats that are suitable for the machine learning pipeline. Feature Engineering is crucial to the process of training a Random forest classifier because the right features can ease the process of creating a model [66]. As already explained the electromagnetic spectrum of sunlight has different amounts of each wavelength, and each of them interacts in different ways with the various objects with which they come into contact. For example, many types of clouds absorb a significant amount of near-infrared and therefore make shadows stand out in a near-infrared image. At the same time, plants re-emit the infrared falling on them, and they also stand out a lot more in near-infrared images. Blue has the smallest wavelength of the visible spectrum and so is easily scattered than green and red. Green is the most abundant wavelength in the visible spectrum and since it is less easy to scatter a significant amount of green makes it through the upper atmosphere to reach the clouds. For this reason, the contrast between cloudy and non-cloudy parts of a green band image is extreme. Red and blue images seem to be scattered by almost the same amount. These facts were used to make the classifier more adept by engineering two new features to complement the red, green, blue and near-infrared features that were already available; the Normalised Difference Vegetation Index (NDVI) and the Normalised Difference Water Index (NDWI). These were computed using equations (1.2) and (1.3) respectively.


Truecolor image of Landsat image on 25/03/2019
0
1000
2000
3000
4000
5000
6000
7000
Normalized Difference Water Index of Landsat-8 Image
0.2
0.0
−0.2
−0.4
−0.6

(a) (b)

Figure 2.4: Bottom of atmosphere corrected true color image stacked next to an engineered feature; NDWI (Normalised Difference Vegetation Index). NDWI differentiates water bodies from the rest of the surroundings. Notable is how lake Kariba stands out in bright orange in the lower right corner of the NDWI image.


Truecolor image of Landsat image on 25/03/2019
0
1000
2000
3000
4000
5000
6000
7000
Normalized Difference Vegetation Index of Landsat-8 Image
0.6
0.4
0.2
0.0
−0.2

(a) (b)

Figure 2.5: Bottom of the atmosphere corrected true color image stacked next to an engineered feature; NDVI (Normalised Difference Vegetation Index), NDVI is good for differentiating high level chlorophyll crops from the rest of the surroundings as can be observed from the bright red color of the sugarcane fields at Nakambala in Kafue district in the center of the NDVI image.

The four spectral bands (red, blue, green, and near-infrared) were used together with the two spectral indexes (NDVI and EVI) in training the machine learning model in order to make use of certain land cover classes that stand out very well in them. Having this many features gave the algorithm a lot of data to use when training the classification model. The spectral indexes helped the algorithm to more accurately identify certain features, thus giving output results that are of high quality. In the new features (spectral indexes), plants stood out the most in the NDVI raster as can be observed in figure 2.4, and water bodies stood out in the NDWI raster as can be observed in figure 2.5. By making these land cover classes stand out the engineered features also made it easier for the RFC to classify them.

2.4.2 Land cover class extraction

In order to create a model for classifying pixels, land cover classes, e.g. clouds, shadows etc. needed to be concatenated into a feature vector that could be used for training and testing the classification model. The feature vector comprised of seven dimensions, six of which contained the pixel values of the four spectral bands and the two engineered features. The seventh one contained class numbers designated to each landcover class using ground truth data of the pixel values of each class. Pixels fitting a certain land cover class, e.g. clouds, shadows and crops were fed into the algorithm by clipping them from the feature images using the procedure illustrated in subsection 2.4.4. The shapefiles were given attributes that identified them as belonging to a certain landcover class. Care was taken to make sure that pixels labelled as belonging to a certain class did not overlap pixels from other classes, e.g. pixels of the landcover class “water” did not have pixels of another class, e.g. “crops” among them. Once all the features (different spectral bands and reflectance indexes) were ready, the various pixels of Sentinel-2 and Landsat-8 images were extracted from them using shapefiles made in QGIS according to three classes; “clouds”, “shadows”, and “data”. Pixels of PlanetScope were extracted according to two classes, “no data”, standing for shadows and clouds and “data” standing for the pixels of the field that were important to the vegetation index analysis. To improve the number of different classes available in a PlanetScope raster four rasters was merged together in QGIS to form a bigger raster containing many more pixels and land cover classes to sample from. The rasters were chosen in such a way that they belonged to different times of the year so that they possessed different land cover types, particularly different types of clouds, shadows bare land and crops. When extracting the land cover classes, great attention was given to pixels of clouds and shadows by getting more of them of different types when sampling (clipping) so as to enhance the model’s ability to classify them. Cross-validation was applied by leaving 25% of all samples randomly so that after training on the training set (75% of samples), the model was tested on this prediction set (25% of samples). Independent validation was performed using a different raster.

To save on time, the classification models so formed were saved using a Python module called Pickle, so that they were loaded whenever they were needed for classifying other images of the respective satellites. This ensured that there was no need to train the model every time it was needed. When the classification model was ready, and the field was clipped from the rest of the image, clouds were masked by doing away with pixels classified as either clouds or shadows for Landsat-8 and Sentinel-2, or no data, for PlanetScope images.

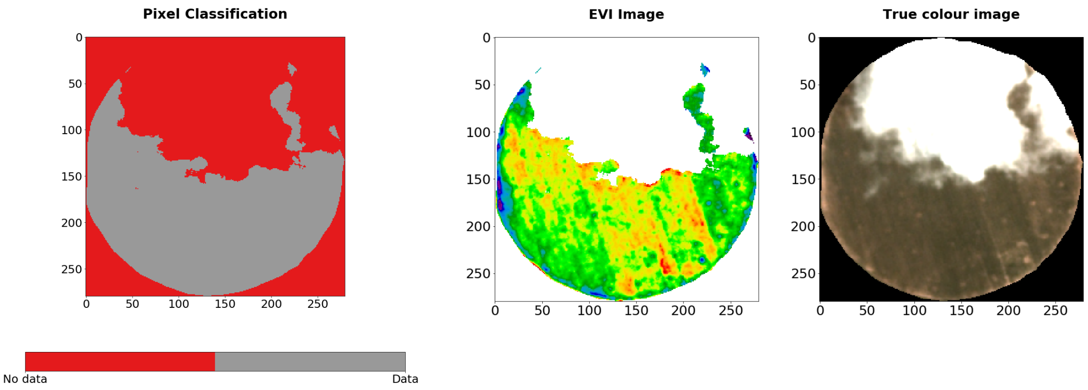


Figure 2.6: Cloud, shadow and no data masking in an EVI image by a pixel classification developed using random forest classifier. The edges of the clouds still have a some clouds left due to complication when sampling them to train the model.

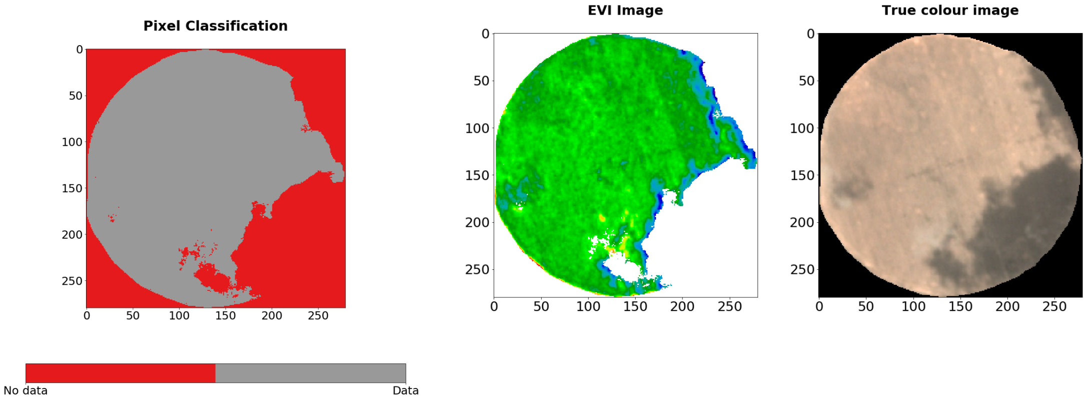


Figure 2.7: Masking of "Shadow" and "no data" values in an EVI image using a pixel classification formed using Random Forest Classifier, the shadows in this image were accurately classified by the model, but it still left out some clouds at the edges. This is due to the complicated process of sampling the edges of clouds accurately for training the model.

After the cloud and shadow masking, the average indexes (NDVI and EVI) of the field were computed by summing the values of the pixels in the area and dividing the result by the number of pixels that were left in the area. After a field was analysed for a single day, the resulting average vegetation index and corresponding date of observation were appended to a Pandas dataframe containing the compounded daily average values of that index for a particular field and that particular variety of soybean.

Once all average vegetation indexes of the days of observation of a field were appended to a dataframe, a graph was plotted to observe the growth trend of the chlorophyll levels of the crops with time. The abscissa of the time-series graph so obtained contained the dates of observation of the satellite images. The ordinate contained the average vegetation index values of the field corresponding to the observation dates. Cubic Spline and Gaussian Process models were used to plot the growth trend of the chlorophyll levels of the field. Figures 2.6 and 2.7 show the pixel classification array and the subsequent masked EVI and true colour images of a field. Notice that the edges of the clouds are not completely well masked. This resulted from the complication of trying to save more fields, i.e. there were fields that had light haze in them and were still analysable. Training the algorithm in such a way that it cropped the edges of the clouds completely clean resulted in a loss of such fields, and so in order to prevent this, the edges of clouds that did not coincide with a cloud shadow were not used to train our models.

2.4.3 Procedure for Atmosphere correction of Sentinel-2 rasters in QGIS

Since the Sentinel-2 rasters needed to be atmosphere corrected before they were ready for use in the vegetation index analysis. The following is the procedure for their atmosphere correction using QGIS.

1. If not already installed, install Semi-Automatic Classification Plugin in QGIS. To do this:
   - Open QGIS.
   - On the top menu select **Plugins** > **Manage and Install Plugins**.
   - In the search bar that appears, search for the **Semi-Automatic Classification Plugin** and click on it to start the installation.
   - Wait for the installation to end. Restart QGIS.
2. On the top menu of QGIS select **SCP** > **Preprocess** > **Sentinel-2**.
3. In the prompt that opens, find the Directory containing the rasters to be corrected in the search bar entitled **Directory containing Sentinel-2 bands**.
4. Select the meta data file in the search bar entitled **Select metadata file (MTD_MSILIC)**.
5. Tick **Apply Dos1 Atmosphere correction**.
6. Click on **Run** and provide a directory into which the Atmosphere corrected rasters should be written.
7. Wait for the process to finish running

2.4.4 Procedure for making shapefiles for field and feature extraction in QGIS

1. Open the georeferenced image that contains the features or field of interest in QGIS.
2. Select **Layer** > **Create Layer** > **New Shapefile Layer** to create the new empty layer for the vector feature. A prompt appears asking to confirm the Coordinate Reference System (CRS) for the new shapefile layer, if no new CRS is provided, QGIS by default gets the CRS of the image from which the shapefile is made [67, 68].
3. A prompt then appears to confirm:
   - The type of feature that needs tracing, these can either be **Points**, **Lines** or **Polygons**. To clip the fields and landcover features used in this study; **Polygon** was selected.
   - The name of the attribute that is being traced, in the case of feature extraction for the machine learning model, this was either "Clouds and Shadows" or "data". In the case of field clipping the name of the field was given here.
   - The type of the attribute, eg. **Text** and **Whole Number**, this was left blank when clipping a field and was given a **Whole number** type when clipping features for the cloud and shadow classifier. Clouds and shadows shapefiles were given a value of 1, while data shapefiles were given a value of 2.
   - The name to save the shapefile layer as.
4. Use the **Toggle editing** button to add or edit features. Another way is to right-click on the shapefile layer in the table of contents and select **Toggle editing**.

5. The polygons around the feature of interest are captured by clicking on the **Add feature** button and then clicking on a set of points along the perimeter of the field or the feature that needs clipping.

6. Once the final point is reached the polygon is completed by right-clicking and providing an ID number in the prompt that appears. After this, the project needs to be saved by clicking on **Save Layer Edits**.

7. In the case of fields, to finally clip the field's shapefiles, click on **Vector** > **Geoprocessing Tools** > **Clip**.

8. In the case of the features for the cloud and shadow classifier, the final clipping is done by clicking on **Vector** > **Geometry Tools** > **Multiparts to singlepart**.

2.4.5 Image histogram

An image histogram shows the distribution of pixel Digital Numbers (DNs), reflectance values or spectral index values by showing how frequently each value occurs in the image. Thus, in this study, image histograms were very useful in monitoring how the vegetation index values of the soybean canopies in the fields were distributed in their images. When creating image histograms of fields, the total range of the pixel value dataset from minimum to maximum vegetation index was divided into 30 equal intervals. These equal parts are known as bins or class intervals. The frequency of each of these intervals was plotted on the histogram. Since the vegetation index images were given a nipy_spectral colour scheme, looking at the histogram of the image was enough to judge the entire tonal distribution at a glance and hence also judge the growth stage of the soybean crops in it. If the histogram values were concentrated toward the left of the histogram, it meant that the crops were young and the image was more green in the nipy_spectral colour tone. If they were concentrated in the middle, it meant that the crops were larger more chlorophyll rich and covering more ground and therefore the image was more yellow in tone. If they were concentrated toward the right, it meant that the canopy had covered allmost the entire ground with a chlorophyll rich biomass, and the image was red in the nipy_spectral colour tone. Figures 2.8 and 2.9 show two growth stages of soybean; the vegetative and reproductive stages and the histogram produced for each stage.

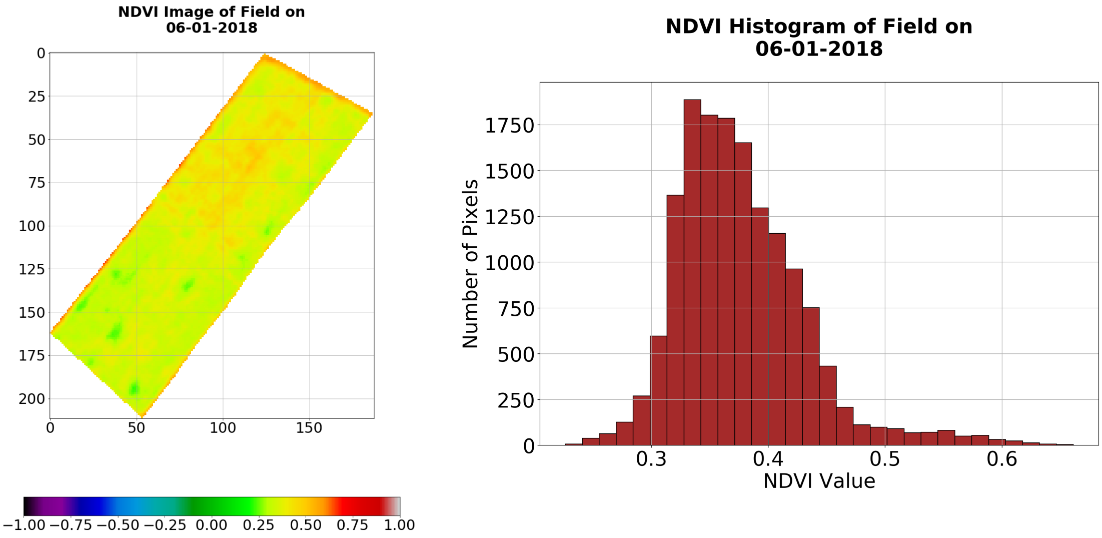


Figure 2.8: Image and histogram of a Field at the start of the farming season (vegetative stage). Due to imperfections in field clipping some pixels have larger NDVI values from the vegetation in the area surrounding the field.

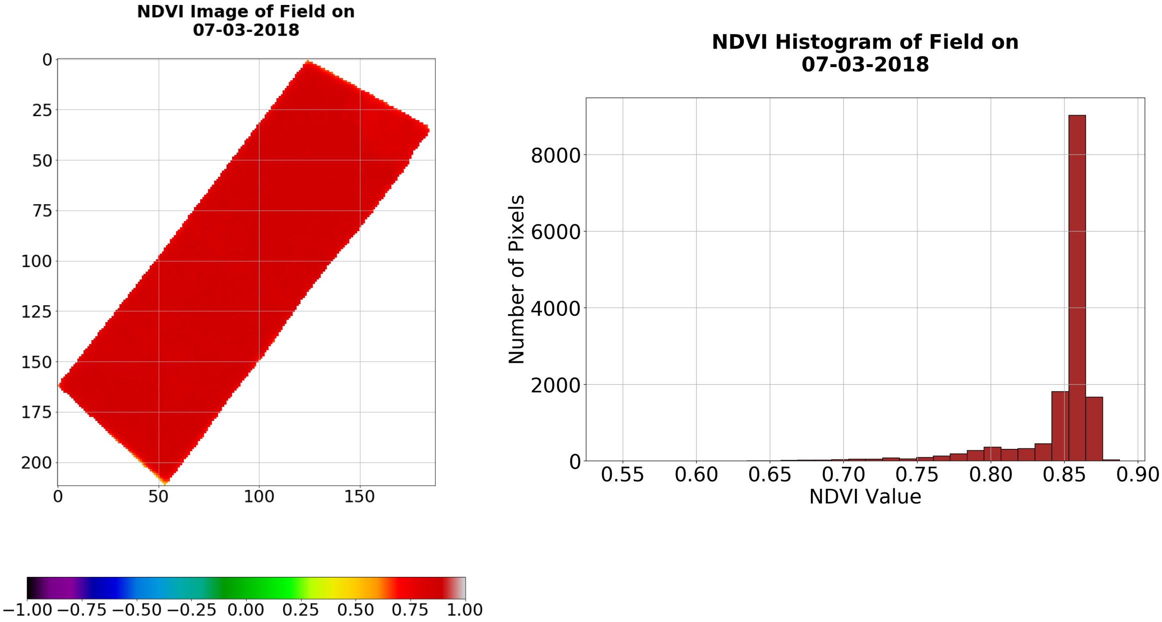


Figure 2.9: Image of a field at the peak of the Soybean's reproductive stage, the crops in the field have matured and the canopy has the highest level of chlorophyll concentration it will attain during the entire farming season. The Histogram shows that more than 8000 of the pixels have NDVI values greater than 0.85.

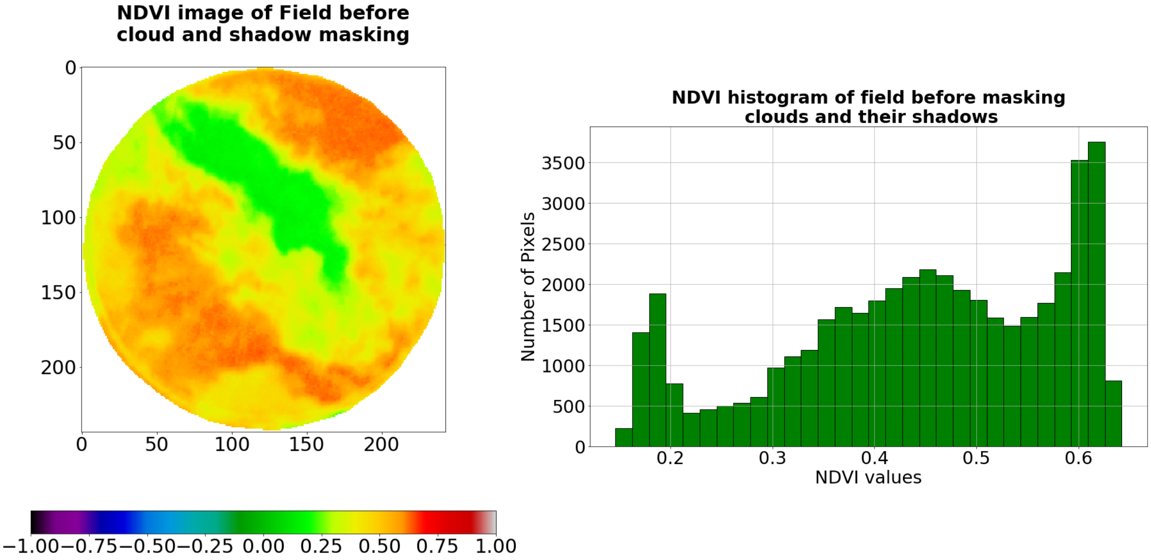


Figure 2.10: NDVI image and histogram of a field on a cloudy day without applying a cloud mask. The contamination of the pixels by cloud and shadow values results in a low average NDVI of the field.

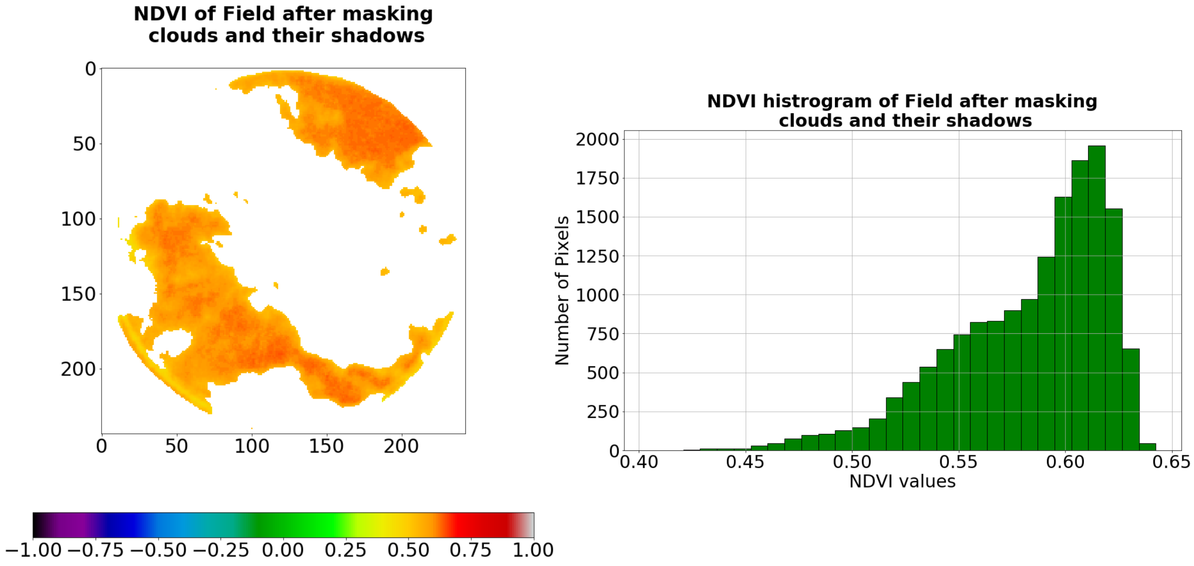


Figure 2.11: NDVI image and image histogram of a field after masking clouds and their shadows.

Histograms were helpful in refining the cloud masking process by showing how well undesired pixels were masked from the field's image, figures 2.10 and 2.11 show such a scenario. The unmasked image's histogram shows much lower NDVI values which are caused by the presence of clouds, whereas the histogram in 2.11 shows that most of the lower NDVI values are not included in the image after masking. Histograms can also be used in observing how well the image's pixel values have been distributed in the process of image histogram equalization.

2.5 Curve fitting of vegetation indexes against time

Curve fitting models were used to track the photosynthetically active vegetation per unit area of the canopy of the soybean crops in the fields in the form of average EVI and NDVI per m^2. They also helped to interpolate where data was missing and so were very essential in determining characteristics of the growth curve such as the peak average NDVI and EVI, area under the curve, the gradient of the increase or decrease in chlorophyll concentration. They were useful in determining when the crops germinated, matured, and dried up in readiness for harvest, all without need to visit the fields in person [69].

The plotting of the average vegetation index values of the soybean fields against the days of observation (which coincided with a particular farming day for each image) was done for all fields for which ground truth data was obtained during the placement with IITA Zambia. Analysing such a complex system as plant growth and its relationship to yield requires nonlinear models to be able to capture the accurate trajectory of the data points and eliminate the noise caused by atmospheric phenomena, clouds, and ground conditions. What was needed are curve interpolation methods that account for an environment where the data is dynamic, noisy and observation is costly and therefore, sparse. Before the data was plotted, it needed some preprocessing to remove Nan values and any recurring date values since some plotting algorithms will not accept data points that have recurring x-axis values. Two models were chosen to plot the best-fit curves of the average vegetation index versus time plots in this study. These are cubic spline Interpolation, and Gaussian process modelling. They were chosen for their fame at accurately interpolating and predicting data distributions of the sort produced in this study according to many sources [69, 70, 71, 72, 73]. This did not exhaust the types of models that can be used to achieve this goal. Other notable models that can be used include, convolution, double logistic, Fourier series, and polynomial fitting [69]. In both cubic spline interpolation and Gaussian process modelling, Simpson's algorithm was used to determine the area under the curve so that it was included among the factors of the curve that were analysed for correlation with soybean yield.

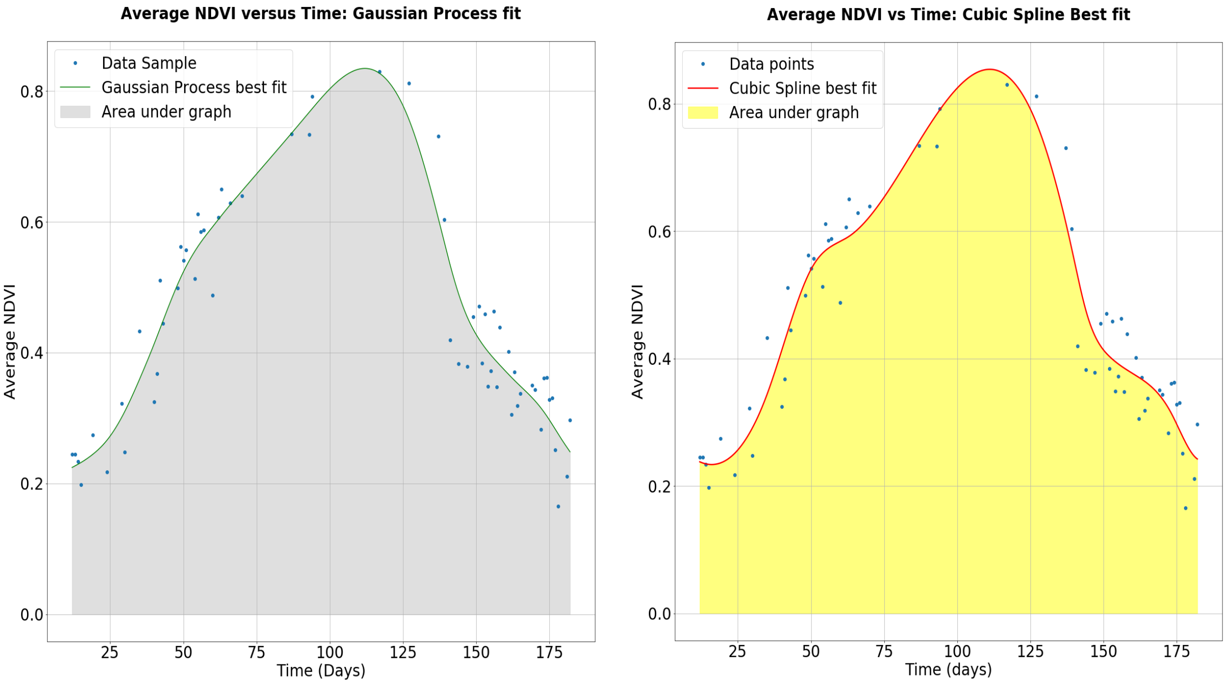
2.5.1 Cubic spline curve fitting

A cubic spline is simply a piece-wise cubic curve with a continuous second derivative. The curvature at any point of the curve depends on the second derivative there; it is best used for fitting curves were increasing the order of the polynomial does not significantly improve the curve fit. It is plotted by using a different polynomial curve between each two data points, i.e. cubic spline is made of different cubic curves attached to each other. The coefficients on the cubic polynomials are the weights used to interpolate the data. They bend the fitting line so that it passes through each of the data points without any erratic behaviour and breaks in the continuity. Cubic splines can closely approximate the actual function of data points without a high degree of divergence from it [69]. The most complicated part about cubic spline interpolation is determining how much detail it must capture, that is how much fitting is over-fitting. In creating a cubic spline best fit in Python, the s value was placed at 0.12. This value was chosen because it gave the best fit to the data distribution of the NDVI and EVI versus time data plots without too much overfitting nor underfitting. In this study, the cubic spline interpolation model was implemented using Python's scipy.interpolate module.

2.5.2 Gaussian Process Modelling

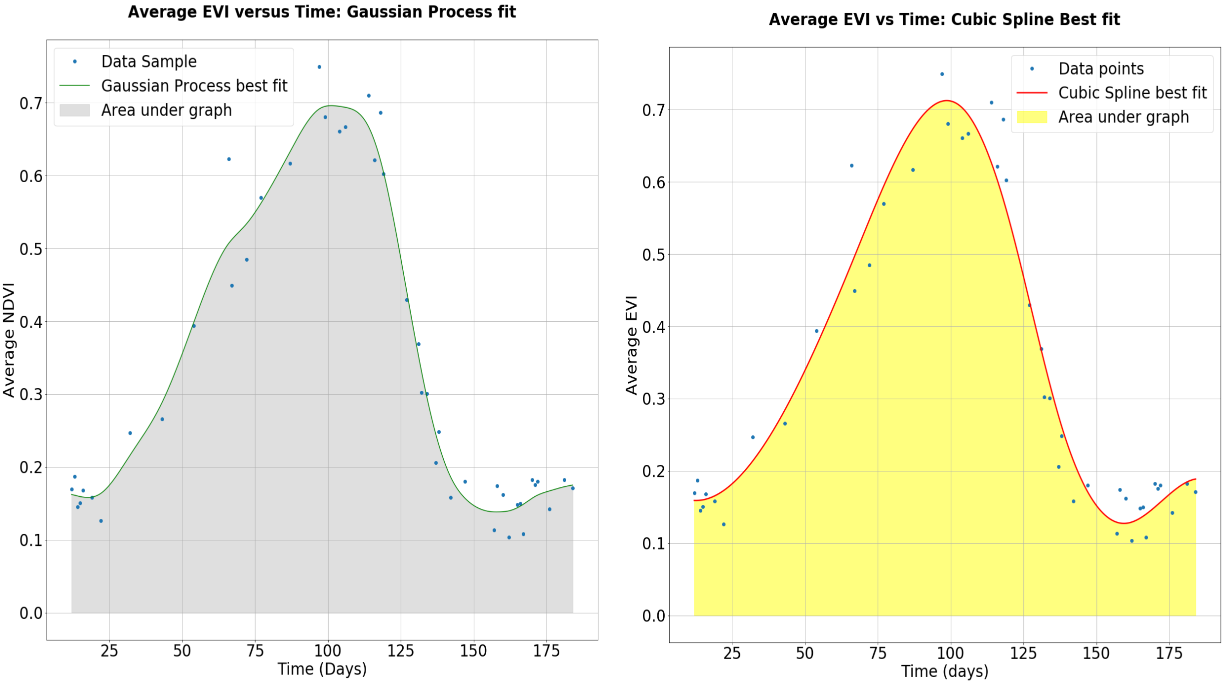
Gaussian Process Modelling is a supervised machine learning algorithm that uses a prior over function which can be used for Bayesian regression. It uses a measure of the similarity between data points (the kernel function) to predict the values of missing data points from the training data. Gaussian processes are beneficial in inferring a distribution over function. The noise in the data is assumed to be having the same finite variance and to be normally distributed [74]. The Matern kernel was implemented for this study using Scikit Learn in Python with different values of *nu* and *scale–length*. The parameter *nu* controls the smoothness of the learned function. The smaller *nu* is, the less smooth the approximated function will be. Values between 0.6 and 0.9 were used for *nu* to mimic an absolute exponential Kernel which is a type of universal kernel that can be integrated against most functions, the *scale–length* was kept at a constant 1.0 which is its default value in Scikit Learn. For each value of *nu* a Least Squares process was carried out for the yield values against the maximum average NDVIs, and the resulting coefficient of determination was observed. Since different parameters can be set in determining the best fit line in both Gaussian process modelling and cubic spline interpolation to best fit the data, it became imperative to come up with a standard for each case which would not change for different fields that were analysed so that all field were analysed using the same standard while keeping in mind that a one fits all type of parameter may actually not exist for either of the models.

Most of the time-series profiles produced in this study showed a perculia phenomenon at the end of the farming season. The NDVI or EVI dipped to a low level as was expected but suddenly steadily rose (see figures 2.13 and 3.4). It was not determined in this study what caused this sudden rise. This dip and sudden rise does not seem plausible from a biological standpoint. The crops in the field at this time of the farming cycle must be dried up and therefore there shouldn't be a sudden rise in NDVI and EVI at this point. A probable cause could be the inability of the cubic spline and Gaussian Process modelling algorithms to smooth out properly at the end of the farming season. Hence, causing a hump in the EVI and NDVI time-series profile.



(a) **(b)**

Figure 2.12: **(a)** Time series profile of the average NDVI for the crops in a field plotted using Gaussian Process Modelling. **(b)** Time series profile of the average NDVI for the crops in a field plotted using cubic spline interpolation.



(a) **(b)**

Figure 2.13: **(a)** Time-series profiles of the average EVI for the crops in a field plotted using Gaussian Process Modelling. **(b)** Time-series profile of the average EVI for the crops in a field plotted using cubic spline interpolation.

2.6 Automation of the image analysis pipeline

Monitoring chlorophyll levels of the canopies of soybean fields using the two vegetation indexes as described in the processes above proved to be a Big Data challenge since each field needed to be separated from very large rasters, was monitored on many days across the farming seasons, and there were many fields that needed to be analysed. This necessitated the Automation of the pipeline to improve the speed of processing the images while providing a standard method for preprocessing and analysing the data obtained from these fields to avoid bias. The Python code that was developed automatically read the files that contained the images, cropped the field of interest from them, masked the clouds and shadows and computed its average vegetation index which it wrote to a file that was named according to the farm and field's names and year of observation. Sometimes a field was not imaged enough to get a solid plot. A criterion was set so that each field studied in this project needed no less than 40 data points (days observed) to be considered for plotting. Fields that did not meet this criterion were only analysed after more of their satellite images were downloaded from the various image distributors, failure to find more images of the particular field resulted in its disqualification.

2.6.1 Flow chart of the automated image analysis pipeline

The automated image analysis pipeline followed the flow chart shown in figure 2.14. It started by searching for raster files in the directory where all the rasters of the virtual constellation were supposed to be stored. It used the 'os' module's 'walk' function to do this. Once it found a raster file, it read it in readiness for analysis. Each raster and its metadata were read using rasterio; the essential factors in the metadata were the raster's transform and CRS. The transform allowed the program to project a referenced raster image from one CRS to another and write out a new image in any of the supported raster formats. After this, the program went on to look for a shapefile which was prepared beforehand using QGIS and a PlanetScope image that contained the field of interest in it, because of this the shapefile possessed an epsg:32735 CRS. If such a shapefile was not found, the program prompted the user to make the shapefile first and retry to run the program, and the program stopped. If the shapefile was found its CRS was changed to match that of the raster that was being analysed in order to make it possible to clip the field. The program then checked if the shapefile overlapped the raster if so, the field for which the shapefile was made was clipped from the rest of the raster, and it underwent cloud masking as described in section 2.4 after which its average vegetation index was computed. The average vegetation index and date of observation were saved, and the program began to search for the next raster in the directory. Since the files were not read in chronological order the dates and average vegetation indexes of the fields were written randomly, and so a sorting library was used to arrange the dates and their respective vegetation indexes before they were plotted using Gaussian and Cubic Spline processes. If no more rasters were found the program attempted to plot the vegetation index time series graph and computed its properties such as the area under the curve, the gradient at specific points of interest on the curve and the maximum average vegetation index.

All the codes used for this image analysis pipeline for data analysis can be found on the GitHub account; MasterPhysicist; `https://github.com/MasterPhysicist/Satellite-Image-Analysis`

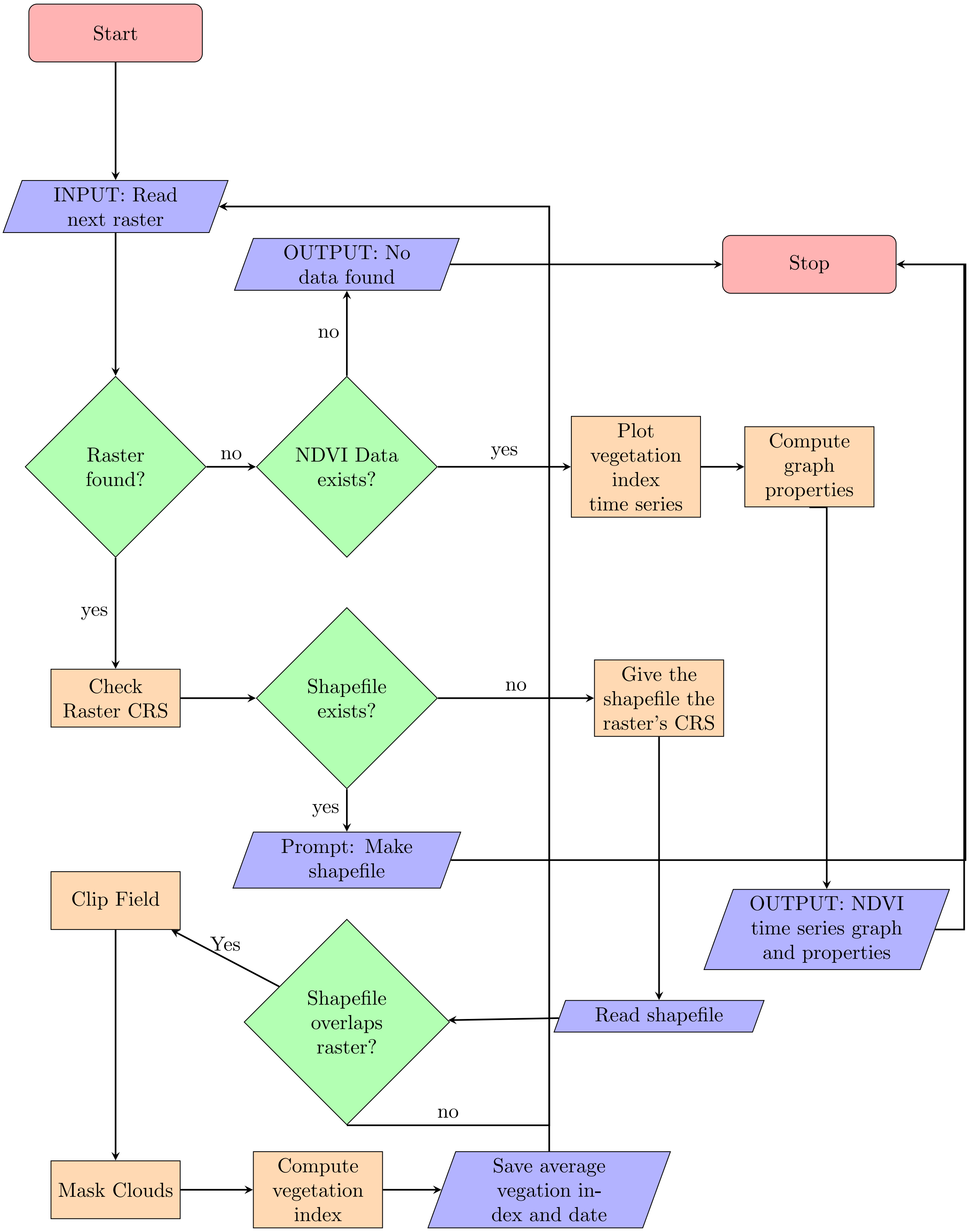


Figure 2.14: Flow chart of automated image analysis pipeline as developed in python. The cloud masking was achieved using a Random Forest classifier while the shapefile was made using QGIS. The NDVI time series were plotted using Cubic Spline and Gaussian Process modelling.

Chapter 3

Analysing EVI and NDVI time-series profiles of soybean canopies

3.1 Introduction

The sprouting, anthesis, senescence, and resulting returns of a soybean variety in terms of seed quantity, bulk, and oil and protein content are a result of its genetic potential and how this interacts with the environment in which the crop grows. Soybean farmers manipulate this environment using techniques such as timely weeding, insect control, and many other proven managerial practices to provide the best possible growth conditions so that they can maximize yield from the crops. A common strategy employed by farmers and breeders in increasing production in soybean is increasing the conversion rate of the incident photosynthetic photon flux density into the crop's biomass [75]. One meaningful way this is achieved is through ample spacing when planting rows so that the plants do not obstruct each other from sunlight while they are growing. Narrow spacing between rows when planting soybean boosts crop productivity and eases the harvesting process [76]. Farmers also consider disease pressure and weed control methods when planting in narrow row spacing. They need crops to be far apart enough to prevent disease and pest transmission while being close enough to protect the soil from too much evaporation. All the fields in this study practiced a narrow spacing of 70cm apart between rows. This fact helped in ensuring that all the fields could be considered together without worrying about the difference in spacing between soybean rows in different fields under our study.

In this chapter, we monitor the average EVI and NDVI of the canopies of Dina, SC-Safari, and SC-Spike across their growing seasons and extract the features of their EVI and NDVI time-series profiles to study their relationship to yield. We will use these profiles to determine which mathematical models distinguish the three varieties from each other. We will attempt to classify them as either high or low yielding relative to each other using these mathematical models. We will also attempt to determine the minimum chlorophyll levels that the canopy of each variety needs in order to remain productive. We will then go on and asses which of the two vegetation indexes is most suitable for analyzing this relationship and which mathematical models best explain the correlation between yield and the properties of the chlorophyll time profiles from EVI and NDVI analysis. Another important thing that we will investigate from these profiles is the time it takes for their particular value at germination to recur when the crop senesces. In essence, this will enable us to analyze the maturity period of each variety. After creating an image analysis pipeline and optimizing it, as explained in chapter 2, the next step involved collecting yield data and geographical positions of farms in our study area in readiness for analysis.

3.2 Visiting Soybean fields and collecting their yield data

During the placement with the IITA's Southern Africa Research and Administration Hub (SARAH) in Lusaka Zambia, farmers were visited in order to collect historic yield data from them for the 2015/2016, 2016/2017, 2017/2018 and 2018/2019 soybean farming seasons. Five commercial farms and four small scale farms comprising 66 soybean fields all together in Lusaka and Chongwe districts were visited. After the fields were observed they were assessed so as to determine whether they qualified to be analysed using the image analysis pipline developed in chapter 2. The following were the criteria used for a field to meet the standard for analysis using the image analysis pipeline used in this study.

- The field was supposed to be in Lusaka or Chongwe districts in Zambia near the IITA headquarters for easy reach when collecting data, and the Kenneth Kaunda International Airport for accurate historic weather data.
- The farmer agreed to provide the data of their yield and they had a good historic documentation of the yield for the 2015/2016, 2016/2017, 2017/2018 and 2018/2019 farming seasons.
- The variety of the soybean crops in the field was known. Mixed cropped fields, i.e fields with two or more soybean varieties grown together in the same field were not analysed unless the varieties could clearly be separated in the satellite images i.e they occupied different parts of the field and the yield reported by the farmer clearly indicated how much yield each of those parts where the respective varieties were grown produced. However, this was very rare in our data.
- The field was required to have an area of no less than 20 x 20 m surface area the minimum possible area that a landsat-8 image can resolve spatially.
- No less than 30 satellite image observations of the field from the combined virtual constellation of Landsat-8 PlanetScope and Sentinel-2 with a good coverage of the peaks of and bases of the average EVI/NDVI curve.
- Farms that experienced extreme crop failure were disqualified.

Of all the farms visited, only three commercial farms; Gaulunia, Syngenta and Chartonel, were chosen to be studied as they were the only ones that met all these criteria, all the small holder farmers visited did not meet these qualifications primarily the historic records and extreme crop failure criteria. It shows one of the greatest challenges one would face in trying to make the process used in this study more widespread in Africa. Therefore, only 39 fields qualified to be analysed out of the 66 fields visited in 2019. This also resulted in 3 soybean varieties taking the center stage as each of the three farms had a predominant variety they planted throughout the years. The three varieties that the three farms grew the most were Dina for Syngenta farm, SC-Spike for Gaulunia farm, and SC-Safari for Chartonel farm. The yield data that was collected was recorded in metric tons per hectare of the grains after threshing.

The failure of small holder farmers to meet these criteria speaks to the difficulty that would be faced in trying to make remote analysis of farms more widespread in Zambia. Harvest data for the fields in Lusaka and Chongwe were collected at the end of May 2019 so as to include yield for the 2018/2019 farming season. Yield data was collected by phoning, emailing or meeting in person the farmers who agreed to provide their data. Table 3.1 below shows the farms and respective 39 fields that were used in this study, the color scheme gives a visual representation of the distribution of the three varieties and how they were spread among the three farms during the 4 farming seasons over which this study was conducted.

Table 3.1: Fields visited and the soybean varieties under their cultivation per year.

Farm	**Field**	**Area (m^2)**	**Soybean variety per farming season**			
			2015/2016	2016/2017	2017/2018	2018/2019
Gaulunia	CVD	44	Safari	Spike	Spike	Spike
	CVD ext	11	Safari			
	CVD(H)	13	Safari	Spike	Spike	Spike
	CVE(P1)	28	Safari	Spike	Spike	Spike
	CVE(P2)	28	Safari	Lukanga	Spike	Spike
	CVW	79	Spike	Shungu &Lukanga	Spike	Spike
	CVW ext	14	Spike			
	RE	50	Spike	Spike	Spike	Spike
	NGW P1	36	Safari	Spike	Spike	
	NGW P2	36		Spike	Safari	Spike
	Ndk 1	49	Spike	Safari	Spike & Safari	Spike
	Ndk 2	48			Spike	Spike
	D/Dale F1	23	Lukanga	Spike		Spike
	D/Dale P1	14		Lukanga & Safari		Spike
	D/Dale P2	14			Spike	
Syngenta	P1	22			Dina	Dina
	P2	20		Dina	Dina	Dina
	P3	50		Dina	Dina	
	P4	50		Dina	Dina	Dina
	P5	12		Dina	Dina	Dina
	P6	38		Dina	Dina	Dina
	P6 B	25		Dina	Dina	Dina
	P9	16		Dina	Dina	Dina
	P10	38		Dina	Dina	Dina
	P11					
	P12	40		Dina	Dina	Dina
	P13	35		Dina	Dina	Dina
	P14	65		Dina	Dina	Dina
	P15	65		Dina	Dina	Dina
	P16	30		Dina	Dina	Dina
	P 16 B	30		Dina	Mila	Mila
	P17	20		Dina	Dina	Dina
Chartonel	P1	40				Safari
	P2	40				Safari
	P3	40				Safari
	P5	40				Safari
	P11	40				Safari
	P12	40				Safari
	P14	40				Safari

The following is the color scheme for the table; yellow represents SC-Spike, blue represents Dina, Green represents SC-Safari, while red represents missing data or fields that were not analysable due to mixed cropping or too few satellite image data points when creating the average EVI and NDVI time series profile of the field. This was especially so in 2015/2016 when there were much fewer satellites in the PlanetScope constellation and only Sentinel-2A in the Sentinel constellation. There were other reasons why some fields were assigned a red color. In some cases, the yield data of the fields were not well recorded or the crops in the field were, in fact, a mixture of two varieties that had their yield reported as one field as was the case with Gaulunia CVW and Gaulunia Diamondale P1 in the 2016/2017 farming season. This was different from Gaulunia Ndk1 in 2017/2018 farming season which was split into two halves with Spike and Safari grown in each separate half which was clearly distinguishable from each other in satellite images and each variety's yield was recorded separately. Empty cells represent missing yield data or a year when that field was not farmed with soybean. Some satellite images in some years were missing auxiliary data and so could not be processed to surface reflectance thereby reducing the number of images available for the analysis of certain fields. such fields were also excluded from the analysis.

Studies have shown that planting soybean in narrow rows (<80cm) increases yields. This is because narrow rows allow the crop canopy to intercept more light during the growing season, thus boosting growth rates and reducing soil moisture loss. It also eases the process of harvesting [77]. However, farmers need to take into account disease pressure and weed control options when planting in narrow rows. Therefore, most farmers aim for row spacing that is short enough to allow high yield and easy harvesting but long enough to allow for weeding and prevent disease transmission [76]. The three farms that provided their yield data in this study all had a narrow row spacing between 30 and 40cm. This was very convenient for our satellite observation as it meant we did not have to factor in row spacing when doing our analysis.

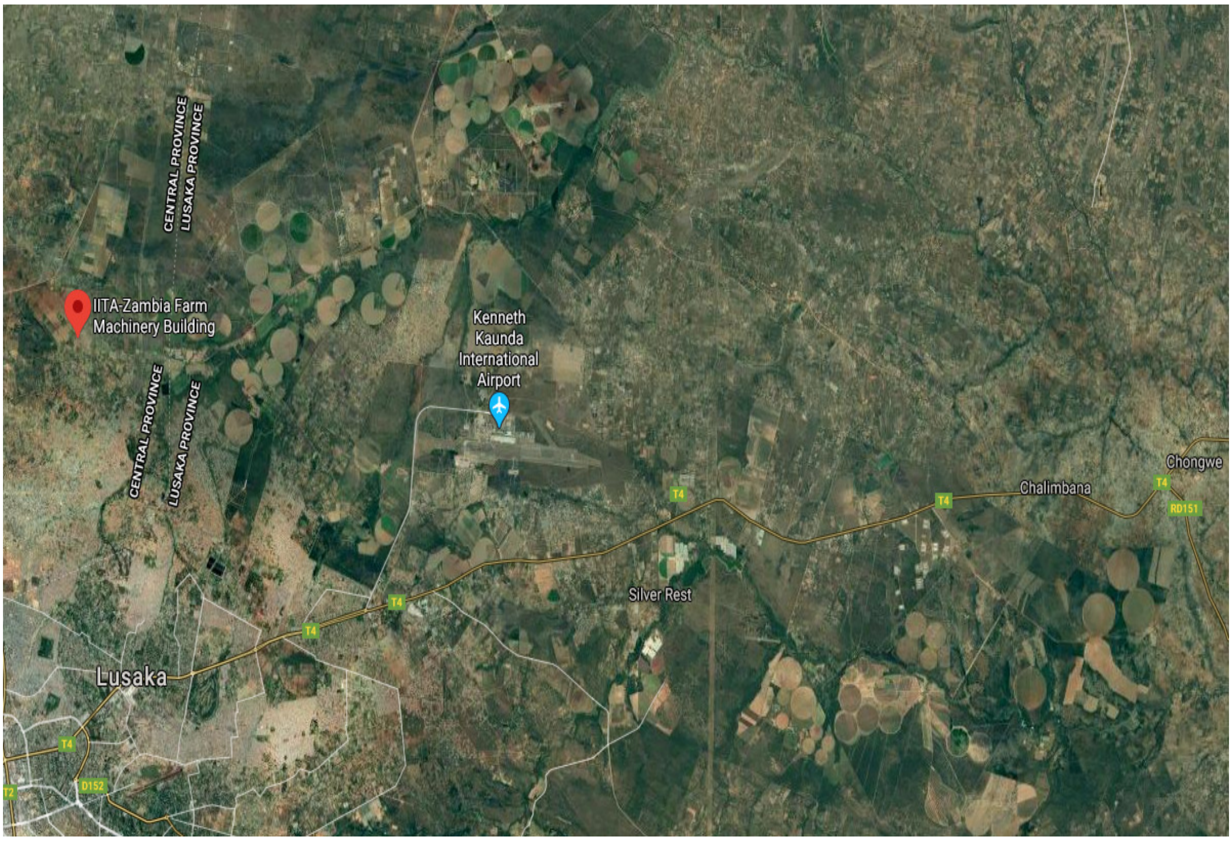


Figure 3.1: Area of Interest, encompassing IITA Southern Africa Research Hub (red pin), Kenneth Kaunda International airport(blue pin), Zamseed farms; immediately north of IITA, Gaulunia farm; immediately east of IITA, Chartonel farm; far northeast of IITA and Syngenta farm in the far southeast of the image. All the 66 fields that where visited during the placement with IITA lie in this area. Image courtesy of google maps.

Key agronomic attributes of the three soybean varieties under study

The three soybean varieties in our study were developed specifically for the Zambian environment by the breeding institutions that developed them. Farmers choose a particular soybean variety for its agronomic attributes that fit in well with their managerial skills and the environment of their fields. In this section, we discuss each variety under study and its various agronomic attributes as advertised by its developers.

1. **Dina**:
   This is a determinate non-promiscuous variety that has good tolerance to the acidic soils found in most parts of Lusaka and Chongwe districts. It was released for the Zambian environment in 2003 by Syngenta, working in conjunction with the Maize Research Institute. Dina is reported to have a maturity period of 102 days by its developer Syngenta on their website. It is also reported to have great tolerance to shattering once the pods are ready for harvest, it is drought-resistant and has a maximum potential yield estimated at 4.5 tons/ha [78].

2. **SC-Spike:**
   This is a determinate variety that was developed by SeedCo Zambia Limited for a locally adapted high yielding variety, it is a highly prolific variety and is reported to having a maximum yield potential that exceeds 5 tons/ha. It is widely grown by commercial farmers in the region under study (Chongwe and Lusaka districts). It has great resilience to bacterial blight, downy mildew and red leaf blotch [79]. Farmers spoken to during the placement with IITA said that it is not very drought resistant. Its maturity period was not found on any of SeedCo's websites nor in any literature reviewed during the course of this research.

3. **SC-Safari:**
   This is an indeterminate, non-promiscuous, quick maturing cultivar, with a maturity period of 125 days [79]. It was released by SeedCo Zambia Limited in 2004 for Zambian and Zimbabwean farmers. Safari has a yield potential of up to 4.5 tons/ha under good management. It has low resistance for Red Leaf Blotch and Frogeye Leaf Spot and has good pod clearance and high resistance to pod shattering [79].

3.2.1 Weather Patterns for Chongwe and Lusaka districts (2017-2019)

Most farmers in Lusaka and Chongwe districts plant their soybean in November, at the start of the rainy season. In the case of small scale farmers, the rainfall is usually their only hope for watering their plants. Commercial farmers on the other hand, use the rains to supplement their irrigation schemes. Therefore, the total rain that falls during the farming season directly indicates how well soybean does in the two districts and indeed across the country. Figure 3.2 shows the daily weather patterns of the farming seasons in the area under study. As shown in Table 3.1, there was no need to collect weather data for the 2015/2016 farming season as none of the fields were analysed for that season for reasons discussed in section 3.2. The total precipitation that fell in the two districts during the 2018/2019 soybean farming period was 522.0 mm. This is far below 1078.1 mm; the average total rainfall expected in the region. Figure 3.2 (a) shows that it was erratic rainfall and thus resulted in severe dry spells in the area and massive crop failure for most of the small scale soybean farmers. All the small scale farmers who were visited during the placement with IITA experienced enormous crop failure. Therefore, yield data from these small scale farmers was not collected for 2018/2019; most of them also did not have well documented historical yield data. For these reasons, only data from the three commercial farmers mentioned earlier was collected. The 2017/2018 farming period saw a much higher total precipitation than 2018/2019, reaching 926.4 mm. Although it was much better rain, its erratic nature, as evidenced from Figure 3.2 (c), must have adversely affected the soybean farmers, for example, there was no rainfall from day 45-60 of the farming season. That is 15 days of no rain at a time when the crops needed to be growing vegetatively. The 2017/2018 farming season also saw some rainfall in May at the time when soybean crops would have been drying up in the field in readiness for harvest; such rain could have added to the adverse impacts the yield of some farms experienced that year. Total precipitation was even higher in the 2016/2017 farming period, with 1030.4 mm of rainfall, which picked during the period from day 50-125. This was a good rainy season for the farmers provided they planted their crops at the end of November. The trend showed an overall reduction in precipitation with time over the three years observed in this study. The average soil temperature rose from 24.41^oC to 24.74^oC to 26.27^oC while the average air temperature rose from 21.20 to 21.57 to 22.53 for the 2016/2017, 2017/2018 and 2018/2019 farming seasons respectively. There was also a considerable increase in average daily solar irradiation across the three farming seasons from 18.41 W/m^2, to 19.20 W/m^2, to 20.33 W/m^2 respectively.

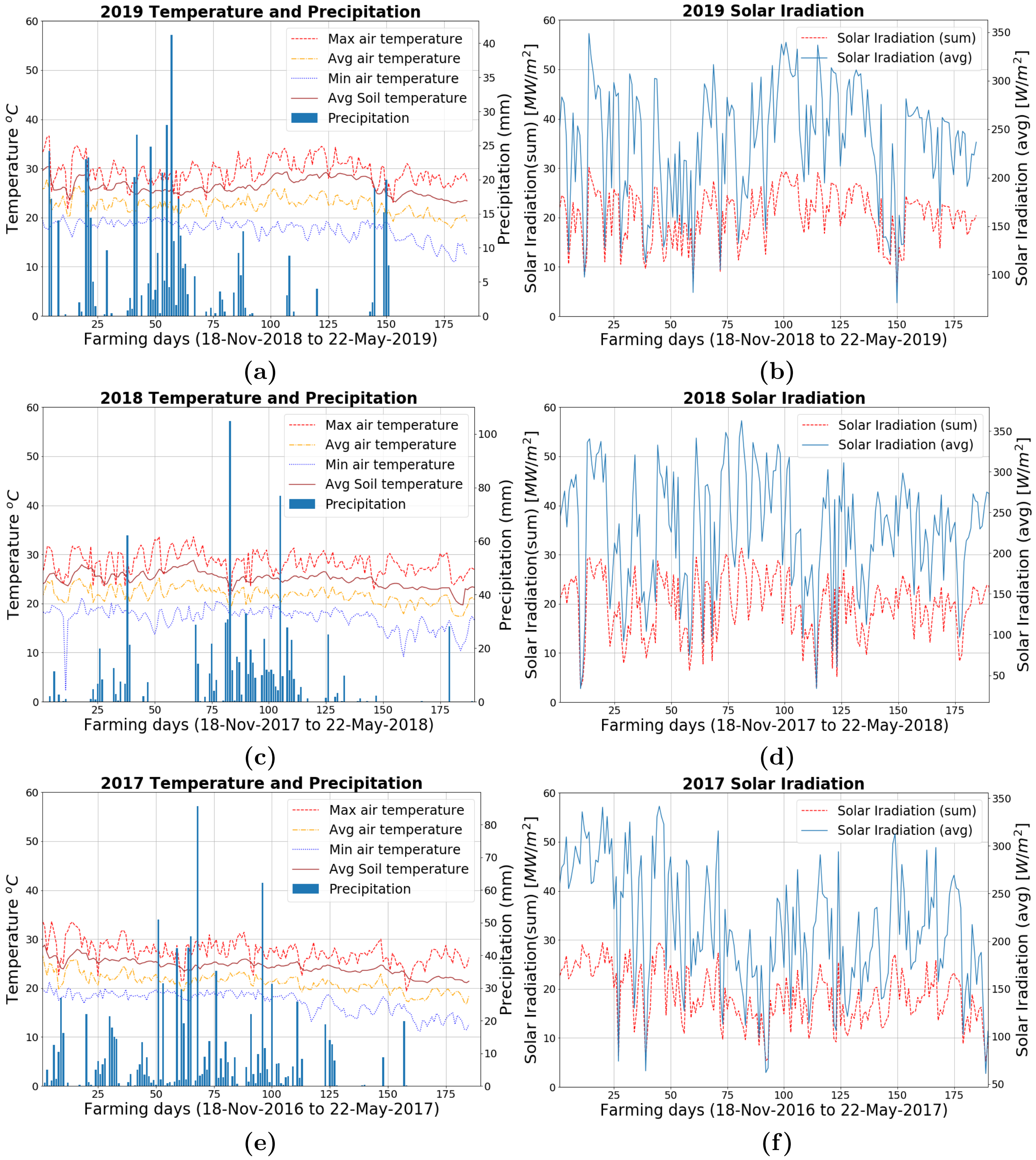


Figure 3.2: Daily minimum, maximum and average air temperature, soil temperature, total daily precipitation, average solar irradiation and total solar irradiation during the 2016/2017, 2017/2018 and 2018/2019 farming seasons of Lusaka and Chongwe districts. The dates of observation are from 18th November to 22nd May for each year, coinciding with the soybean farming period of the two districts

3.3 Accuracy of the cloud and shadow masking model

Before the image analysis pipeline could become viable for use and automation it was vital to make sure that the cloud masking process implemented using Random Forest Classification was accurate enough in classifying cloud and shadow-filled pixels from good ones that contained data without it being too stringent on haze filled pixels. Images with small amounts of haze were considered as still eligible for analysis in this study so as to increase the number of observations. A criterion was set to have no less than 40 canopy EVI and NDVI observations across a farming season for a graph to be considered good enough to acquire information from it.

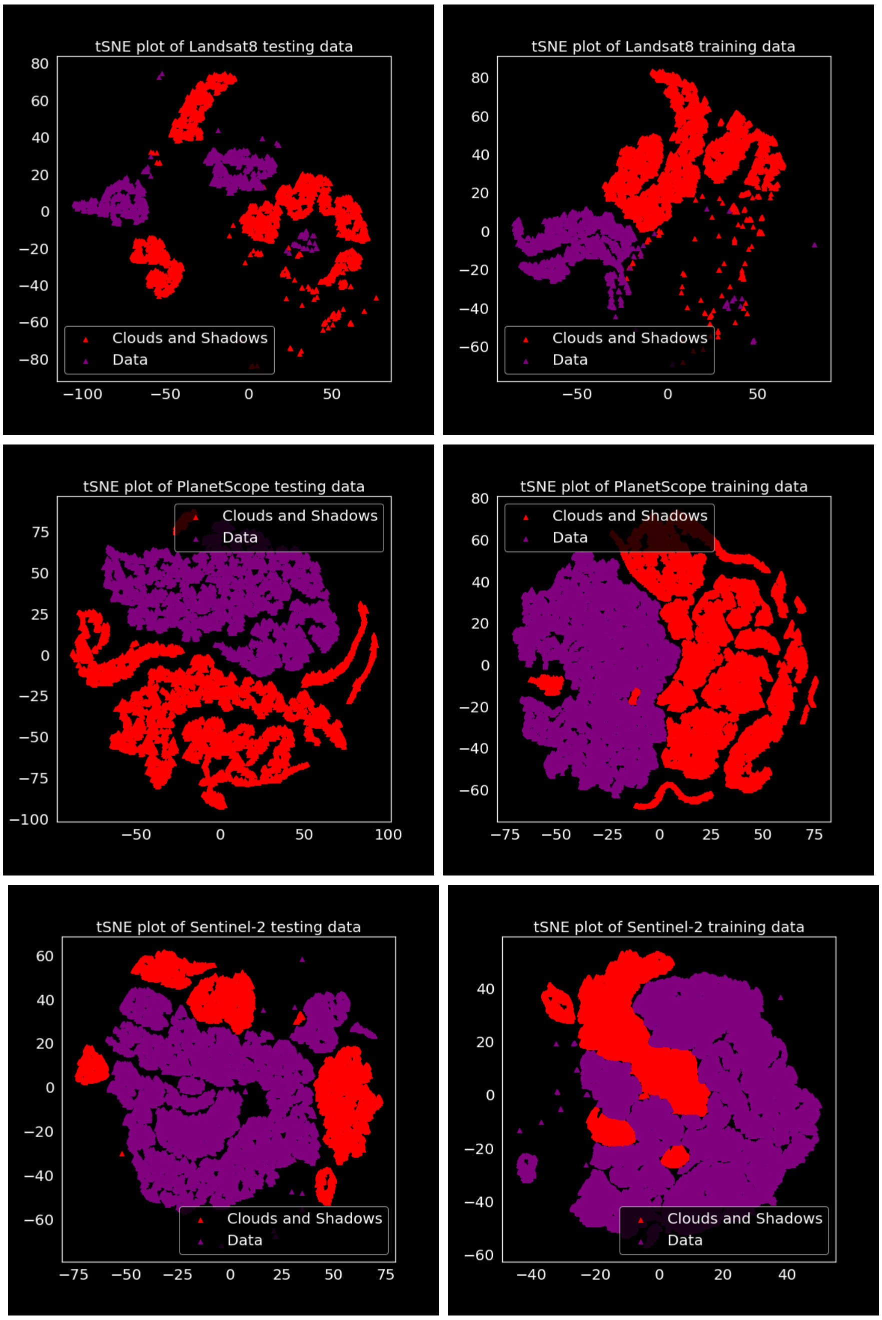


Figure 3.3: t-distribution Stochastic Neighbor Embedding (tSNE) plots visualizing the data clustering of the training and testing data sets used to create the cloud and shadow classifiers used to mask clouds and shadows. The plots show that the unwanted pixels were distinct from the wanted pixels with very minimal overlaps between the two clusters.

A nonlinear dimension reduction unsupervised machine learning algorithm called T-distributed Stochastic Neighbor Embedding (tSNE) was used to visualise how the training and testing datasets were clustered relative to each other, as shown in Figure 3.3. The tSNE is particularly well suited for the visualisation of high-dimensional datasets like the ones used in the training and testing of our cloud and shadow classifier to see how closely related the "cloud and shadow" pixel values, represented in red, were to the "data" pixel values, represented in purple. The training and testing datasets used in this study had seven columns and millions of rows; it reduced them into two dimensions to visualise how they were clustered [80].

The main reason for developing separate cloud and shadow masking models for each satellite constellation was because the masking processes provided by the respective distributors of the satellite images were very stringent when masking haze, clouds and shadows. They did away with any haze which therefore resulted in very few observations during certain parts of the farming seasons. Lusaka and Chongwe districts are very cloudy and hazy during the rainy seasons, particularly at the end of December and beginning of January during the peak of the rainy season. The downside to allowing more haze-filled images into the EVI and NDVI time profile analysis was that it brought some noise into the graphs which in turn brought about some noise in the features of the profiles that were analysed for correlation with yield.

To ensure that the cloud masking was accurate, the model that was developed after training the algorithm was cross-validated by testing it with data that was not used in its training and its accuracy was scored using sklearn's *accuracy_score* method. A separate classification model was developed for each satellite constellation to avoid any confusions that could result from the small differences between the satellites. Only classification models that showed high accuracy scores were saved for use with more images. The classifier developed for Landsat 8 showed an accuracy score of 99.72%, while that developed for Sentinel-2 had an accuracy score of 99.13% and the one developed for PlanetScope reached an accuracy score of 99.70% after cross-validation with data from the testing set that was collected from the same images as the training set. By default the sklearn function *train_test_split* carved off 25% of the data cropped from the images for testing, the downside to this is that even if *train_test_split* chose the data randomly, the data chosen for testing the classifier could still have been very similar to the data that was chosen for training hence giving a high cross-validation score and yet the model was not as accurate with data from a different image. The fact that each classifier was built separately and showed slightly different accuracy scores means that there may have been some discrepancies in the vegetation index time series plots which may have spilt over into the yield analysis plots that followed.

3.4 Which features of the EVI/NDVI time-series profile correlate with yield?

The relationship between soybean yield and various features of the NDVI and EVI time-series profiles was investigated by looking at how yield linearly correlates with these features. After the average EVI and NDVI of the canopies of respective soybean fields were computed using the image analysis pipeline presented in chapter 2, the distribution of data points were modelled using cubic spline interpolation and Gaussian process modelling. These gave the best approximation of the trend for the two vegetation indexes separately across the farming seasons; five different features were extracted from the resulting time series profiles for this investigation. Figure 3.4 shows the five features of interest that were analysed for linear correlation with yield. The area under the graph shown in yellow was computed using the Simpson algorithm, implemented in Python using the simps method of the scipy.intergrate module, it integrated the function of the curve over the period of observation (18^{th} November to 22^{nd} May) to get the area under the curve. The dimensions for the area under the curve are EVIdays and NDVIdays. The average maximum vegetation index was obtained by getting the maximum value in the Gaussian process modelled, and cubic spline interpolated array of the canopy's average EVI and NDVI versus time profile using the numpy module's max method. The full width at half the maximum (FWHM) of the curve was obtained by dividing the maximum average vegetation index by 2 to get its half value, after which the time interval between the two points of the curve where this value occurs was determined. It is this time interval that is the FWHM. The average vegetation indexes at the dates of the FWHM were also noted. The gradient of the graph at FWHM dates for both the ascending side and descending sides of the curves were computed using numpy and analysed for linear correlation with yield. All the varieties were analysed together for the linear correlation of these features of their EVI and NDVI curves with yield irrespective of variety.

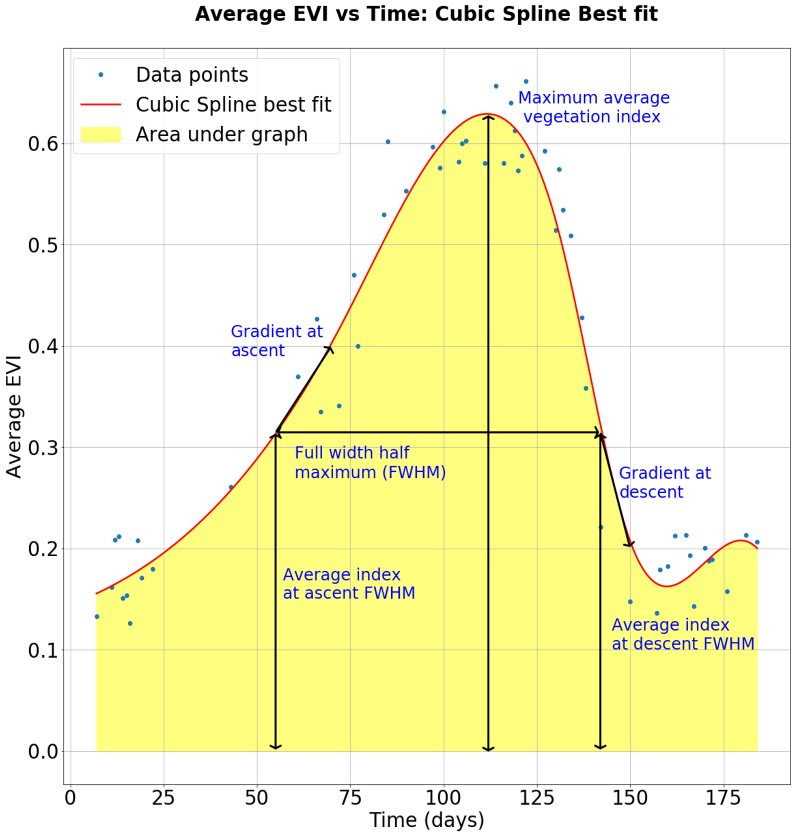


Figure 3.4: Features of the average vegetation index time series profile that were analysed for correlation with soybean yield.

The features that showed a high linear correlation with yield were analysed further by plotting their regression against yield using linear (y = mx + c) , logarithm (y = a·ln(bx) + c), and power (y = ax^b) functions. Where, y is the yield of the field after the crops are threshed, measured in tons/ha and x is the value of the feature of interest of the average NDVI or EVI time series profile in either the cubic spline or the Gaussian process model of the two vegetation indexes. Their regression plots were made to infer more information about their relationship, such as how the regression models fit well into the data distribution on the plots.

The values of area under the NDVI time series curves ranged between 80 and 90 NDVIdays for all the three varieties,. The area under the EVI time profile curve ranged between 60 and 70 EVIdays across all the three varieties. The area under the curve did not change much between

Gaussian process modelling and cubic spline interpolation. The full width at half the maximum of the EVI/NDVI ranged between 70 and 115 days. It showed no linear correlation with yield even though it was expected to, since when looking at all the varieties together, each variety has a different growth period. Such a correlation was also expected in the area under the curve based on the same premise i.e., this was expected to translate into different areas under the curve with respect to variety. However no significant difference was observed that distinguished the varieties from each other in a linear correlation with yield.

The resulting coefficients of determination of the linear correlation between yield and the various features of these curves are recorded in table 3.2 for the cubic spline features and table 3.3 for the Gaussian process features. It can be observed from the two tables that only the maximum average EVI and NDVI showed a significant linear correlation with the yield.

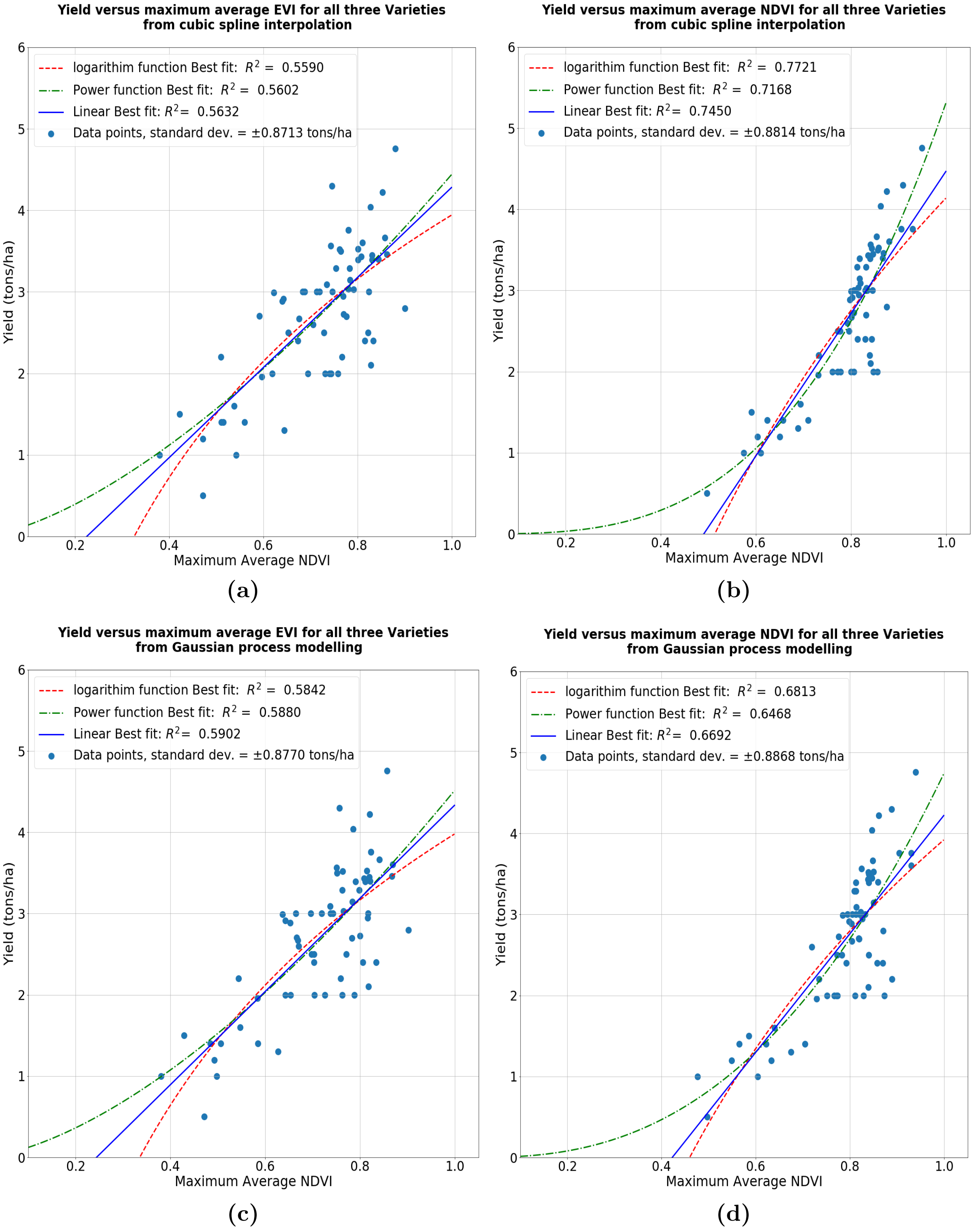
Table 3.2: Coefficients of determination (R^2) values for linear regression of the yield as a function of the area under the graph, the full width half maximum, the gradients of ascent and descent there, and the maximum of the cubic spline interpolated EVI/NDVI time series graphs.

Variety	**Area under graph**		**Ascent gradient**		**Descent gradient**		**FWHM**		**Maximum average**	
	NDVI	**EVI**	**NDVI**	**EVI**	**NDVI**	**EVI**	**NDVI**	**EVI**	**NDVI**	**EVI**
All three	0.0387	0.0126	0.0328	0.0185	0.1827	0.0531	0.2451	0.1903	0.6692	0.5902
Dina	0.0268	0.0152	0.0919	0.0307	0.0413	0.2629	0.2383	0.3409	0.7246	0.6646
Safari	0.0131	0.0420	0.0942	0.0317	0.0522	0.0270	0.3021	0.1484	0.6728	0.8050
Spike	0.0023	0.0398	0.0420	0.1086	0.0750	0.0302	0.1301	0.1773	0.7297	0.2872

Table 3.3: Coefficients of determination (R^2) values for linear regression of the yield as a function of the area under the graph, the full width half maximum, the gradients of ascent and descent there, and the maximum of the Gaussian process interpolated EVI/NDVI time series graphs.

Variety	**Area under graph**		**Ascent gradient**		**Descent gradient**		**FWHM**		**Maximum average**	
	NDVI	**EVI**	**NDVI**	**EVI**	**NDVI**	**EVI**	**NDVI**	**EVI**	**NDVI**	**EVI**
All three	0.0631	0.0193	0.0294	0.0242	0.0023	0.0076	0.3201	0.3792	0.7450	0.5632
Dina	0.0834	0.0193	0.0307	0.0256	0.1064	0.0452	0.2297	0.2096	0.6285	0.7855
Safari	0.0213	0.0192	0.0601	0.0107	0.0144	0.0039	0.0084	0.0463	0.7249	0.8387
Spike	0.0166	0.0515	0.0144	0.0395	0.0512	0.0301	0.3275	0.1940	0.7882	0.3796

The features of the graphs that showed a linear correlation with yield were plotted against it to visualise the trend of their relationship. Figure 3.5 shows the resulting linear, logarithm and power plots of the yield versus the maximum average EVI and NDVI from cubic spline and Gaussian process interpolation. All three varieties are included in these plots. It can be observed from Figure 3.5 that the maximum average EVI and NDVI have a positive correlation with the yield. Maximum average NDVI showed higher R^2 values with yield in both cubic spline and Gaussian process models than maximum average EVI. The plots also show that the power function model gave the best fit to the shape of the data distribution for the maximum average vegetation indexes obtained from both the cubic spline and Gaussian process models when the three varieties where plotted in one graph. This was especially so in the maximum average NDVI plots. The fact that the maximum average EVI and NDVI show this correlation means that the EVI and NDVI values at the Full With of the Half Maximum (FWHM) also correlate with yield since they are simply half of the maximum average EVI and NDVI values of these curves. The next important thing is to observe how this correlation between yield and the maximums average EVI and NDVI varies when looking at each variety separately.



Yield versus maximum average EVI for all three Varieties from cubic spline interpolation

(a)

Yield versus maximum average NDVI for all three Varieties from cubic spline interpolation

(b)

Yield versus maximum average EVI for all three Varieties from Gaussian process modelling

(c)

Yield versus maximum average NDVI for all three Varieties from Gaussian process modelling

logarithim function Best fit: R^2 = 0.6813
Power function Best fit: R^2 = 0.6468
Linear Best fit: R^2= 0.6692
Data points, standard dev. = ±0.8868 tons/ha
Yield (tons/ha)
Maximum Average NDVI
0
1
2
3
4
5
6
0.2
0.4
0.6
0.8
1.0

(d)

Figure 3.5: Linear, logarithmic and power function best fit curves of the cubic spline interpolated, and Gaussian process modeled, average EVI/NDVI versus time(days) series curves of Dina SC-Safari and SC-Spike for the farming seasons of 2016/2017, 2017/2018 and 2018/2019.

3.5 How does the correlation vary per variety?

Building on the principle that different soybean varieties have different traits in terms of yield and phenotype, it became imperative to observe the correlation between yield and maximum average EVI and NDVI for each variety separately. Each variety was analysed by mathematically modelling the yield versus maximum average EVI and NDVI relationship using linear (y = mx + c), logarithm (y = aln(bx) + c) and power (y = ax^b) models. The resulting models were plotted on the same graph for each analysis of the cubic spline and Gaussian process time series profiles to compare how well they fitted the data distribution in the plots. Coefficients of determination (R^2) values were also computed for each of the mathematical models to measure how well they correlated with the data points. Also noted was how well the trendlines of these functions fitted the data point distribution in these plots and how the extrapolated trendlines faired in the EVI and NDVI range that plants exist in. Figures 3.6, to 3.11 below show the trends of the mathematical models of the yield versus the maximum average EVI and NDVI of each of the varieties when analysed separately. All the plots showed a positive correlation between yield and maximum values of these vegetation index time-series profiles. Analysing single varieties like this resulted in a significant reduction in the standard deviation of both the yield and maximum average EVI and NDVI values compared to those observed in the analysis of all three varieties together. This reduction signifies that these values were much less dispersed when looked at within a single variety. Therefore, the reduction in the standard deviation of these plots shows that separate analysis of the varieties improves the correlation of the mathematical models with the data points. Plots of yield versus both cubic spline and Gaussian process modelled maximum average NDVI values of SC-Spike showed high R^2 values, in sharp contrast with the low values observed in the equivalent EVI analysis. These low R^2 values were caused by data from a few fields that diverged greatly from the trend followed by the rest of the fields. It can be seen from the equations of the plots in figures 3.6, to 3.11 that the different varieties have different equations from each of the mathematical models used to plot the correlation of yield with maximum average EVI and NDVI. Each equation, therefore, describes the specific relationship between yield and maximum average EVI and NDVI of a particular variety, i.e. each of them is a characteristic equation of a specific variety. The smallest standard deviation of yield values was observed in the SC-Spike regression, while the largest was observed in the Dina regression.

3.5.1 Logarithm regression

Logarithm functions fitted the data point distribution of yield versus maximum average NDVI and EVI very well. They gave R^2 values ranging from 0.2869 to 0.7943 in their yield versus maximum average EVI plots. In the case of the logarithm plots, yield versus maximum average EVI derived from Gaussian process modelled time-series profiles gave higher R^2 values than those obtained using cubic spline interpolated ones except in the case of SC-Spike where the reverse was observed. The R^2 values of the logarithm best-fit lines in the yield versus maximum average NDVI analysis ranged between 0.6464 and 0.8085 a significant improvement from the range observed in the EVI analysis. Yield versus maximum average NDVI derived from cubic spline interpolation time-series profiles gave higher R^2 values than those obtained from Gaussian process model ones. Forward extrapolation of the logarithm trendline gave the lowest predictions of maximum potential yields than the linear and power models. The logarithm model did not go below 0.2 EVI and NDVI when it was extrapolated backwards.

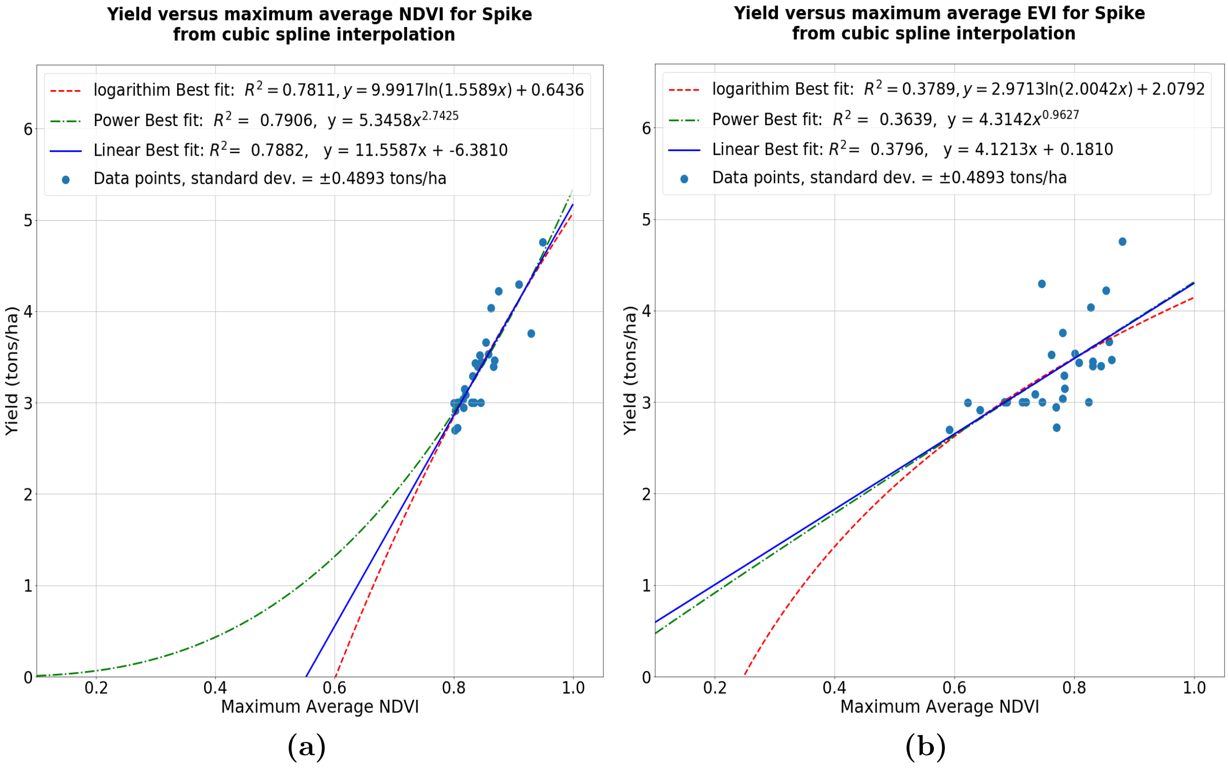
3.5.2 Linear regression

Linear functions showed a good fit to the data point distribution of the yield versus maximum average EVI and NDVI data-point distribution, as shown in figures 3.6 to 3.11 below. Yield versus Maximum average EVI linear best fits gave R 2 values ranging from 0.2872 to 0.6646. Linear plots of yield versus maximum average EVIs derived from Gaussian process modelling showed higher R^2 values than those of maximum average EVIs derived from cubic spline interpolation. The exception was the yield versus maximum average EVI and NDVI plots for SC-Spike where the reverse was observed. Yield versus maximum average NDVI linear best fits saw R^2 values ranging from 0.6728 to 0.8387, a great improvement compared to the EVI version of the analysis. Linear plots of yield versus maximum average NDVI derived from cubic spline interpolation showed higher R^2 values than those for which the maximum average NDVI were derived from Gaussian process modelling. Different varieties showed different gradients of the linear function. Unlike the logarithm function, the linear function when extrapolated backwards went below EVI = 0.20 as shown in figures 3.6 (b), 3.7 (b) and 3.8 (b) below. It stayed above this point in the linear plots of the yield versus maximum average NDVI.

3.5.3 Power regression

The power functions showed an excellent fit to the data point distribution of the yield versus maximum average NDVI and EVI with high R^2 values. Of all the plots in figures 3.6 to 3.11, the highest R^2 values were observed for the power regression's yield versus maximum average NDVI from cubic spline interpolation time-series profiles. Power functions also gave the best fit to the shape of the yield versus maximum average NDVI from cubic spline interpolation data points in figure 3.5 (b) and (d) above when the regressions were plotted for all the varieties together. Yield versus maximum average EVI power best fits gave R^2 values ranging from 0.2821 to 0.8322. Yield versus maximum average NDVI power best fits saw R^2 values ranging between 0.7028 and 0.8540. Just like the linear and logarithm functions, cubic spline interpolation yield versus maximum average EVI power plots gave higher R^2 values than Gaussian process modelled ones except in the case of SC-Spike where the reverse was observed. Power function yield versus maximum average NDVIs obtained from cubic spline interpolation gave higher R^2 values than those obtained from Gaussian process modelled NDVI time series profiles. Forward extrapolation of the power function gave the highest maximum potential yield predictions. Backward extrapolation of power best fit lines went below 0.2 EVI and NDVI.

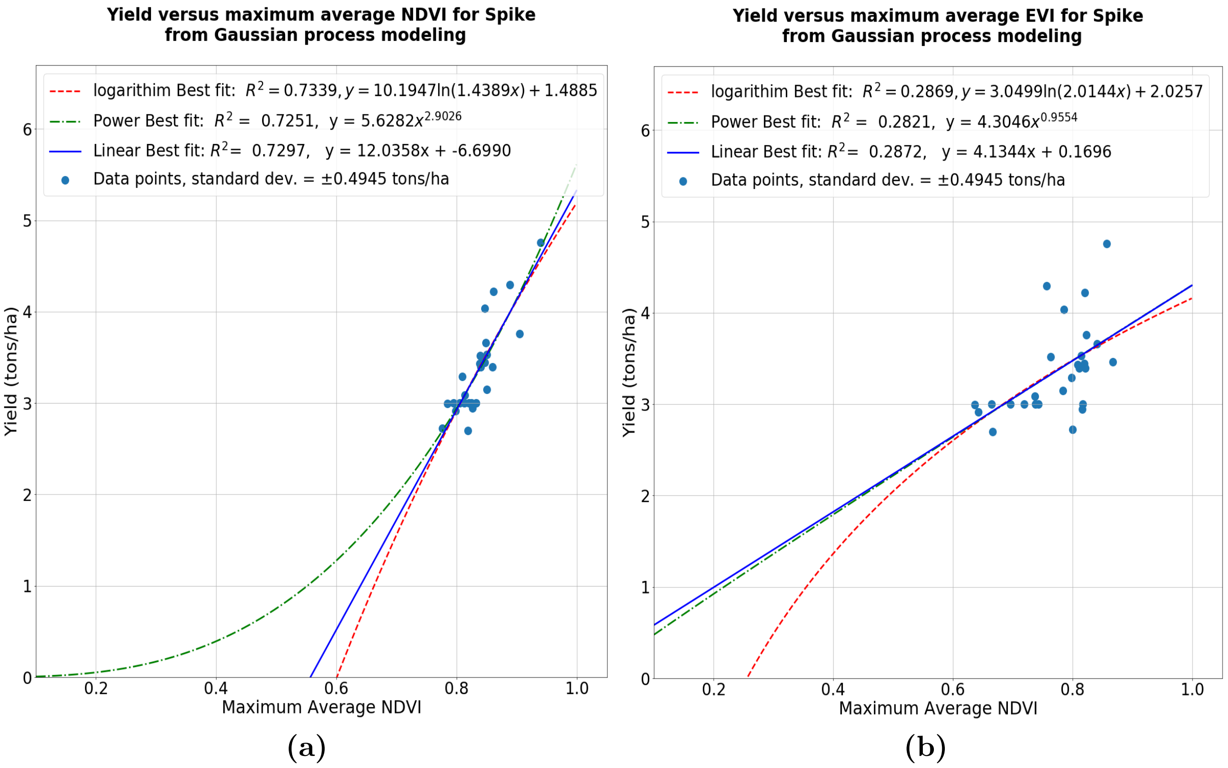
All three mathematical functions showed good R^2 values in the yield versus maximum average EVI and NDVI plots. The yield versus maximum average NDVI obtained from cubic spline interpolation showed higher correlation to yield than the yield versus maximum average NDVI derived from Gaussian process modelling of the time-series profiles. Conversely, in the EVI analysis, yield versus maximum average EVI obtained from Gaussian process modelling showed a higher correlation with yield than yield versus maximum average EVI obtained from cubic spline interpolation. Power functions showed higher R^2 values than logarithm and linear functions. In the next section, we will determine what extrapolation of the trendlines of the logarithm, linear and power functions of yield versus maximum average EVI and NDVI can show about Dina SC-Spike and SC-Safari.


Yield versus maximum average NDVI for Spike from cubic spline interpolation

Yield versus maximum average EVI for Spike from cubic spline interpolation

(a) (b)

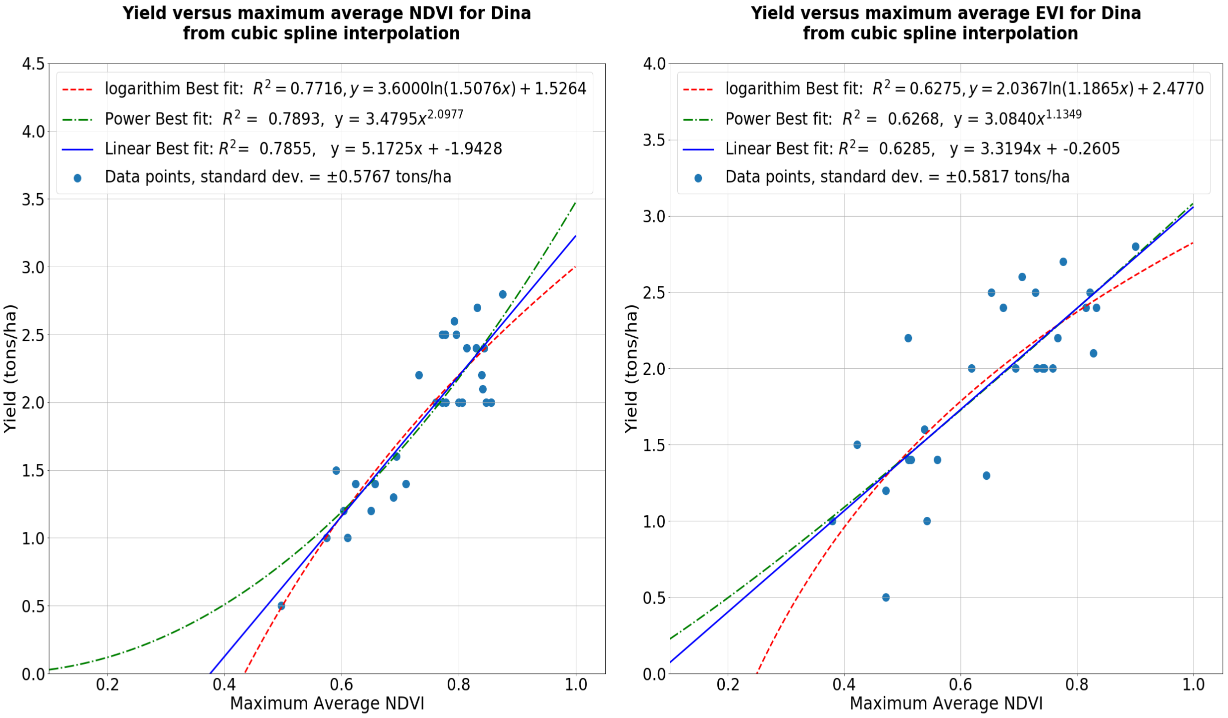
Figure 3.6: Linear, logarithm and power regression plots of the cubic spline interpolated maximum average EVI and NDVI versus time(days) of SC-Spike for the 2016/2017, 2017/2018 and 2018/2019 farming seasons.


Yield versus maximum average NDVI for Spike from Gaussian process modeling

Yield versus maximum average EVI for Spike from Gaussian process modeling

(a) (b)

Figure 3.7: Linear, logarithm and power regression curves of Gaussian process modeled average EVI versus time(days) for SC-Spike for the 2016/2017, 2017/2018 and 2018/2019 farming seasons.

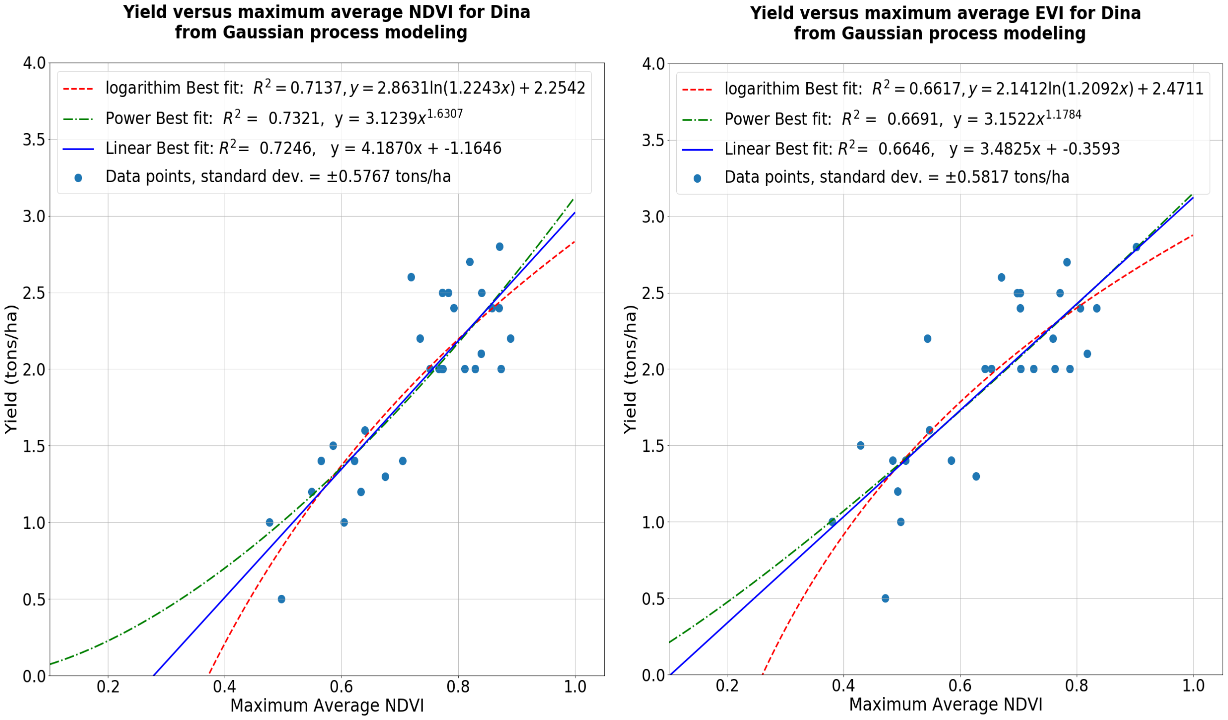

Yield versus maximum average NDVI for Dina from cubic spline interpolation

(a)

Yield versus maximum average EVI for Dina from cubic spline interpolation

(b)

Figure 3.8: Linear, logarithm and power regression curves of the cubic spline interpolated maximum average EVI and NDVI versus time (days) for Dina for the 2016/2017, 2017/2018 and 2018/2019 farming seasons.

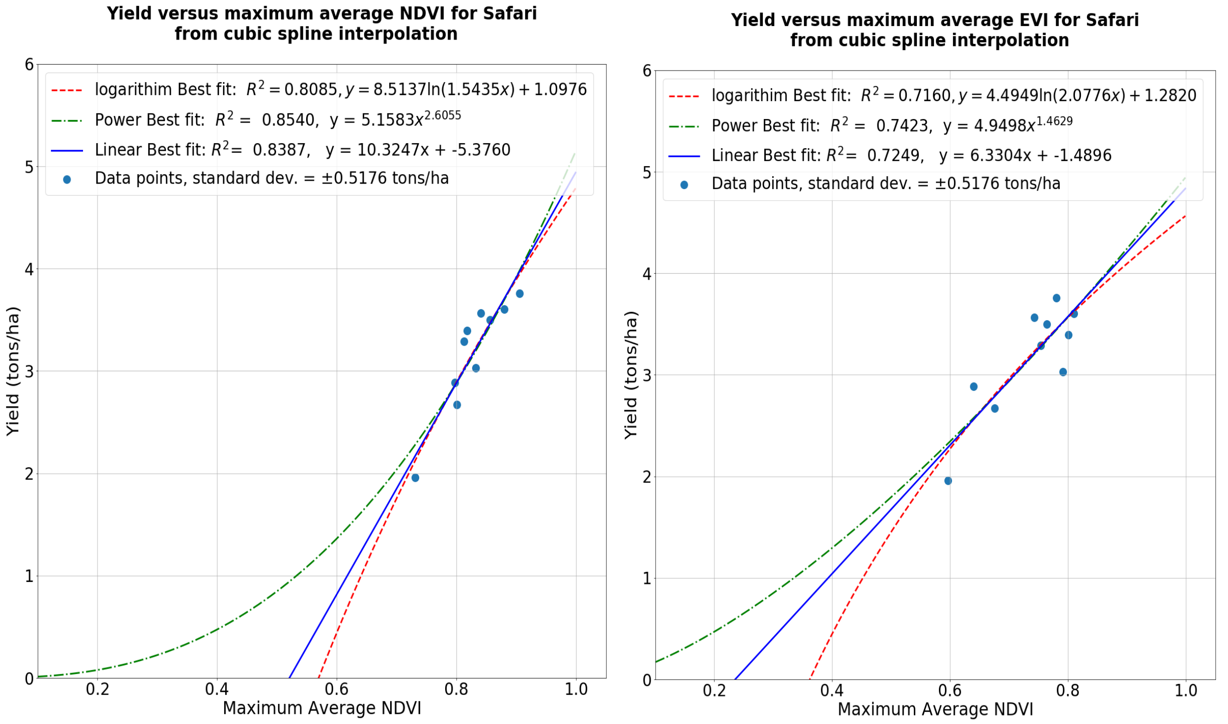

Yield versus maximum average NDVI for Dina from Gaussian process modeling

(a)

Yield versus maximum average EVI for Dina from Gaussian process modeling

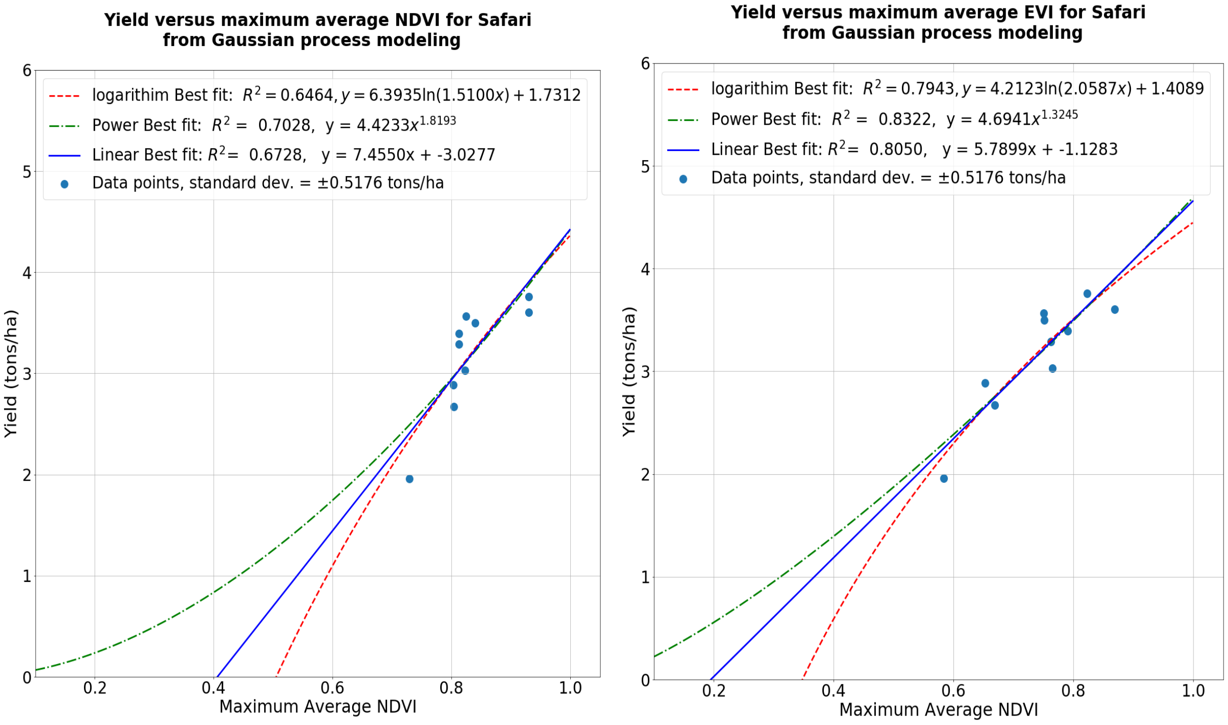
(b)

Figure 3.9: Linear, logarithm and power regression curves of the Gaussian process interpolated maximum average EVI and NDVI versus time (days) for Dina variety for the 2016/2017, 2017/2018 and 2018/2019 farming seasons.



(a) (b)

Figure 3.10: Linear, logarithm and power regression curves of the cubic spline interpolated maximum average EVI and NDVI versus time (days) for SC-Safari variety for the 2016/2017, 2017/2018 and 2018/2019 farming seasons.



(a) (b)

Figure 3.11: Linear, logarithm and power regression curves of the Gaussian process interpolated maximum average EVI and NDVI versus time (days) for SC-Safari variety for the 2016/2017, 2017/2018 and 2018/2019 farming seasons.

3.6 What can extrapolation of the regression models tell us about each variety?

Forward and backward extrapolation of the linear, logarithm and power regression curves were made in order to observe how they behave at the maximum and minimum possible EVI and NDVI and the minimum possible yield value (0 tons/ha) of plants. Since yield can not go below 0 tons/ha, it is worthless to extrapolate backward beyond this point. By analysing these plots we observe that the value of the average maximum EVI and NDVI values at 0tons/ha is a significant property of a soybean variety, i.e., it is the theoretical maximum average EVI or NDVI below which the variety fails to give any yield. On the other hand, the maximum possible EVI and NDVI value; 1.0, was chosen for analysis because it represents the highest level of near-infrared reflectance and red absorbance a canopy can achieve when analysing it using these two vegetation indexes. It is, in essence, the maximum photosynthetic ability that the canopy can attain and therefore represents its most productive state. Thus, the forward extrapolation was used to predict the maximum potential yield of each of the three varieties at the point when the leave sare at their most productive state.

It is important to note that NDVI and EVI are normalised physical quantities that range between -1.0 and 1.0, with green plants ranging between 0.2 and 1.0. Anything with an NDVI or EVI value outside the 0.2 to 1.0 range is not a green plant. Hence, the maximum possible chlorophyll level that a photosynthetically active canopy of any plant can attain is 1.0 EVI and NDVI. Conversely, the least possible chlorophyll level of any plant canopy is 0.2 EVI and NDVI. This means extrapolation of the linear, logarithm and power functions best-fit curves must stay within this interval for them to remain relevant. Since yield versus maximum average EVI and NDVI plots show that yield increased with increased EVI and NDVI values of soybean canopies, it is fair to conclude that the theoretical maximum potential yield returns of the varieties occur at 1.0 EVI and NDVI, which is the maximum possible EVI and NDVI value. The positive correlation of yield with canopy reflectance has been documented in one form or another in many studies for soybean and other grain crops, including those done by Emmanouil et al. 2016 [79], Israel Zelitch 1982 [80], Zhang et al. 2019 [81] and B. Ma 2001 [37]. These studies used either Unmanned Aerial Vehicles (UAVs) or Earth observation satellites. They did not attempt to analyse the separate varieties, analyse many different mathematical functions, nor extrapolate the regression curves to try and infer more information from them.

Forward extrapolation of the power functions gave the highest maximum potential yield predictions out of the three mathematical models used. Linear forward extrapolation gave the second-highest and logarithm forward extrapolation gave the lowest maximum potential yields for all plots when varieties were analysed separately. Backward extrapolation of the logarithm function gave the highest EVI, and NDVI values at which the yield was 0 tons/ha, while linear, backward extrapolation gave the second-highest, and backward power function extrapolation gave the least in all cases. According to figures 3.6 to 3.11, power functions only extinct to yield = 0 tons/ha at 0.0 maximum average EVI and NDVI values, but no green plant exists below 0.2 EVI and NDVI. Hence, only the 0.2 to 1.0 interval of the NDVI and EVI interval is valid in the forward and backward extrapolation of the power function. Observing yield values at 0.2 EVI and NDVI of a power function suggests that a field that has these values still gives some yield. This is in sharp contrast to the Logarithm and Linear backward extrapolations that suggest that a 0.2 maximum average EVI and NDVI value results in a yield = 0 tons/ha in most of the backward extrapolations of figures 3.6 to 3.11. Data from failing soybean fields is necessary to ascertain which of these models is accurate in this regard. The power function gave the best fit to the shape of the distribution of data points in figure 3.5 (b) and (d) when analysing all the varieties together. This excellent fit of a power function to yield versus NDVI was also observed by B.

L. Ma et al. in a 2001 study [37]. They found a power relationship between soybean yield and NDVI when they observed at the R5 stage of the soybean crop's growth cycle.

Table 3.4: Comparison of the maximum potential yield observed by each variety's developers to that predicted using linear, logarithm and power regression of yield versus maximum average EVI/NDVI obtained from Gaussian process modelling for each of the three soybean varieties.

Variety	**Observed max. yield (tons/ha)**	**Extrapolated maximum yield(tons/ha)**					
		EVI			**NDVI**		
		linear	logarithm	power	linear	logarithm	power
Dina	4.5	3.12±0.58	2.88±0.58	3.15±0.58	3.02±0.58	2.83±0.58	3.12±0.58
Safari	above 5.0	4.66± 0.52	4.45±0.52	4.69 ±0.52	4.43±0.52	4.37±0.52	4.42±0.52
Spike	above 4.5	4.30±0.49	4.16±0.49	4.31±0.49	5.34±0.49	5.20±0.49	5.63±0.49

Table 3.5: Comparison of the maximum potential yield observed by each variety's developers to that predicted using linear, logarithm and power regression of yield versus maximum average EVI/NDVI obtained from cubic spline modelling for each of the three soybean varieties.

Variety	**Observed max. yield (tons/ha)**	**Extrapolated maximum yield(tons/ha)**					
		EVI			**NDVI**		
		linear	logarithm	power	linear	logarithm	power
Dina	4.5	3.06±0.58	2.83±0.58	3.08±0.58	3.23±0.58	3.00±0.58	3.50±0.58
Safari	above 5.0	4.82± 0.52	4.57±0.52	4.95 ±0.52	4.95±0.52	4.80±0.52	5.16±0.52
Spike	above 4.5	4.30±0.50	4.14±0.50	4.31±0.50	5.18±0.50	5.08±0.50	5.35±0.50

The behaviour of the linear and logarithm functions at yield = 0 tons/ha was analysed. A look at plots of yield versus maximum average EVI and NDVI in figures 3.6 to 3.11 reviews that the NDVI model predicts that SC-Spike has a higher maximum average NDVI at which it gives yield = 0tons/ha compared to SC-Safari. Dina has the least. What this means is that an SC-Spike variety's canopy will need to reach a higher average NDVI in order to give yield. In comparison, a Dina variety will produce yield at a much lower average NDVI, thereby making Dina need less chlorophyll levels in its canopy to remain productive compared to the other two varieties. It could also mean Dina is more resilient to stresses that diminish chlorophyll levels in soybean canopies than Spike and Safari. These observations are in keeping with the fact that Dina is reported to be very drought-resistant by MRI-Syngenta, its developers [76]. The EVI equivalent of this analysis by looking at figures 3.7, 3.9 and 3.11 show extrapolations in EVI going lower than 0.2 to EVI levels that are not possible for plants to exist in. The EVI analysis, in general, also showed lower R^2 values compared to the NDVI ones. A variety like SC-Spike is chiefly desired for its high yield potential at the expense of stress resilience. In contrast, a variety like Dina is desired for its biotic and abiotic stress resilience at the expense of some yield potential. Tables 3.6 and 3.7 show the maximum average NDVI and EVI values at which the three mathematical models gave a potential yield value of 0tons/ha according to backward extrapolation of linear and logarithm functions of the yield versus maximum average EVI and NDVI.

The behaviour of power functions at 0.2 EVI and NDVI values in figures 3.6 to 3.11 was observed for each variety in order to determine their yields at the lowest EVI and NDVI value of their canopies. The analysis of SC-Spike for both the cubic spline and Gaussian process model's yield versus maximum NDVI placed the yield for 0.2 NDVI at 0.1tons/ha. The EVI equivalent of this analysis had much lower R^2. The analysis of Dina using cubic spline yield versus maximum average NDVI placed the yield at 0.2 NDVI at 0.2 tons/ha. The Gaussian process yield versus maximum average NDVI for Dina placed the yield at 0.2 NDVI at 0.25 tons/ha. The EVI equivalent of this places the yield of Dina at 0.2 EVI at 0.5 tons/ha for both the Gaussian process and cubic spline interpolation analyses. The analysis of SC-Safari using the cubic spline yield versus maximum average NDVI placed the yield at 0.2 NDVI at 0.1 tons/ha. The Gaussian process yield versus maximum average NDVI for Dina placed the yield at 0.2 NDVI at 0.3 tons/ha. The EVI equivalent of this placed the yield of SC-Safari at 0.2 EVI at 0.5 tons/ha in both the Gaussian process and cubic spline interpolation analyses. Therefore the power function ranking of the three varieties in terms of ability to yield at the lowest possible EVI and NDVI value placed Dina as the highest yielding followed by SC-Safari and SC-Spike respectively.

Table 3.6: Maximum average EVI and NDVI at yield = 0 tons/ha from the linear, logarithm and power regression of yield versus cubic spline interpolated maximum average EVI and NDVI.

Variety	**Extrapolated maximum average EVI/NDVI at yield = 0 tons/ha**					
	EVI			**NDVI**		
	linear	logarithm	power	linear	logarithm	power
Dina	0.079±0.140	0.250±0.140	0.000±0.140	0.376±0.099	0.434±0.099	0.000±0.099
Safari	0.235±0.070	0.362±0.070	0.000±0.070	0.521±0.059	0.570±0.059	0.000±0.059
Spike	0.044±0.073	0.248±0.073	0.000±0.073	0.552±0.038	0.602±0.038	0.000±0.038

Table 3.7: Maximum average EVI and NDVI at yield = 0 tons/ha from the linear, logarithm and power regression of yield versus Gaussian process interpolated maximum average EVI and NDVI.

Variety	**Extrapolated maximum average EVI/NDVI at yield = 0 tons/ha**					
	EVI			**NDVI**		
	linear	logarithm	power	linear	logarithm	power
Dina	0.103±0.136	0.261±0.136	0.000±0.136	0.278±0.117	0.372±0.117	0.000±0.117
Safari	0.195±0.080	0.0.348±0.080	0.000±0.080	0.4061±0.057	0.505±0.057	0.000±0.057
Spike	-0.041±0.064	0.256±0.064	0.000±0.064	0.557±0.035	0.601±0.035	0.000±0.035

The ranking of the three soybean varieties in terms of ability to yield at very low average EVI and NDVI values of their canopies by the backward extrapolation of the linear logarithm and power functions gave similar results as those observed in the literature. The literature placed Dina as the most stress-resilient to environmental stress of these three varieties followed by SC-Safari, SC-Spike was the least stress-resilient [76]. In the next section, we will attempt to determine the maturity periods of the three soybean varieties from their average EVI and NDVI time series profiles.

3.7 What is the Maturity period of each variety?

Since soybean is a short-day crop, its maturity is primarily determined by variety-specific day length requirements that initiate flowering. Most studies that measure the maturity period do it by counting the number of days from planting to physiological maturity, i.e. just when a significant percentage, e.g. 95% of the soybean crops in the field becomes ready for harvest. In this study, we used the time interval between equal EVI and NDVI values of the time series profiles to determine the maturity periods of the three soybean varieties.

Using the FWHM (full width at half the maximum) as a means for determining maturity period proved futile as there was no significant difference in FWHM values between the different varieties. Average FWHM of Safari computed from EVI was found to be 85 days, that of Spike was found to be 83 days, and that of Dina was found to be 86 days. This remained relatively the same in the NDVI analysis with SC-Safari showing giving an average FWHM of 84 days, SC-Spike gave 84 days, and Dina gave 87 days.

A more effective way of measuring maturity period was done by obtaining the lowest possible NDVI and EVI values of the canopies at germination and senescence when most of the ground was not covered by the crops canopy and determining the time interval between them. The resulting time periods for each variety were then averaged to gain their average maturity periods. NDVI and EVI values are not directly equal to each other, i.e., $0.2\text{NDVI} \neq 0.2\text{EVI}$ since EVI corrects for aerosol and soil effects while NDVI does not. Hence, NDVI values tend to saturate much quickly and are usually larger than their corresponding EVI values. For this reason, different values of EVI and NDVI were used to measure the average maturity periods, EVI = 0.2 was used for EVI analysis, while NDVI = 0.3 in the NDVI analysis. NDVI = 0.3 was chosen because it was large enough to avoid the perturbations that occur in the NDVI profile at germination period caused by noise due to irrigation, weeding, and atmosphere effects since NDVI is very susceptible to soil and water effects in the field. Table 3.8 shows the resulting maturity period of the three varieties. Dina showed the most extended maturity period, SC-Spike had the second-longest and Safari had the shortest in all the analyses.

Table 3.8: Average maturity period of each variety obtained from the average EVI and NDVI's cubic spline and Gaussian process time profiles

variety	**EVI = 0.2**		**NDVI = 0.3**	
	Cubic spline	Gaussian Process	Cubic spline	Gaussian Process
Spike	123	121	118	117
Safari	105	100	111	111
Dina	131	136	145	147

Maturity periods of the various soybean varieties in this study were obtained from the websites of the institutions that developed them. The observed maturity period for SC-Safari as reported by SeedCo is 125 days, while that for Dina is 102 days. The maturity period for SC-Spike was not found in any of the literature reviewed during this study. One of the farmers interviewed during the placement with IITA placed it at 120 days. As can be observed by comparing maturity periods reported by the seed developers with those in table 3.8 above, the maturity periods observed using the remote sensing process developed in this study were different from those observed by the seed developers. This is because the process used to measure maturity period in ground observations measures the time interval from germination to a point when a significant percentage of the plants in the field were fully grown, dried up, and ready for harvest. The process developed in this study measured the time interval for the NDVI and EVI at germination to recure during senescence when the crops were ready for harvest. In a practical sense, the process used in this study is far less ambiguous than the process used in ground observations by these seed-producing organisations since it can be hard to approximate for example when 95% of the crops in a field are mature. It gives a definite EVI or NDVI value, unlike the process used by farmers and breeders which depends on eye observation of whether the entire field is dry enough for harvest or not. A field may be too big to observe properly by eye. However, the process developed in this study may be susceptible to noise caused by atmospheric and soil effects, especially when using NDVI. Maturity period of a soybean variety is very important especially since most subsistence farmers rely on rainfed agriculture in an area where the the rainy season is shrinking in size. Thus early maturing soybean varieties are becoming increasingly important.

3.8 Discussion

The image analysis pipeline and the EVI and NDVI times-series profiles derived using cubic spline interpolation, and Gaussian process modelling proved to be a good source of information on remote sensed agronomic attributes of the three soybean varieties that were studied. The analysis made in this study observed attributes that are very close to those observed by the breeders of these varieties. The model did exceptionally well in predicting the attributes of SC-Spike and SC-Safari. However, there were some inconsistencies between the attributes observed here and those reported by the developers of Dina. For example, the yield potential of Dina is reported to be about 4.5 tons/ha, but this study placed it between 2.8 and 3.12 tons/ha. Its maturity period is stated to be 102 days, but this study placed it in the range of 131 to 147 days.

3.8.1 Challenges encountered

Initially, this research was supposed to study experimental soybean lines grown at the IITA soybean breeding facilities. The varieties were grown on thousands of small 3m x 3m plots at their Southern Africa Research and Administration Hub (SARAH) in Lusaka Zambia. These plots would have provided a massive well-documented dataset to study. However, the fact that the plots were 3m x 3m, and were only separated by 80cm from each other made it impossible to study them using any of the satellite constellations that provided affordable images for such research at the time. This was because it is impossible to separate the individual plots from each other using the spatial resolutions that were offered by these satellites. The ever-changing technological landscape provides hope that affordable satellite images will eventually become available that will permit the analysis of such plots.

PlanetScope images had a download cap on them of 10, 000 km^2 per month. This was an enormous challenge that made the image analysis process go very slowly, once the downloading cap was reached, the only time more images could be downloaded was when the next month began. Another major challenge was the slow and sometimes unreliable feedback from some farmers regarding the yield from their fields. Some of the farmers that were visited during the placement with IITA did not report their yield data. The fact that there were very few satellites in the PlanetScope constellation during the 2015/2016 farming season was yet another enormous challenge since it meant a terribly reduced temporal resolution, and so all yield and vegetation index data from the fields that season was not analysed.

3.8.2 Potential sources of error

The low R^2 values observed in the regression of yield versus EVI can be attributed to the fact that the soil and atmosphere correction factors in the formula were not developed for specific use with PlanetScope, Sentinel-2 and Landsat-8 images, they were developed specifically for the MODIS satellite which has a spatial resolution of 250m while PlanetScope, Sentinal-2 and Landsat-8 satellite images used in this study have spatial resolutions of 3m, 10m and 20m respectively. The use of satellites with different spatial resolutions to form a single constellation may on its own have contributed to the errors as satellites with a lower spatial resolutions observe lesser spatial details compare to those with a higher spatial resolution. Other potential sources of error include impurities due to misclassifications by the cloud masking algorithm. In hindsight, the disqualification of fields that experienced extreme crop failure while visiting the fields was a bad idea as these would have better helped to observe the mathematical models at 0 tons/ha. The fact that each variety was dominated by a single farm is a possible primary source of error. As far as this study is concerned the managerial techniques carried out by all the farmers seemed to be fairly the same. However, it is also possible that the small differences in these managerial techniques and environmental conditions also influenced some of the traits observed by in these varieties.

3.8.3 Potential improvements to the yield prediction model

One of the best and most obvious ways to improve this canopy analysis pipeline is to increase the number of satellites in the observing constellation and their spatial resolutions so that the fields can be tracked with higher frequency and greater spatial detail. Another reason to improve their spatial resolutions is so that smaller fields can be included in the analyses. Increasing temporal resolution would be especially useful for Gaussian process modelling because it requires many more data points than cubic spline interpolation for it to make the best predictions. One satellite constellation that can prove useful with improving temporal resolution for large fields is MODIS. It has an almost perfect revisit time. The two satellites in its constellation, Terra MODIS and Aqua MODIS view the entire Earth in 1 to 2 days in 36 spectral bands. Its high spectral resolution would help include other spectral indexes so that more biophysical, ecophysiological and biochemical properties of a canopy can be studied. It can not, however, be used for small fields because of its spatial resolution of 250m per pixel. Increasing the number of satellites can also help capture more cloud and haze-free days, thereby improving the accuracy of the analysis during cloud infested periods of the farming seasons. One other potential improvement to the model is the inclusion of a leaf water spectral index, Normalized Difference Water Index (NDWI). This could improve accuracy in determining the maturity period since it tracks the water content of the canopy. It was not possible to use this spectral index in this study with PlanetScope images as they did not observe in the shortwave infrared spectral band, which is vital to computing this index. The study suffered severely from minimal ground level observations. More ground truth data collection would inevitably improve the data collection process as it would compliment the remote sensing. Other ways of observing soybean fields that can improve the results include observing the EVI and NDVI per unit area of the fields and the median EVI and NDVI of a field instead of the average EVI and NDVI used in this study. Such analyses would produce cubic spline and Gaussian process curves that are much like the ones observed in this study but could reduce the noise in the data. Another crucial thing to improve upon in future is to make the same analysis in this research for different varieties under precisely the same growing conditions. The best place to find many different varieties under the same managerial and environmental conditions is a breeding farm with many experimental trial plots.

3.8.4 Potential Utilisation of the yield and maturity prediction model

The characteristic equations that resulted from the curve fitting of yield against average EVI and NDVI in this study can be an essential tool in assessing the yield potentials of different soybean varieties. In a situation where a satellite constellation with adequate spatial resolution exists, the processes developed here can be useful in improving the speed of phenotypically identifying the best performing soybean varieties with minimal resources. They can be an excellent tool for farmers, insurance officers, and breeders alike in crop monitoring for yield evaluation. However, these yield prediction are very susceptible to wrong predictions in conditions when disease, pests, or weeds attack a field understudy after the maximum average EVI and NDVI, especially in the case of small research plots and smallholder farms. Although such a scenario would result in a strange decent of the vegetation index time series curve, which would serve as an alarm to show that the crop became stressed after reaching its maximum average EVI or NDVI and the results from that observation would not be trusted. The techniques employed in this study also provide a secure and effective means of estimating the maturity period of new varieties of soybean. Breeders only need to analyse the average time it takes for one variety of soybean grown on several plots to move from a specific EVI or NDVI value at the start of the growing season to recur at the end of the growing season. The maturity periods of each experimental plot so obtained are then averaged to get the average maturity period of that variety. Taking all these factors into account, the image analysis pipeline and EVI and NDVI time profile analysis presented in this study could significantly reduce the time it takes for new soybean varieties to be released to the general public if organisations such as IITA, Zamseed and Syngenta incoperated it into their soybean breeding processes.

Chapter 4

Summary, Conclusions and Recommendations

4.1 Summary and Conclusions

In this research we developed an analysis procedure that can be developed further for use in establishment of a rapid variety breeding process for soybean, it was observed that combining Landsat-8, Sentinel-2, and PlanetScope satellites into a virtual constellation provided a good, accurate, and straightforward way of tracking the daily chlorophyll levels of a soybean canopy using EVI and NDVI. We were severely restricted by low spatial and temporal resolution. As earth observation satellites continue to evolve this limitation will become a thing of the past and it will become possible to spatially and temporally resolve the smaller breeding plots used in the soybean breeding process. Given the complexity of the interaction of soybean canopies with photosynthetic photon flux density, it was shown in this study that NDVI and EVI spectral indexes simplify the process by measuring the rate of reflectances of the photon flux density on a scale ranging from -1 and 1 with plants existing in the range 0.2 to 1.0 for both vegetation indexes. Cubic spline interpolation and Gaussian process modelling both proved suitable for tracking these vegetation indexes across all the farming seasons.

The various features of the EVI and NDVI time-series profiles such as the area under the curve, the FWHM, the gradients of ascent and descent at the FWHM and the NDVI and EVI values at the FWHM and maximum values of the time series profiles were analysed for linear correlation with the yield for all the three varieties together. Only the maximum EVI and NDVI values of the time-series profiles showed a significantly high positive correlation with the yield for the linear, logarithm and power models. R^2 values of these plots ranged between 0.5590 and 0.5632 for maximum EVI obtained from cubic spline interpolation versus yield and 0.5842 and 0.5902 for maximum EVI obtained from Gaussian process modelling versus yield. The R^2 values of the NDVI equivalent of this analysis were slightly higher. Ranging from 0.7168 to 0.7721 in the NDIV cubic spline interpolations, and 0.6468 and 0.6813 in the NDVI Gaussian process regression. Thus NDVI showed much higher correlation than EVI when all the varieties were analysed together using linear logarithm and power regression. Cubic spline maximum EVI and NDVI showed higher correlation than their Gaussian process counterparts. By default, the NDVI and EVI values at the FWHM were also correlated to yield since they are simply half of the Maximum values. Full-width half maximum showed a small correlation that did not exceed $R^2 = 0.3$.

Analysing the varieties one by one resulted in much smaller standard deviations than when they were analysed together. The computed maximum potential yields of the three varieties using the linear, logarithm, and power regressions were close to those reported by the companies that developed them. This was especially so for SC-Spike and SC-Safari. According to the backward and forward extrapolations of the three mathematical models when analysing each variety singularly, the logarithm and linear functions were more accurate processes to use for modelling the relationship between yield and the maximum average NDVI and EVI. They did not overshoot beyond 0.2 NDVI and EVI in most of the extrapolations. This was particularly so in graphs that showed R^2 values greater than 0.5. The forward extrapolations of these mathematical models placed SC-Spike as having the highest yield potential and Dina as having the lowest. On the other hand, backward extrapolation placed Dina as the variety that would produce yield with the most moderate Chlorophyll content in the biomass of its canopy. At the same time, backward extrapolation of these mathematical models predicted that SC-Spike needs much more chlorophyll content in its biomass than SC-Safari in order to yield. In turn, SC-Safari needs more Chlorophyll content in its canopy than Dina in order to yield.

The maturity period prediction model ranked Dina as requiring the longest time to reach maturity, followed by SC-pike while SC-Safari needed the least time. Table 4.1 ranks the three soybean varieties according to maximum yield potential and maturity period. One is the highest rank, and three is the lowest rank. So that in the case of maximum potential yields, one is the highest yielding, and three is the least. In the case of maturity periods, one is the earliest maturing, and three is the latest. Early maturing varieties are desirable since the rainy season is shortening in Zambia.

Table 4.1: Ranking the three varieties according to maximum yield potential and maturity period based on the average NDVI exploratory data analysis with 1 being the highest score and 3 being the lowest

Variety	**Maximum potential yield**	**Maturity period**
Spike	1	2
Dina	3	3
Safari	2	1

4.2 Recommendations

The findings of this study underscore the importance of using high temporal, spatial and spectral resolution remote sensing techniques in soybean yield estimation and in improvement of the speed of breeding new varieties that are compatible with Sub-Saharan Africa's conditions. These techniques promise to alleviate some of the challenges breeders face logistically and financially. They can help them determine specific agronomic attributes of new varieties much faster than is currently done. This study demonstrates the potential applications of NDVI and EVI and other spectroscopic analyses as a means of diagnosing the state of crops in a field and predicting yield potential and maturity period of newly bred soybean strains. In future work, it's recommended to monitor other varieties to see if the same traits observed for Dina SC-Safari, and SC-Spike can be observed with them and to incorporate the techniques in this study into field monitoring apps for farmers and breeders in Sub Saharan Africa. This type of application can be expanded to include other grain crops such as maize and wheat. There is an immense entrepreneurial opportunity in coming up with such a piece of software. The profiling of the chlorophyll levels of soybean canopies measured using EVI and NDVI is useful for observing how the plants of a field are doing and at what stage of growth they are. With the current increase in earth observation satellite technology, this will become ever more readily available so that farmers and extension officers will be tracking the photosynthetic capability of their fields remotely. Such an analysis will also make it possible for smaller groups of extension workers to analyse more farms even without having to visit the farms in person for a long time. Satellite data is becoming ever more available for such analysis and their spatial spectral and temporal resolutions are becoming better with every passing year. A very important study that can be considered using the kind of remote sensing analysis done in this study is the relationship between the details of yield, i.e., oil content, protein content, the number of grains, etc. with farm additives like fertilisers, rhizobia inoculants, pesticides, soil water content, etc. Such an analysis will look more at the causes of the correlation between yield and EVI and NDVI. Parts of fields can be sectioned and studied separately, tracking their chlorophyll levels of each section across the farming season while monitoring the levels of the various nutrients and pest/stress levels in them.

The possibilities for the economic emancipation of Africa through agriculture are as vast as the continent itself. To the vibrant people that work tirelessly to see this future realised the author of this thesis quotes Alan Kay (1971) “The best way to predict the future is to invent it”.

Bibliography

[1] Adelodun Kolapo. *Soybean: Africa's Potential Cinderella Food Crop.* 04 2011.

[2] Julius H. Kotir. Climate change and variability in sub-saharan africa: a review of current and future trends and impacts on agriculture and food security. *Environment, Development and Sustainability*, 13(3):587–605, Jun 2011.

[3] Jane Payumo, Evelyn Lemgo, and Karim Maredia. Transforming sub-saharan africa's agriculture through agribusiness innovation. *Global Journal of Agricultural Innovation, Research and Development*, 4:1–12, 07 2017.

[4] Deepak Ray, Nathaniel Mueller, Paul West, and Jonathan Foley. Yield trends are insufficient to double global crop production by 2050. *PloS one*, 8:e66428, 06 2013.

[5] Stephen Long, Amy Marshall-Colon, and Xin-Guang Zhu. Meeting the global food demand of the future by engineering crop photosynthesis and yield potential. *Cell*, 161:56–66, 03 2015.

[6] Enoch Sapey Dalia Mohamedkheir Khojely, Seifeldin Elrayah Ibrahim and Tianfu Han. History, current status, and prospects of soybean production and research in sub-saharan africa. *The Crop Journal*, 2018.

[7] J. O'S. Farming in africa, the boundless scope for the continent's agricultural expansion. *The Economist*, 2013.

[8] Food and Agriculture Organization of the United Nations. Production share of soybeans by region average 1994 - 2017. http://www.fao.org/faostat/en/data/QC/visualize, 12 2019.

[9] Christoph Müller. Climate change impact on sub-saharan africa: an overview and analysis of scenarios and models. *Discussion Paper / Deutsches Institut für Ent-wicklungspolitik*, March 2009.

[10] Abhilash Singh Chauhan, Raman Sharma, Aashish Kumar, Kapil Malik, and Harender Dagar. *APPLICATIONS OF REMOTE SENSING IN AGRICULTURE*, pages 141–146. 01 2018.

[11] William Shurtleff and Akiko Aoyag. History of soy in africa - part 1, http://www.soyinfocenter.com/HSS/africa1.php, 2019.

[12] Antony Chapoto, Brian Chisanga, and Mulako Kabisa. Zambia agriculture status report 2017, 01 2018.

[13] TechnoServe. Soy value chains. http://espanol.technoserve.org/project/soy-value-chains, 01 2020.

[14] Brivery Siamabele. Soya beans production in zambia: Opportunities and challenges. *American Journal of Agricultural and Biological Sciences*, 14:55–60, 01 2019.

[15] Aditya Pratap, S. Gupta, Jitendra Kumar, and Ramesh Solanki. *Soybean*, volume 1, pages 293–321. 01 2012.

[16] Sichilima I., Mebelo Mataa, and Alice Mweetwa. Morpho-physiological and yield responses associated with plant density variation in soybean (glycine max l. (merrill)). *International Journal of Environment, Agriculture and Biotechnology*, 3:274–285, 01 2018.

[17] Seifeldin Ibrahim. Agronomic studies on irrigated soybeans in central sudan: I. effect of plant spacing on grain yield and yield components. *International Journal of AgriScience*, 2:733–739, 08 2012.

[18] A.P. Uriyo F.R. Ntare. *Soybean Training Manual*, volume 10. IITA, 1970.

[19] Jonas Chianu, E M. Nkonya, Franklin Mairura, Justina Chianu, and Festus Akinnifesi. Biological nitrogen fixation and socioeconomic factors for legume production in sub-saharan africa: A review. *Agronomy for Sustainable Development*, 31:139–154, 01 2011.

[20] Naoki Matsuo, Tetsuya Yamada, Yoshitake Takada, Koichiro Fukami, and Makita Hajika. Effect of plant density on growth and yield of new soybean genotypes grown under early planting condition in southwestern japan. *Plant Production Science*, 21:1–10, 02 2018.

[21] Mariangela Hungria and Iêda Mendes. *Nitrogen Fixation with Soybean: The Perfect Symbiosis?*, pages 1005–1019. 07 2015.

[22] Anna Chlingaryan, Salah Sukkarieh, and Brett Whelan. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, 151:61–69, 08 2018.

[23] Joel Ransom, Abdel Babiker, and George Odhiambo. *Integrating crop management practices for Striga control*, pages 213–228. 06 2007.

[24] T R. Sinclair, L C. Purcell, Vincent Vadez, Rachid Serraj, Andy King, and R Nelson. Identification of soybean genotypes with n fixation tolerance to water deficits. *Crop Science*, 40:1803–1809, 01 2000.

[25] Chaoyang Wu, Zheng Niu, Quan Tang, and Wenjiang Huang. Estimating chlorophyll content from hyperspectral vegetation indices: Modeling and validation. *Agricultural and Forest Meteorology*, 148:1230–1241, 07 2008.

[26] Cunxiang Wu, Qibin ma, Kwan Yam, Ming-Yan Cheung, Yunyuan Xu, Tianfu Han, Hon-Ming Lam, and Kang Chong. In situ expression of the gmnmh7 gene is photoperiod-dependent in a unique soybean (glycine max [l.] merr.) flowering reversion system. *Planta*, 223:725–35, 04 2006.

[27] Peter Sexton handiwe Nleya and Kyle Gustafson. *Soybean Growth Stages.* O'Reilly Media, Inc., 1st edition, 03 2019.

[28] SeedCo. Soyabean growers guide, 2017.

[29] Frederick Baijukya, Harun Murithi, and Fred Kanampiu. *Improving cultivation practices for soybeans in sub-Saharan Africa*, pages 105–122. 01 2018.

[30] Jegor Miladinovic, Joe Burton, Svetlana Balesevic Tubic, Dragana Miladinović, Vuk Djordjevic, and Vojin Đukić. Soybean breeding: Comparison of the efficiency of different selection methods. *Turkish Journal of Agriculture and Forestry*, 35:469–480, 10 2011.

[31] Hailu Tefera, Alpha Kamara, Baffour Asafo-Adjei, and K. Dashiell. Improvement in grain and fodder yields of early-maturing promiscuous soybean varieties in the guinea savanna of nigeria. *Crop Science*, 49, 11 2009.

[32] Paul Pinter, Jerry Hatfield, J.s Schepers, Edward Barnes, M Susan Moran, Craig Daughtry, and Dan Upchurch. Remote sensing for crop management. *Photogrammetric Engineering and Remote Sensing*, 69, 06 2003.

[33] Jegor Miladinovic, Joe Burton, Svetlana Balesevic Tubic, Dragana Miladinović, Vuk Djordjevic, and Vojin Đukić. Soybean breeding: Comparison of the efficiency of different selection methods. *Turkish Journal of Agriculture and Forestry*, 35:469–480, 10 2011.

[34] Xiaoyan Zhang, Jinming Zhao, Yang Guijun, Jiangang Liu, Jiqiu Cao, Chunyan Li, Xiaoqing Zhao, and Gai Junyi. Establishment of plot-yield prediction models in soybean breeding programs using uav-based hyperspectral remote sensing. *Remote Sensing*, 11:2752, 11 2019.

[35] Neil Yu, Liujun Li, Nathan Schmitz, L.F. Tian, Jonathan Greenberg, and Brian Diers. Development of methods to improve soybean yield estimation and predict plant maturity with an unmanned aerial vehicle based platform. *Remote Sensing of Environment*, 187, 12 2016.

[36] Thomas R Sinclair, Larry C Purcell, and Clay H Sneller. Crop transformation and the challenge to increase yield potential. *Trends in plant science*, 9:70–5, 03 2004.

[37] L. T. Evans and R. A. Fischer. Yield potential: Its definition, measurement, and significance. *Crop Science*, 39(6):1544–1551, 1999.

[38] Russell G. Congalton Rebecca L. Dodge. *Meeting Environmental Challenges with Remote Sensing Imagery.* American Geosciences Institute, 2012.

[39] J.G.P.W. Clevers. A simplified approach for yield prediction of sugar beet based on optical remote sensing data. *Remote Sensing of Environment*, 61:221–228, 1997.

[40] Xiangjin Wei, Junfeng Xu, Hongnian Guo, Ling Jiang, Saihua Chen, Chuanyuan Yu, Zhenling Zhou, Peisong Hu, Huqu Zhai, and Jianmin Wan. Dth8 suppresses flowering in rice, influencing plant height and yield potential simultaneously. *Plant Physiology*, 153(4):1747–1758, 2010.

[41] David Helman, Idan Bahat, Yishai Netzer, Alon Ben-Gal, Victor Alchanatis, Aviva Peeters, and Y. Cohen. Using time series of high-resolution planet satellite images to monitor grapevine stem water potential in commercial vineyards. *Remote Sensing*, 10, 10 2018.

[42] James Campbell and Randolph Wynne. *Introduction to Remote Sensing.* 01 2011.

[43] Edward Kairu. An introduction to remote sensing. *GeoJournal*, 6:251–260, 05 1982.

[44] Planet Labs. Planet imagery product specifications, August 2018.

[45] European Space Adminstration. Sentinel-2: Esa's optical high-resolution mission for gmes operational services, March 2012.

[46] Feng Gao, Martha Anderson, Craig Daughtry, and David Johnson. Assessing the variability of corn and soybean yields in central iowa using high spatiotemporal resolution multi-satellite imagery. *Remote Sensing*, 10:1489, 09 2018.

[47] Department of the Interior U.S. Geological Survey. *Landsat 8 (L8) data users handbook.* EROS Sioux Falls, South Dakota, April 2019.

[48] B. Ma, Lindsay Dwyer, Elroy Cober, and Malcolm Morrison. Early prediction of soybean yield from canopy reflectance measurements. *Agronomy Journal - AGRON J*, 93, 11 2001.

[49] Robert N. Colwell. Determining the prevalence of certain cereal crop diseases by means of aerial photography. volume 26, pages 223–286, 1956.

[50] B. R. BUTTERY and R. I. BUZZELL. The relationship between chlorophyll content and rate of photosynthesis in soybeans. *Canadian Journal of Plant Science*, 57(1):1–5, 1977.

[51] R. Kumar and L. Silva. Light ray tracing through a leaf cross section. *Appl. Opt.*, 12(12):2950–2954, Dec 1973.

[52] Xue Jinru and Baofeng Su. Significant remote sensing vegetation indices: A review of developments and applications. *Journal of Sensors*, 2017:1–17, 01 2017.

[53] Craig Daughtry, Vern Vanderbilt, and V. Pollara. Variability of reflectance measurements with sensor altitude and canopy type1. *Agronomy Journal - AGRON J*, 74, 01 1982.

[54] Tyler Nigon, David Mulla, Carl Rosen, Y. Cohen, Victor Alchanatis, Joseph Knight, and Ronit Rud. Hyperspectral aerial imagery for detecting nitrogen stress in two potato cultivars. *Computers and Electronics in Agriculture*, 112, 01 2015.

[55] Zhenong Jin, George Azzari, Marshall Burke, Stephen Aston, and David B. Lobell. Mapping smallholder yield heterogeneity at multiple scales in eastern africa. *Remote Sensing*, 9:931, 09 2017.

[56] G. Garik Gutman. Vegetation indices from avhrr: An update and future prospects. *Remote Sensing of Environment*, 35:121–136, February–March 1991.

[57] Naila Yasmin, Muhammad Khokhar, Sundus Tanveer, Zafeer Saqib, and Waseem Khan. Dynamical assessment of vegetation trends over margalla hills national park by using modis vegetation indices. *Pakistan Journal of Agricultural Sciences*, 53(4), 09 2016.

[58] Alfredo Huete. A soil-adjusted vegetation index (savi). *Remote Sensing of Environment*, 25:295–309, 08 1988.

[59] Bunkei Matsushita, Wei Yang, Jin Chen, Onda Yuyichi, and Qiu Guoyu. Sensitivity of the enhanced vegetation index (evi) and normalized difference vegetation index (ndvi) to topographic effects: A case study in high-density cypress forest. *Sensors*, 7, 11 2007.

[60] Zhangyan Jiang, Alfredo Huete, K. Didan, and Tomoaki Miura. Development of a two-band enhanced vegetation index without a blue band. *Remote Sensing of Environment*, 112:3833–3845, 10 2008.

[61] Python Software Foundation. General python faq. https://docs.python.org/3/faq/general.html, March 2020.

[62] Sebastian Raschka. *Python Machine Learning*. Packt Publishing, 2015.

[63] Sean Gillies. Rasterio documentation. 1, June 2019.

[64] Hankui Zhang, David Roy, L Yan, Zhongbin Li, Haiyan Huang, E Vermote, Sergii Skakun, and Jean-Claude Roger. Characterization of sentinel-2a and landsat-8 top of atmosphere, surface, and nadir brdf adjusted reflectance and ndvi differences. *Remote Sensing of Environment*, 04 2018.

[65] Mats Töpel. Tutorial for creating polygons in qgis, September 2014.

[66] Alice Zheng and Amanda Casari. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists.* O'Reilly Media, Inc., 1st edition, 2018.

[67] Siddhartha Khare and Sanjay Ghosh. Training module for qgis. 07 2017.

[68] Niels Raes and Maarten Zelfde. Introduction to qgis 2.14 - mapping species distributions. 08 2016.

[69] Andrey Chernov Natalya Vorobiova. Curve fitting of modis ndvi time series in the task of early crops identification by satellite images. In *Procedia Engineering*, volume 201, pages 184–195, 2017.

[70] Hankui Zhang, Jing Chen, Bo Huang, Huihui Song, and Yiran Li. Reconstructing seasonal variation of landsat vegetation index related to leaf area index by fusing with modis data. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 7:950–960, 03 2014.

[71] Vera Maiorova, Alexey Bannikov, Dmitriy Grishko, Igor Jarenov, Victor Leonov, Alexey Toporkov, and Alexander Harlan. Monitoring condition of agricultural fields based on prediction of ndvi with the use of multi-spectral and hyper-spectral data from space imagery. *Science and Education of the Bauman MSTU*, 13, 07 2013.

[72] Jus Kocijan. Gaussian process models for systems identification. 01 2008.

[73] Najmuddin Ahmad and khan Deeba. The study of new approaches in cubic spline interpolation for auto mobile data. *Journal of Science and Arts*, 17:401–406, 09 2017.

[74] M. Kuss. *Gaussian Process Models for Robust Regression, Classification, and Reinforcement Learning.* PhD thesis, Biologische Kybernetik, 2006. passed with distinction, published online.

[75] Berkley Walker, Darren Drewry, Rebecca Slattery, Andy VanLoocke, Young Cho, and Donald Ort. Chlorophyll can be reduced in crop canopies with little penalty to photosynthesis. *Plant Physiology*, 176:pp.01401.2017, 10 2017.

[76] Luiz Garcia, Ivo Frare, Thiago Inagaki, Pedro Weirich Neto, Marcos Martins, Maghnom Melo, Leandro Nadolny, Marcos Rogenski, Newton Filho, and Ezequiel Oliveira. Spacing between soybean rows. *American Journal of Plant Sciences*, 09:711–721, 01 2018.

[77] Jason Debruin and Palle Pedersen. Effect of row spacing and seeding rate on soybean yield. *Agronomy Journal - AGRON J*, 100, 05 2008.

[78] Syngenta Zambia. Soya bean varieties. https://www.syngenta.co.zm/soya-beans-varieties, 02 2020.

[79] Seedco Zambia Limited. Key agronomic attributes of seedco soyabean varieties. https://www.seedcogroup.com/zw/media/blog/key-agronomic-attributes-seed-co-soyabean-varieties, 02 2020.

[80] George C. Linderman and Stefan Steinerberger. Clustering with t-sne, provably. *CoRR*, abs/1706.02582, 2017.