Predicting soybean (*Glycine max (L.) Merr*) grain yield using remote sensing

Siphokazi Ruth Gcayi

A Dissertation submitted to the Faculty of Natural and Agricultural Sciences, at the University of the Free State, in fulfilment of the academic requirements for the degree of Master of Science in Geography

January 2019

Supervisor: Dr S.A. Adelabu

Co-supervisor: Dr J. G. Chirima

Abstract

Accurate yield statistics of soybean (*Glycine max (L.) Merr*) grain are important for planning the management of crops prior to harvests as well as putting together logistics covering transport of grain after harvest. These statistics are essential to farmers, government and other policy-makers for guiding important decisions related to expected yields. Conventional methods currently used in South Africa to obtain such crop yields statistics are unreliable, subjective and labour intensive. As such, they pose a risk on food security as decisions concerning total agricultural production in the country rely on them. Remote sensing as part of precision agriculture technologies can overcome challenges experienced in acquiring crop statistics. Remote sensing techniques offer real-time, objective, accurate and reliable crop statistics that can be used to derive yield information. The present study sought to examine the utility of remote sensing in predicting soybean grain yields. Specifically, the study investigated the utility of hyperspectral remote sensing data for predicting soybean grain yields. To realize this aim, the study was restricted to the following objectives: (i) evaluating the potential of narrow-band indices to predict soybean grain yield (ii) determining the suitable growth stage to predict soybean grain yield using hyperspectral data and (iii) assessing the ability of Sentinel-2 Multispectral Instrument (MSI) to estimate soybean grain yield from resampled hyperspectral data. Firstly, an evaluation of the potential of narrow-band indices in predicting soybean grain yield was achieved by comparing NDVI, SR and EVI, vegetation indices, derived from hyperspectral data. The results showed that the suitable bands to predict soybean grain yield were combinations situated in the red-edge (680-750 nm), NIR and largely on the MIR (1300 to 2399 nm) of the electromagnetic spectrum. Similarly, the results showed that SR better predicted soybean grain yield ($R^2 = 0.843$) as compared to NDVI and EVI that yielded an $R^2 = 0.841$ and $R^2 = 0.537$ respectively. Secondly, as a way of determining the most suitable growth stage for predicting soybean grain yield, the study investigated the flowering, pod formation, and seed filling stages. The results showed that the most suitable growth stage to predict soybean grain yield was during the flowering stage as shown by both the NDVI ($R^2=0.863$) and the SR ($R^2=0.865$). Finally, the study assessed the potential of the new generation multispectral sensor Sentinel-2 MSI compared to Landsat 8 OLI and WorldView-2 in predicting soybean grain yield by resampling the hyperspectral data. The sensitivity testing of the multispectral bands revealed that sensitive spectral bands to soybean grain yield for Sentinel-2 MSI were the blue, red and re-edge bands whereas for Landsat 8 OLI and WorldView-2 included the red, blue and coastal blue bands. Sentinel-2 MSI yielded better results when predicting soybean grain yield than Landsat

8 OLI and WorldView-2. The study demonstrated a huge potential of hyperspectral remote sensing data in predicting soybean grain yields. In addition, the results showed the potential of new generation multispectral sensors to provide useful data in resource-poor countries. The findings of this study also demonstrated the utility of using remote sensing data during the flowering stage to predict soybean grain yield to assist in decision-making and overcome challenges confronting the use of conventional methods.

Declaration

This research work was undertaken in the Geography department at the University of the Free State, Qwaqwa campus, from April 2016 to January 2019, under the Supervision of Dr Samuel A. Adelabu (University of the Free State), and Dr George J. Chirima (Agricultural Research Council - Soil Water and Climate), in partial fulfilment of the requirements for the degree of Master of Science.

I declare that the research work reported in this dissertation has never been submitted in any way to any other university. It represents my original work, except where acknowledgements are made.

……..

Siphokazi R. Gcayi

…….

Dr S.A. Adelabu

……

Dr J. G. Chirima

Dedication

I dedicate this work to my parents and my siblings who fully supported this work.

Acknowledgements

I thank God in the Heavens who made this possible and provided me with the strength to complete my Masters dissertation.

Special thanks go to the Agricultural Research Council and the National Research Foundation for having funded this research.

To my supervisors, Dr Adelabu and Dr George Chirima, thank you for believing in me, mentoring, guiding and supporting me throughout the course of this study. I am grateful to Dr Khaled Abutaleb for his contribution and guidance in statistical analysis. Dr Solomon and Mr Eric Economon, thank you for your assistance with acquiring spectral signatures. To the ARC-ISCW Pedology division, thank you for allowing me to work in your soybean experimental farms, especially Patience Chauke, Sinawo Tsipinana and Bukanani Manina whom I worked with at the soybean experimental farms.

To my colleagues from the Agricultural Research Council, thank you for your encouragement and support during difficult times. The small talks we shared made things more endurable. My colleagues from the University of the Free State Geography department, thank you for your questions and criticism during our seminars, they helped me understand my research better.

I am grateful to my parents Mr Vulindlela Elliot Gcayi and Mrs Nobelungu Margaret Gcayi. You supported me and allowed me to grow and explore my capabilities. To my siblings, Luyanda, Gcobani, Siziwe and Siyabonga, thank you for your support. I especially acknowledge my brothers Gcobani and Siyabonga and my sister Siziwe who played an important role in my education. As my brother Gcobani always says, “Family takes care of each other”. To my sister Siziwe, thank you for your friendship and sisterhood. I always know that you have my back. To my brother Siyabonga, thank you for your support. Your advices make me challenge myself and think of the next step.

To everyone that encouraged me in one way or another and might have not mentioned by name, many thanks to you.

“God is not human, that He should lie, not a human being, that He should change His mind. Does He speak and then not act? Does He promise and not fulfil?

Numbers 23: 19

Table of Contents

List of Figures

List of Tables

Chapter 1

General Introduction

1.1 Background

Soybean (*Glycine max (L.) Merr*) is one of the important crops in the world grown on approximately 6 % of the fertile lands (Hartman *et al.*, 2011). In Africa, the largest producers of soybean include Nigeria, South Africa, and Uganda, producing about 39%, 35% and 14% of the continent's total yield respectively (Kolapo, 2011). Soybean in Africa enables tackling hunger and malnutrition problems as it is an economical source of protein (Kolapo, 2011). Also, soybean is important for its role in fixing nitrogen into the soil (Mabulwana, 2013a). Because of this important function , farmers in South Africa rotate soybean with maize in order to obtain better yields (Bahta and Willemse, 2016). In South Africa, soybean is grown in all the nine provinces that includes Gauteng, Mpumalanga, Limpopo, North West, Free State, KwaZulu Natal, Northern Cape, Eastern Cape and the Western Cape. However, Mpumalanga, KwaZulu Natal, and Free State provinces are the main growers (Magagane, 2012, Mabulwana, 2013b). South Africa produces about 100 000 to 800 000 tons of soybean per year averaging 1.7 to 2 tons per hectare (DAFF, 2014). About 8% of the produce is for human consumption; while about 32% is for oil and oilcake, and 60% is for animal feeds in particular the broiler and egg industries (DAFF, 2014). In the past years, soybean production and demand in South Africa have been increasing (Sihlobo and Kapuya, 2016). Owing to the increasing demand, soybean production in South Africa is foreseen to increase by 2020 (Dlamini *et al.*, 2014). The current crop produce does not meet the demand of South Africans (Sihlobo and Kapuya, 2016). As such, South Africa imports large quantities of soybean from Argentina (Dlamini *et al.*, 2014) and countries in Africa (DAFF, 2014). The increasing consumption of soybean products by South Africans suggests a need to produce higher soybean grain yields to meet the population demands.

Attaining higher yields for soybean requires expansion of the area planted or using more fertilizers (Bahta and Willemse, 2016). Expansion of area planted needs continuous monitoring using reliable methods that can produce instantaneous information from which yield predictions can be derived. Yield information can assist policy-makers and farmers in deciding on the handling of yields before and after harvest (Noureldin *et al.*, 2013). Also, yield predictions guide the market value of agricultural goods, the amount of imports and exports in case of shortfall and surplus,

transportation and trade between countries (Esfandiary *et al.*, 2009, Monteiro *et al.*, 2012, Rajah *et al.*, 2017). Presently, crop yield predictions rely on physical field visits (Noureldin *et al.*, 2013), agricultural census, manual field surveys (Fermont and Benson, 2011), and physical calculation of yield from numerous sampling areas (Wang *et al.*, 2013). Specifically, in South Africa, yield information is acquired through surveys conducted via telephone, post office mail and email (FAO, 2016). However, these methods are often biased, costly to carry out, prone to large inaccuracies (Noureldin *et al.*, 2013) and susceptible to human error. Information acquired through these methods may be available very late to avoid food shortfalls in case of poor yields (Noureldin *et al.*, 2013). This indicates that there is a need for developing reliable methods such as remote sensing based approaches that can be used to constantly monitor crop status, growth and development.

Remote sensing techniques provide cost effective and alternative methods, which can reduce the exhaustive manual field sampling (Adam, 2010). Compared to manual field survey methods, remote sensing offers objectivity, efficiency, and acquisition of statistical data that enhances crop yield estimation at regional, national, and global scales (Zhao *et al.*, 2007b, Ahmad *et al.*, 2014). Prediction of crop yields utilising remote sensing is based on the theory of reflectance of green vegetation acquired as spectral data from satellite imagery depending on conditions under which the images were captured, and the composition and the make-up of the crop (Sapkota *et al.*, 2016). The spectral reflectance is an illustration of vital factors that influence agricultural crops and other cumulative environmental factors that affect crop growth and development (Sawasawa, 2003). Measured spectral reflectance can be utilised to derive numerous vegetation indices such as the Normalised Difference Vegetation Index (NDVI) (Sapkota *et al.*, 2016) that can be effectively utilised to monitor crops and estimate crop yield (Mashaba *et al.*, 2017).

Commonly, researchers compute vegetation indices from broadband multispectral datasets in order to predict crop yields (Mutanga *et al.*, 2013, Noureldin *et al.*, 2013, Mashaba *et al.*, 2017). Multispectral data is characterised by broadbands, which are found in the visible and the near infrared regions of the electromagnetic spectrum (Govender *et al.*, 2007). In addition, multispectral data have the advantage of acquiring data at large spatial areas (Sibanda *et al.*, 2015). Datasets like Landsat 8 OLI and Sentinel-2 MSI provide additional advantages (European Space Agency, 2015, USGS, 2016) to developing countries such as South Africa because they can be readily and freely acquired from open source platforms. However, using multispectral data to derive vegetation

indices to predict crop yield is a challenge because they provide inadequate information on agricultural characteristics such as yield (Thenkabail *et al.*, 2002). Besides vegetation indices derived from multispectral data require optimised processing because they saturate when dealing with high biomass of crops (Mutanga and Skidmore, 2004, Adam *et al.*, 2014) such as soybean. Besides , broadband multispectral data have disadvantages when observing vegetation that have high spectral differences and shadows resulting from canopy and background (Adam *et al.*, 2014). These disadvantages often make it difficult to produce a precise biomass prediction model (Adam *et al.*, 2014). Although these limitations pose formidable challenges, they can still be overcome by using high-resolution hyperspectral data to predict crop yield with high biomass such as soybean.

Remote sensing hyperspectral data have the ability to subdue the obstacles encountered by broadband multispectral data because they have higher spectral resolution (Mashimbye, 2013). Hyperspectral data contains numerous and contiguous spectral bands in the visible, near infrared (NIR), middle infrared (MIR) and thermal infrared bands of the electromagnetic spectrum (Govender *et al.*, 2007, Adjorlolo, 2013). Data acquired through hyperspectral sensors enhance detailed study of the earth's attributes at levels of details and accuracy that are normally not attainable from broadband multispectral sensors (Govender *et al.*, 2007). In addition, information acquired by using instruments such as spectroradiometers produce high quality data of the vegetation condition and biomass compared to multispectral instruments (Kumar *et al.*, 2002). For vegetation, hyperspectral data facilitate determination of the health status, biophysical and biochemical characteristics of vegetation related to its composition and phenology (Thenkabail *et al.*, 2000, Adelabu, 2013). In application, hyperspectral data have been used to successfully predict biomass for vegetation (Mutanga and Skidmore, 2004, Adam *et al.*, 2014) and crops such as sugarcane, Swiss chard, wheat and cotton (Thenkabail *et al.*, 2000, Prasad *et al.*, 2007, Abdel-Rahman *et al.*, 2013, Abdel-Rahman *et al.*, 2014). However, to the best of my knowledge, hyperspectral data has not been extensively used to estimate soybean grain yields. This is because hyperspectral remote sensing in countries such as South Africa is new and is still being tested (Govender *et al.*, 2007).

When predicting soybean grain yield it is also imperative to know the optimal growth stage that is most appropriate to predict the yields. Tagarakis and Ketterings (2017), in their study predicted corn yield, they noted that the time in which remote sensing data is obtained could influence the

yield predictions of that crop. This is especially important for soybean, because soybean undergoes growth and development in two stages; the vegetative and reproductive stages (McWilliams *et al.*, 1999). The different growth phases respond differently to environmental factors that can influence yield predictions. For example, solar radiation is regarded as an important aspect that determines the yield of soybean as it is a driving factor of photosynthetic activity (Mathew *et al.*, 2000). The amount of radiation captured and used by soybean throughout the growing period differs depending on the growth phase and weather as determined by daily variations of energy received by the plant. Researchers have indicated that hyperspectral data could be used to determine the optimal growth stage to predict yield (Gao *et al.*, 2012, Gutierrez *et al.*, 2012). Ma *et al.* (2001), used multispectral canopy reflectance to determine the suitable stage to predict soybean grain yield. Also, Christenson *et al.* (2016) used hyperspectral data to predict soybean maturity and grain yield. However, a search of the literature revealed that hyperspectral data has not been utilised extensively to determine the suitable crop growth stage to predict soybean grain yield under different environmental conditions including in Africa. Although hyperspectral data have been used extensively to predict maize yields, their use in predicting soybean yield has been very limited. The reason for this could be that soybean is not as important as maize is in African countries because it is not a staple food. However, the increasing interest in soybean-based foods justifies concerted attempts to objectively predict soybean yields.

Although the use of hyperspectral data has traditionally been constrained by restricted access because of the high processing costs involved (Mutanga *et al.*, 2012). Nevertheless, advancements in remote sensing have produced new resolution multispectral sensors such as Sentinel-2 MSI, Landsat 8 OLI and WorldView-2 that have made access increasingly affordable. These new generation sensors have a combination of multispectral and hyperspectral attributes. Sentinel-2 MSI and WorldView-2 have a finer spectral resolution compared to traditional multispectral sensors (Zheng *et al.*, 2018). Sentinel-2 MSI and Landsat 8 OLI freely provide NDVI, leaf area index (LAI) as well as biophysical vegetation condition indicators (European Space Agency, 2015). These instruments are an advantage to resource-limited countries.

Hyperspectral data poses challenges when processing due to the high volumes of data involved (Dye *et al.*, 2011, Adjorlolo, 2013). As such, obtaining relevant information for predicting yield is difficult. Due to this limitation, researchers have suggested the use of advanced statistical methods such as the random forest algorithm. The random forest algorithm is a machine learning

technique that is able to overcome high dimensionality and redundancy problems of hyperspectral data (Dye *et al.*, 2011, Abdel-Rahman *et al.*, 2013). Researchers have observed that random forest performs better in comparison to other machine learning systems such as support vector machine and neural network because of its robustness against overfitting (Liaw and Wiener, 2002, Dye *et al.*, 2011, Abdel-Rahman *et al.*, 2013, Adelabu, 2013, Adam *et al.*, 2014). In addition, random forest provides more accurate regression results compared to support vector machine and artificial neural network and other commonly used algorithms (Wang *et al.*, 2016). This study used the random forest regression method to predict soybean grain yield from hyperspectral data. The study used field hyperspectral data to determine the optimal period to predict soybean grain yield during growth and development. The study resampled hyperspectral data to Sentinel-2 MSI, Landsat 8 OLI and WorldView-2 multispectral resolutions to test their ability in predicting soybean grain yield.

1.2 Aim

The aim of this study was to examine the utility of remote sensed hyperspectral data in predicting soybean grain yield.

1.3 The objectives of the study were to:

1. Evaluate the potential of narrow-band indices to predict soybean (*Glycine max (L.) Merr*) grain yield in the Free State and Mpumalanga provinces of South Africa.
2. Determine the suitable growth stage to predict soybean (*Glycine max (L.) Merr*) grain yield using hyperspectral data
3. Assess the ability of Sentinel-2 Multispectral Instrument (MSI) to estimate soybean (*Glycine max (L.) Merr*.) grain yield from resampled hyperspectral data.

1.4 Problem Statement

Soybean is an important crop that plays a crucial role in human nutrition, animal nutrition, industries and soil fertility. Although soybean contributes substantially to different sectors of the country's economy, South Africa remains a net importer of soybean because it does not produce enough grain to meet the country's ever-increasing demands. This indicates the need for the country to be able to produce enough soybean to meet the demand. Because South Africa is not able to produce enough soybean grain to meet its requirements, there is need for a dependable and affordable method to predict yields in order to guide recurrent estimations of how much the country has to import. Unfortunately, abilities to do so remain constrained by

lack of affordable techniques that are capable of enhancing a timely provision of the required information. Methods commonly used to monitor crops include manual ground field surveys and ground-based data reports, which have proved to be unreliable. This is because information acquired through these methods are highly susceptible to inaccuracies, bias and expensive to carry out. Remote sensing technology is growing into a widely used method in the field of agriculture. The technology is reliable, convenient, and cheaper. Broadband multispectral data has been widely used in crop monitoring and predictions of yields. However, it has limitations due to low spatial and spectral resolutions. It is therefore important to aim to provide work around techniques that be used to overcome these limitations in order to enhance the utility of remotely sensed data in aiding the prediction of soybean yields. For this study, this was done through assessing the ability of narrow-band indices in predicting soybean yield, determining the suitable stage to predict soybean and by resampling hyperspectral data to multispectral resolutions.

1.5 Dissertation Outline

This dissertation consists of five chapters. Chapter 1 gives a background of the study, followed by three chapters packaged as standalone manuscripts each addressing a specific objective, while the last chapter is a synthesis of the research. Each of the standalone manuscripts has an individual introduction, materials and methods, results, discussion and conclusion sections. Although the methods used were similar, these are repeated in every chapter to enhance clarity by providing space for detailed elaboration of how different datasets were processed and analysed. Because some contents including study areas are similar, the thesis has relied on cross-referencing several sections. These manuscripts will be submitted to journals for publication.

Chapter 2

Evaluates the potential of narrow-band indices to predict soybean (*Glycine max (L.) Merr*) grain yield in the Free State and Mpumalanga provinces of South Africa by identifying significant narrow-bands, narrow-band indices, and comparing NDVI, SR and EVI using random forest regression models. The results showed that relevant wavelengths in predicting soybean were combinations situated in the red-edge (680-750 nm), NIR and the MIR (1300 to

2399 nm) of the electromagnetic spectrum. Findings in this chapter show that the SR better predicts soybean grain yield compared to NDVI and EVI.

Chapter 3

Determines the most suitable growth stage to predict soybean (*Glycine max (L.) Merr*) grain yield using hyperspectral data by comparing the performance of NDVI and SR indices derived from hyperspectral data acquired during the flowering, pod formation and seed filling stages of soybean. Results indicated that the flowering stage is the best stage to predict soybean using hyperspectral data.

Chapter 4

Assesses the ability of Sentinel-2 Multispectral data to estimate soybean (*Glycine max (L.) Merr.*) grain yield from resampled hyperspectral data through systematic comparison with Landsat 8 OLI and WorldView-2 multispectral data. Results of this comparison revealed that Sentinel-2 MSI is better able to predict soybean yield than Landsat 8 OLI and WorldView-2.

Chapter 5

This chapter provides an overview of the findings of the research and summative conclusion and recommendations for future research placing the findings of this research in a broader context.

References

Abdel-Rahman, E. M., Ahmed, F. B. & Ismail, R. 2013. Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data. *International Journal of Remote Sensing,* 34**,** 712-728.

Abdel-Rahman, E. M., Mutanga, O., Odindi, J., Adam, E., Odindo, A. & Ismail, R. 2014. A comparison of partial least squares (PLS) and sparse PLS regressions for predicting yield of Swiss chard grown under different irrigation water sources using hyperspectral data. *Computers and Electronics in Agriculture,* 106**,** 11-19.

Adam, E. 2010. *The remote sensing of Papyrus vegetation (Cyperus papyrus L.) in swamp wetlands of South Africa.* Doctor of Philosophy in Environmental Sciences, University of KwaZulu-Natal.

Adam, E., Mutanga, O., Abdel-Rahman, E. M. & Ismail, R. 2014. Estimating standing biomass in papyrus (Cyperus papyrus L.) swamp: exploratory of in situ hyperspectral indices and random forest regression. *International Journal of Remote Sensing,* 35**,** 693-714.

Adelabu, S. 2013. *The Remote Sensing of Insect Defoliation in Mopane Woodland.* Phd, University of KwaZulu-Natal.

Adjorlolo, C. 2013. *REMOTE SENSING OF THE DISTRIBUTION AND QUALITY OF SUBTROPICAL C3 AND C4 GRASSES.* University of KwaZulu-Natal, Pietermaritzburg, South Africa.

Ahmad, I., Ghafoor, A., Bhatti, M. I. & Akhtar, I.-U. H. 2014. Satellite Remote Sensing and GIS based Crops Forecasting & Estimation System in Pakistan. *Crop monitoring for improved food security*.

Bahta, Y. T. & Willemse, J. 2016. The comparative advantage of South Africa soybean production. *OCL,* 23**,** A301.

Christenson, B. S., Schapaugh, W. T., An, N., Price, K. P., Prasad, V. & Fritz, A. K. 2016. Predicting Soybean Relative Maturity and Seed Yield Using Canopy Reflectance. *Crop Science,* 56**,** 625-643.

Daff 2014. Soybean Market Value Chain Profile. Pretoria: Department of Agriculture, Forestry and Fisheries.

Dlamini, T. S., Tshabalala, P. & Mutengwa, T. 2014. Soybeans production in South Africa. *OCL,* 21**,** D207.

Dye, M., Mutanga, O. & Ismail, R. 2011. Examining the utility of random forest and AISA Eagle hyperspectral image data to predict Pinus patula age in KwaZulu-Natal, South Africa. *Geocarto International,* 26**,** 275-289.

Esfandiary, F., Aghaie, G. & Mehr, A. D. 2009. Wheat yield prediction through agro meteorological indices for Ardebil District. *World Academy of Science, Engineering and Technology,* 49**,** 32-35.

European Space Agency, E. 2015. SENTINEL-2 User HandBook.

Fao 2016. Crop Yield Forecasting: Methodological and Institutional Aspects. Food and Agriculture Organiszation of the United Nations Rome.

Fermont, A. & Benson, T. 2011. Estimating yield of food crops grown by smallholder farmers. *International Food Policy Research Institute, Washington DC***,** 1-68.

Gao, J.-X., Chen, Y.-M., Lü, S.-H., Feng, C.-Y., Chang, X.-L., Ye, S.-X. & Liu, J.-D. 2012. A ground spectral model for estimating biomass at the peak of the growing season in Hulunbeier grassland, Inner Mongolia, China. *International journal of remote sensing,* 33**,** 4029-4043.

Govender, M., Chetty, K. & Bulcock, H. 2007. A review of hyperspectral remote sensing and its application in vegetation and water resource studies. *Water Sa,* 33**,** 145-151.

Gutierrez, M., Norton, R., Thorp, K. R. & Wang, G. 2012. Association of spectral reflectance indices with plant growth and lint yield in upland cotton. *Crop science,* 52**,** 849-857.

Hartman, G. L., West, E. D. & Herman, T. K. 2011. Crops that feed the World 2. Soybean—worldwide production, use, and constraints caused by pathogens and pests. *Food Security,* 3**,** 5-17.

Kolapo, A. 2011. *Soybean: Africa's Potential Cinderella Food Crop,* INTECH Open Access Publisher.

Kumar, L., Schmidt, K., Dury, S. & Skidmore, A. 2002. Imaging spectrometry and vegetation science. *Imaging spectrometry.* Springer.

Liaw, A. & Wiener, M. 2002. Classification and regression by randomForest. *R news,* 2**,** 18-22.

Ma, B., Dwyer, L. M., Costa, C., Cober, E. R. & Morrison, M. J. 2001. Early prediction of soybean yield from canopy reflectance measurements. *Agronomy Journal,* 93**,** 1227-1234.

Mabulwana, P. T. 2013a. *Determination of Drought Stress Tolerance Among Soybean Varieties Using Morphological and Physiological Markers.* Master of Science, University of Limpopo.

Mabulwana, P. T. 2013b. *Determination of drought stress tolerance among soybean varieties using morphological and physiological markers.* University of Limpopo.

Magagane, T. G. 2012. *Genotype by environment interactions in soybean for agronomic traits and nodule formation.* FACULTY OF SCIENCE AND AGRICULTURE, UNIVERSITY OF LIMPOPO, SOUTH AFRICA.

Mashaba, Z., Chirima, G., Botai, J. O., Combrinck, L., Munghemezulu, C. & Dube, E. 2017. Forecasting winter wheat yields using MODIS NDVI data for the Central Free State region. *South African Journal of Science,* 113**,** 1-6.

Mashimbye, Z. E. 2013. *Remote sensing of salt-affected soils.* Stellenbosch: Stellenbosch University.

Mathew, J. P., Herbert, S. J., Zhang, S., Rautenkranz, A. A. & Litchfield, G. V. 2000. Differential response of soybean yield components to the timing of light enrichment. *Agronomy Journal,* 92**,** 1156-1161.

Mcwilliams, D., Berglund, D. & Endres, G. 1999. Soybean growth and management quick guide. *North Dakota State University and University of Minnesota*.

Monteiro, P. F. C., Angulo Filho, R., Xavier, A. C. & Monteiro, R. O. C. 2012. Assessing biophysical variable parameters of bean crop with hyperspectral measurements. *Scientia Agricola,* 69**,** 87-94.

Mutanga, O., Adam, E. & Cho, M. A. 2012. High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *International Journal of Applied Earth Observation and Geoinformation,* 18**,** 399-406.

Mutanga, O. & Skidmore, A. K. 2004. Narrow band vegetation indices overcome the saturation problem in biomass estimation. *International Journal of Remote Sensing,* 25**,** 3999-4014.

Mutanga, S., Van Schoor, C., Olorunju, P. L., Gonah, T. & Ramoelo, A. 2013. Determining the best optimum time for predicting sugarcane yield using hyper-temporal satellite imagery. *Advances in Remote Sensing,* 2**,** 269.

Noureldin, N., Aboelghar, M., Saudy, H. & Ali, A. 2013. Rice yield forecasting models using satellite imagery in Egypt. *The Egyptian Journal of Remote Sensing and Space Science,* 16**,** 125-131.

Prasad, B., Carver, B. F., Stone, M. L., Babar, M., Raun, W. R. & Klatt, A. R. 2007. Potential use of spectral reflectance indices as a selection tool for grain yield in winter wheat under Great Plains conditions. *Crop science,* 47**,** 1426-1440.

Rajah, P., Odindi, J., Abdel-Rahman, E. & Mutanga, O. 2017. Determining the optimal phenological stage for predicting common dry bean (Phaseolus vulgaris) yield using field spectroscopy. *South African Journal of Plant and Soil,* 34**,** 379-388.

Sapkota, T. B., Jat, M., Jat, R., Kapoor, P. & Stirling, C. 2016. Yield Estimation of Food and Non-food Crops in Smallholder Production Systems. *Methods for Measuring Greenhouse Gas Balances and Evaluating Mitigation Options in Smallholder Agriculture.* Springer.

Sawasawa, H. L. 2003. Crop Yield Estimation: Integrating RS, GIS, and Management Factor. *A case study of Birkoor and Kortigiri Mandals, Nizamabad District India***,** 1-9.

Sibanda, M., Mutanga, O. & Rouget, M. 2015. Examining the potential of Sentinel-2 MSI spectral resolution in quantifying above ground biomass across different fertilizer treatments. *ISPRS Journal of Photogrammetry and Remote Sensing,* 110**,** 55-65.

Sihlobo, W. & Kapuya, T. 2016. South Africa's soybean industry: A brief overview. [Accessed 15/02/2017].

Tagarakis, A. C. & Ketterings, Q. M. 2017. In-season estimation of corn yield potential using proximal sensing. *Agronomy Journal,* 109**,** 1323-1330.

Thenkabail, P. S., Smith, R. B. & De Pauw, E. 2000. Hyperspectral vegetation indices and their relationships with agricultural crop characteristics. *Remote sensing of Environment,* 71**,** 158-182.

Thenkabail, P. S., Smith, R. B. & De Pauw, E. 2002. Evaluation of narrowband and broadband vegetation indices for determining optimal hyperspectral wavebands for agricultural crop characterization. *Photogrammetric Engineering and Remote Sensing,* 68**,** 607-622.

Usgs 2016. Landsat 8 (L8) Data Users Hnadbook.

Wang, L. A., Zhou, X., Zhu, X., Dong, Z. & Guo, W. 2016. Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *The Crop Journal,* 4**,** 212-219.

Wang, Q., Nuske, S., Bergerman, M. & Singh, S. Automated crop yield estimation for apple orchards. Experimental Robotics, 2013. Springer, 745-758.

Zhao, J., Shi, K. & Wei, F. Research and application of remote sensing techniques in Chinese agricultural statistics. Fourth international conference on agricultural statistics, October, 2007. 22-24.

Zheng, Q., Huang, W., Cui, X., Shi, Y. & Liu, L. 2018. New Spectral Index for Detecting Wheat Yellow Rust Using Sentinel-2 Multispectral Imagery. *Sensors,* 18**,** 868.

Chapter 2

Evaluating the potential of narrow-band indices to predict soybean (*Glycine max (L.) Merr*) grain yield in the Free State and Mpumalanga provinces of South Africa

Abstract

Yield information makes it possible for decisions to be taken regarding the management of agricultural production before and after harvest by government and other decision makers. Traditional approaches to collecting yields such as manual surveys and physical computation of yield are costly and take a long time for information to be available. Remote sensing hyperspectral data can be used to provide real-time, fast, and reliable yield information that can be useful in predicting soybean grain yield. Vegetation indices are ratios used to combine multiple band observations of the hyperspectral data into one index and can be applied to derive soybean grain yield. The objective of this study was to evaluate the potential of vegetation indices derived from hyperspectral data to predict soybean grain yield. Soybean hyperspectral data were acquired in March and April summer season of 2017 for Free Sate and Mpumalanga provinces. Hyperspectral data was acquired from the above-mentioned sites using a handheld spectroradiometer with a spectral range of 350 to 2500 nm sites from 72 plots of each site. The random forest regression algorithm was used to predict the soybean grain yield. NDVI, SR and EVI were calculated from the hyperspectral data for all probable bands situated in the 400 nm and 2399 regions of the electromagnetic spectrum. The results showed that relevant wavelengths in predicting soybean were combinations situated in the red-edge (680-750 nm), NIR and largely in the MIR region (1300 to 2399 nm) of the electromagnetic spectrum. Furthermore, regression results showed that SR better predicted the soybean grain yield (R^2 = 0.843) compared to NDVI and EVI that yielded R2 = 0.841 and R^2= 0.537 respectively. Overall, the results of this study suggest that narrow-band indices have the potential to predict soybean grain yield.

Keywords: soybean yield, hyperspectral data, vegetation indices

2.1 Introduction

South Africa is the third largest consumer of soybean in the world (Van De Merwe *et al.*, 2013). Mpumalanga, KwaZulu Natal, and Free State provinces are the largest soybean producers in the country (Magagane, 2012). Over the last decade, soybean production and consumption in South Africa has increased (Van De Merwe *et al.*, 2013, Sihlobo and Kapuya, 2016). Currently, soybean production does not meet South African local demands (Sihlobo and Kapuya, 2016). As a result, South Africa imports large quantities of soybean products (Sihlobo and Kapuya, 2016). Attaining higher yields entails increasing the area planted and/or use of more fertilisers (Mourtzinis *et al.*, 2013). Increasing production in both approaches requires constant crop monitoring using reliable techniques that can provide real-time statistics. Constant monitoring of crops can enhance chances of attaining higher yield through early detection of problems that can potentially affect yield. Soybean yield information in the hands of farmers and policy-makers is important for decisions such as planning for harvesting, yield management and market profiling (Noureldin *et al.*, 2013). Thus, there is a need for an efficient real-time monitoring system to consistently provide information on the status, growth and development of soybean in order to facilitate yield predictions.

Various methods that include the use of agricultural censuses and field surveys have been used to predict grain crop yields (Fermont and Benson, 2011) a (Wang *et al.*, 2013). In South Africa, current yield predictions are based upon field surveys conducted telephonically, via emails, and or by post (FAO, 2016). However prediction methods based on traditional crop yields surveys are frequently subjective, susceptible to large inaccuracies and take a long time for information to be available for the benefit of food security and early planning before and during harvests (Noureldin *et al.*, 2013). In addition, predicted yields influence the pricing of agricultural commodities and the decisions to be taken regarding imports and exports (FAO, 2016). This therefore justifies the need for crop monitoring initiatives that involve the use of reliable techniques such as remote sensing to ensure fair pricing of agricultural commodities and objective decision-making. Remote sensing methods are suitable because they include the acquisition of crop canopy measurements (Ahmad *et al.*, 2014), and can deliver immediate, reliable, measurable evaluations of the ability of plants to capture radiation and photosynthesize (Ma *et al.*, 2001). These canopy spectral measurements are beneficial for estimating crop yield (Ahmad *et al.*, 2014). Research shows that remote sensing spectral bands have strong relationships with vegetation biomass (Adam *et al.*, 2014).

Many researchers have used broadband multispectral data to predict yield of various crops such as maize (Shanahan *et al.*, 2001), rice (Noureldin *et al.*, 2013), soybean (Ma *et al.*, 2001) and wheat (Wang *et al.*, 2014, Mashaba *et al.*, 2017). Broadband multispectral data have advantages as it is applicable to regional areas and also because of numerous revisits of the same area as well as capturing data at large spatial scales in real-time (Sibanda *et al.*, 2015). Despite these advantages, broadband data has drawbacks for vegetation observation such as exhibiting excessive spectral differences and shadows from the above-ground coverage and landscape (Adam *et al.*, 2014). The latter can be a hindrance in producing precise biomass prediction models with the ability to distinguish between soil background and vegetation (Adam *et al.*, 2014). Precise biomass predictions are essential for effective monitoring and management of vegetation (Adam *et al.*, 2014). Furthermore, broadband data does not have specific narrow-bands that precisely focus on biochemical and biophysical characteristics of crops (Thenkabail *et al.*, 2002, Mariotto *et al.*, 2013). This suggests that multispectral broadband datasets exhibit difficulties in monitoring crops with high biomass such as soybean. Although multispectral broadband datasets have these disadvantages, research has shown that these disadvantages can be overcome by the use of vegetation indices (Mutanga and Skidmore, 2004). Vegetation indices eliminate differences caused by soil background, above-ground geometry, sun view angles as well as the influence of atmospheric circumstances when assessing biophysical characteristics of vegetation at above-ground scale (Mutanga and Skidmore, 2004).

Widely used vegetation indices for vegetation monitoring and modelling are calculated using the red and the near infrared (NIR) bands (Cho *et al.*, 2007). The red and NIR bands respond to the biochemical and biophysical properties of crops (Thenkabail *et al.*, 2002, Cho *et al.*, 2007) and are sensitive to the rate of photosynthetic activity in green vegetation (Teillet *et al.*, 1997). The Normalised Difference Vegetation Index (NDVI) (Tucker, 1979) and Simple Ratio (SR) (Jordan, 1969) are commonly utilised indices that are calculated using the NIR and the red bands (Teillet *et al.*, 1997) with applications for crop monitoring. Soybean has been monitored using NDVI modelled from broadband data sets such as AVHRR/NOAA (Lokupitiya *et al.*, 2010, Esquerdo *et al.*, 2011). Locke *et al.* (2000) used SR, NDVI, Soil Adjusted Vegetation Index (SAVI) and Transformed SAVI (TSAVI) to evaluate soybean biophysical properties such as yield, leaf area index (LAI) and biomass (Locke *et al.*, 2000). Furthermore, the SR index is known to be able to decrease the effect of soil background on the spectral reflectance and is also sensitive to changes occurring at prime developmental phases of vegetation (Adelabu *et al.*, 2012). The Enhanced

Vegetation Index (EVI) is another widely used index in agricultural forecasting and is computed using the red and NIR bands with an addition of the blue band (Huete *et al.*, 1994). However, the EVI is insensitive to saturation when faced with high biomass vegetation (Testa *et al.*, 2018). Despite the usefulness of these spectral bands, broadband data is unresponsive to the variation in plant features (Sibanda *et al.*, 2015).

Due to disadvantages encountered by broadband data, researchers encourage the use of hyperspectral data that covers the whole range of the electromagnetic spectrum instead of just two or three bands (Mutanga and Skidmore, 2004). Hyperspectral data provide advantages of handiness, flexibility, controllability and high temporal resolution, which are greatly beneficial in precision agriculture applications as opposed to satellite based platforms (Huang *et al.*, 2016). Also, hyperspectral datasets contain other important spectral bands such as the red edge bands that are useful in the study of vegetation (Mutanga and Skidmore, 2004). The red edge band is highly responsive to variations in biomass of green vegetation (Mutanga and Skidmore, 2004). Narrow bands are important for supplying more information with substantial enhancements compared to broad bands in enumerating biophysical properties of agricultural crops (Thenkabail *et al.*, 2000, Mariotto *et al.*, 2013). In addition, hyperspectral datasets are important for modelling yield features of agricultural crops (Mariotto *et al.*, 2013) such as chlorophyll content, photosynthetic activities and leaf structure (Kumar *et al.*, 2002). Numerous researchers (Thenkabail *et al.*, 2000, Mutanga and Skidmore, 2004, Mariotto *et al.*, 2013) have used hyperspectral data for vegetation monitoring with positive results.

Mutanga and Skidmore (2004), calculated NDVI from hyperspectral data and obtained that regular NDVI including strong chlorophyll absorption bands in the red region and NIR region inadequately predicted biomass (R^2=0.26). Whereas, the modified NDVI (MNDVI) that included bands in the range (700-750 nm) and narrow-bands in the red-edge region (750-780 nm) showed a high predictive ability for biomass (R^2=0.77). Mariotto *et al.* (2013), observed that important bands when modelling biophysical properties of maize, wheat, cotton, rice and alfafa, (about 74% of them) are situated in the 1051-2331 nm regions. The remaining percentage of bands are in the 970 nm region (10%), red-edge region (6%) and the visible region (10%) (Blue region (400-500nm), green region (501-600 nm) and NIR region (760-900 nm). Thenkabail *et al.* (2000) reported that stronger correlations with crop biophysical characteristics were situated in the red region (650-700 nm), shorter wavelengths of the green region (500-550 nm), the NIR region (900-

940nm) and in the moisture sensitive area centred at 982 nm. Similarly, many researchers have used hyperspectral data to predict yield of agricultural crops such as lint (Zhao *et al.*, 2007a), wheat (Babar *et al.*, 2006), maize (Weber *et al.*, 2012) and soybean (Ma *et al.*, 2001). However, for soybean, Ma *et al.* (2001) utilised spectral data acquired using a multispectral hand-held radiometer with a fewer number of bands and obtained a positive correlation between NDVI and soybean grain yield ($R^2 = 0.80$). Overall, however, research has shown that hyperspectral data has enabled estimation of yield of various crops and biomass of several vegetation types but soybean grain yield has not been predicted comprehensively using hyperspectral data.

Hyperspectral data has however some limitations, such as those related to high dimensionality and redundancy (Abdel-Rahman *et al.*, 2014) and the problem of multicollinearity (Adjorlolo, 2013). As a result, identifying suitable bands for modelling is a challenging process. To overcome this problem, researchers encourage the use of advanced statistical methods such as random forest (RF) regression algorithm (Adam *et al.*, 2014). Random forest is a regression algorithm that applies bootstrapping aggregation to create a group of trees based on the randomness of samples taken from the training data (Adelabu, 2013). The random forest algorithm is known to be able to handle the high dimensionality of hyperspectral data and reduce data redundancy (Adjorlolo, 2013). Also, random forest has been noted to perform better than other machine learning algorithms such as support vector machine and neural network because of its robustness against overfitting (Liaw and Wiener, 2002, Dye *et al.*, 2011, Abdel-Rahman *et al.*, 2013, Adelabu, 2013, Adam *et al.*, 2014). The aim of this study was to evaluate the performance of narrow-band vegetation indices derived from hyperspectral data notably NDVI, SR and EVI in predicting soybean grain yield. The vegetation indices selected for the study are those frequently used for biomass or agricultural crop and ecological vegetation studies (Mutanga and Skidmore, 2004) and have been applied successfully in predicting other crops. The first objective of this study was to assess the relationships of narrow-band NDVI, SR and EVI to soybean grain yield. The second objective was to identify narrow-band indices suitable for predicting soybean grain yield. The third objective was to compare the performance of NDVI, SR and EVI random forest models developed from narrow-bands (400 nm to 2399 nm) in predicting soybean grain yield.

2.2 Materials and Methods

2.2.1 Study sites

The research was conducted on two experimental farms located in the Free State Province of South Africa in Phuthaditjhaba (28°25'26"S and 28°56'12"E) and in the Mpumalanga province in Ermelo (26° 45'18" S and 30° 13'55" E) (**Figure 2.1**). The Free State and Mpumalanga provinces experience warm summers with high rainfall and cold winters. Both areas receive approximately 625 mm of precipitation annually with most precipitation occurring in summer (October - March). The soil in Phuthaditjhaba can be characterised as "rich loam" (Koatla, 2012) while the soil in Ermelo can be characterised as "low clay" sandy soil (Sakala *et al.*, 2017) . The different sites were chosen to test if there would be differences in soybean yield since the areas consisted of different soil.

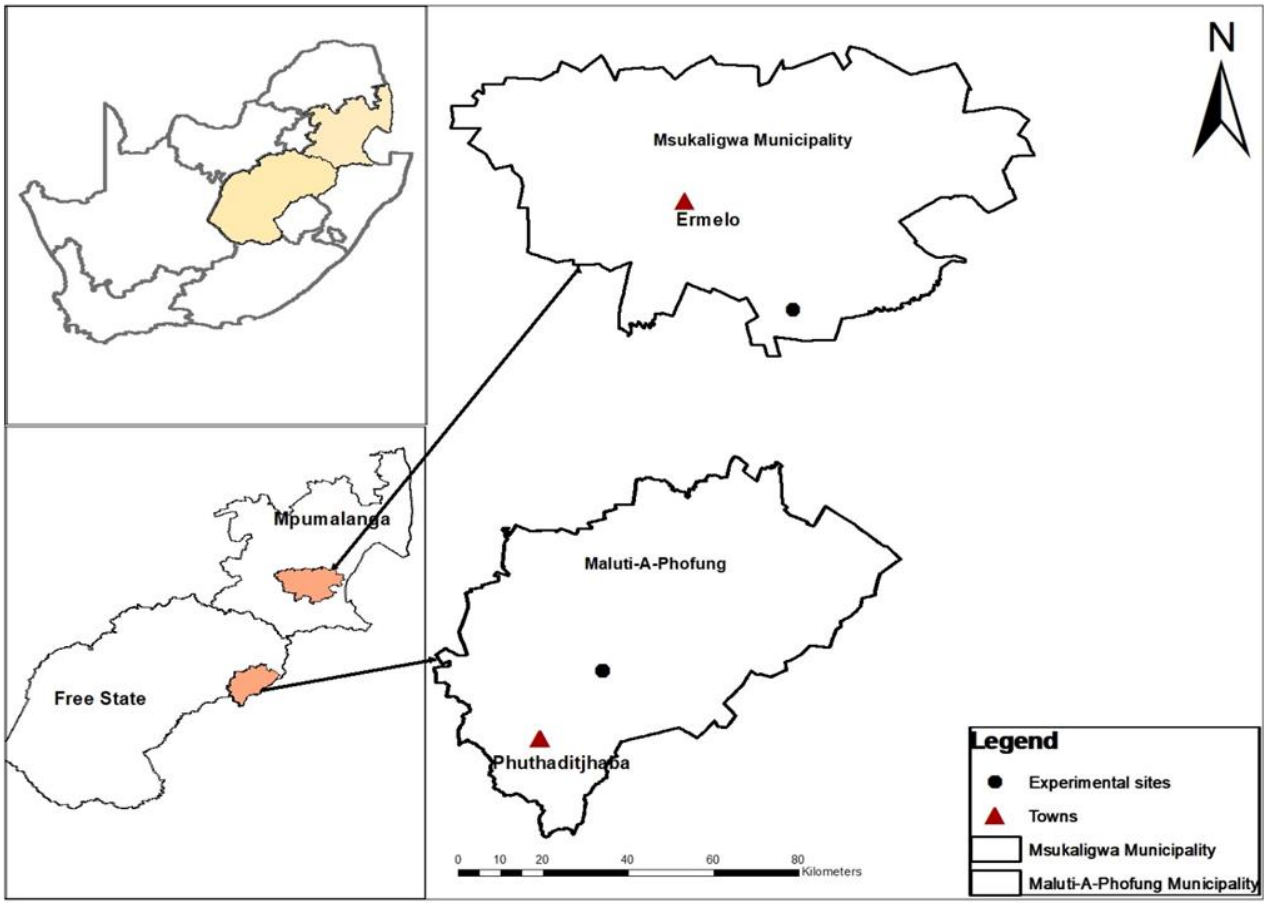


Figure 2.1: Map showing the location of the study sites in Free State (FS) and Mpumalanga (MP) provinces.

2.2.2 Experimental setup

The experiment on both sites followed a split plot Randomized Complete Block Design (RCBD) method. In the two study sites, 72 experimental plots each with a size of 7 m length and 3 m width were used. The plots consisted of 7 rows with 60 cm row spacing. Three soybean cultivars from

Pannar seeds (PANN 1500 R, PANN 1614 R and PANN 1664 R) were sown from 13th to 15th December 2016 in the MP and from 19th to 21st of December 2016 in FS site. Fertilizer treatments of 0 kg, 30 kg and 60 kg of phosphorus (P) were applied to the plots to test if there would be differences in yield based on the different treatments. The experiment consisted of three replicates and the soybean relied on rainwater.

2.2.3 Field spectral measurements

The first set of field spectral measurements in Mpumalanga and Free State were taken in March 2017 and the second set of spectral measurements were taken in April 2017. During these periods, the soybean had reached maximum canopy cover whereby the soil background could have little effect on the spectral measurements. Due to differences in planting date, the soybean in Mpumalanga was in the pod formation stage during the first visit while in the Free State site it was still flowering. Canopy spectral measurements were acquired randomly plot by plot across fertilizer treatments of 0 kg, 30 kg and 60 kg during the flowering, pod formation and seed filling stages. An Analytical Spectral Device (ASD) Field Spec®3 optical sensor (Analytical Spectral Devices, Inc., Boulder, CO, USA) was used to take spectral measurements from 10:00 am to 14:00 pm local time (GMT+2). The spectroradiometer records wavelength ranging from 350 to 2500 nm, measuring radiation at 1.4 nm bandwidths for the spectral region of 350-1000 nm and registers 2 nm intervals for the spectral region of 1001-2500 nm (ASD, 2005).

The spectral measurements were taken under cloud free conditions. In each plot, 5 spectral measurements were taken with the optical cable connected to the spectroradiometer held at about 30 cm above the soybean canopy. Every 10 to 15 minutes a white reference spectralon calibration panel was used to balance any changes in the atmosphere and irradiance of the sun. The spectral measurements were added together to obtain the median spectral measurements for each plot. **Figure 2.2** shows average spectral reflectances of soybean at flowering, pod formation and seed filling stages. The spectral reflectance curves indicate the amount of radiation absorbed and reflected by the soybean at different regions of the spectrum. For soybean, the flowering and pod formation stages are critical stages in which the soybean utilises the absorbed radiation to photosynthesise and form grains (Board and Kahlon, 2011). A higher spectral signature is an indicator of a healthy crop in which high yield can be expected whereas a low spectral signature indicates a low yield (Board and Kahlon, 2011).

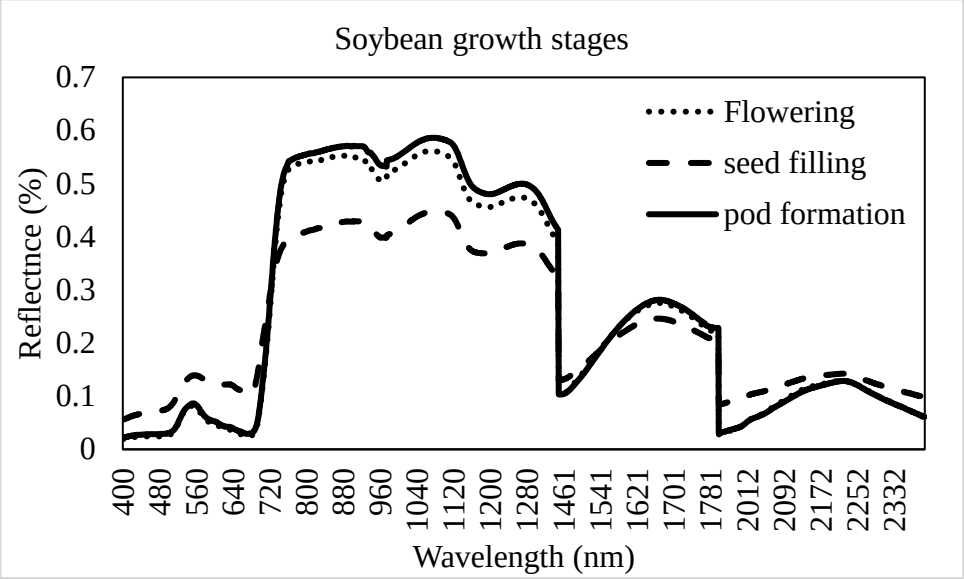


Figure 2.2: Average spectral curves of soybean canopies at flowering, pod formation and seed filling stages

2.2.4 Soybean yield data

To obtain soybean grain yield data, the soybean pods were harvested from the middle 3 rows of each plot at the end of the growing season of May and June 2017. The soybean pods were then crushed to obtain the soybean grains. The soybean grains obtained from each plot were weighed using the LBK1 weighing scale from ADAM Equipment (Adam Equipment, 2017). The grain measurements of specific plots for each site were added to obtain the total yield of the soybean of each site.

2.3 Data analysis

448 Bands ranging from 350 to 399 nm, 1350 to 1450 nm, 1800 to 1950 nm and 2400 to 2500 nm were omitted from the analysis due to atmospheric water absorption and the effect of noise in the reflectance spectra following techniques described by Abdel-Rahman *et al.* (2014) and Adam *et al.* (2014). The remaining 1702 narrow-bands situated between 400 nm and 2399 nm were used to compute the narrow-band indices.

The NDVI, SR and EVI indices were calculated using the standard indices equations (Jordan, 1969, Rouse, 1974, Huete *et al.*, 1994) (**Table 2.1**). These indices were calculated from all probable two-bands combinations including 1702 narrow bands situated between 400 and 2399 nm (Mutanga and Skidmore, 2004, Cho *et al.*, 2007, Adam *et al.*, 2014). The narrow bands are

presented as λ_1 (400-2399 nm) and λ_2 (400-2399 nm) combinations following approaches outlined in (Mutanga and Skidmore, 2004). The calculated vegetation indices were correlated to the soybean yield using the Spearman's correlation coefficient (Mukaka, 2012). The correlations between vegetation indices and soybean grain yield were calculated to assess their relationship.

Table 2.1: Vegetation indices computed from the λ_1 (400-2399 nm) and λ_2 (400-2399 nm) combinations.

Index name	Abbreviation	Formula	Reference
Normalized Difference Vegetation Index	NDVI	$NDVI = \frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2}$	(Rouse, 1974)
Simple Ratio	SR	$SR = \frac{\lambda_1}{\lambda_2}$	(Jordan, 1969)
Enhanced Vegetation Index	EVI	$EVI = \mathrm{G}\frac{\mathrm{N} - \mathrm{R}}{\mathrm{N} + \mathrm{C_1R} - \mathrm{C_2B} + \mathrm{L}}$	(Huete *et al.*, 1994)

2.3.1 Assessing the differences in yields between study sites and fertilizer treatments

Exploratory data analysis was performed to understand the data before any statistical analysis was done. The statistical analysis was performed in STATISTICA 13 software testing for normalcy of the data using Lilliefors test (Dell Inc, 2015). Furthermore, an analysis of variance was performed to determine if there were differences in soybean grain yield means between the two study sites and between the three fertilizer treatments.

2.3.2 Statistical analysis using the Random forest (RF) regression

The random forest regression technique was used to predict the soybean grain yield. RF is a machine learning algorithm developed by (Breiman, 2001) that applies a bootstrap aggregation method in which an ensemble of trees (*ntree*) are developed on the basis of the randomness of samples extracted from the training data. For regression, the random forest permits trees to grow to the highest magnitude without trimming, depending on the bootstrap sample from the training data (Breiman, 2001). At every tree, the RF grows a randomized subgroup of predictors (*mtry*) to

identify the optimum split at every node of the tree (Abdel-Rahman *et al.*, 2013). At the end, the RF averages the outcome of the overall sum of trees in order to obtain the overall estimation (Prasad *et al.*, 2006b). From the bootstrap samples of the training data (2/3), each tree grows randomly and selected independently. The residual original data (1/3) of the excluded samples (called out-of-bag (OOB)) are then used to validate the model and predict variables of importance (Palmer *et al.*, 2007, Powell *et al.*, 2010).

RF requires two parameters to be tuned, these parameters include (i) (*ntree*) the number of trees to grow and (ii) (*mtry*) the number of variables that are split at each node (Abdel-Rahman *et al.*, 2013). The *ntree* and the *mtry* parameters (vegetation indices) were then optimized for the random forest model using the top 20 NDVI, SR and EVI data sets to determine the best index that can be used to predict soybean grain yield. The *mtry* was calculated for all probable band combinations while the *ntree* was evaluated at 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, and 5000 trees. The random forest model was developed from 70% (2/3) of the training data to build a model that can predict soybean grain yield (g/m^2) and 30% (1/3) of the test data was used to validate the model (OOB). Important indices at predicting soybean grain yield were selected by the RF using the permutation variable importance measures (mean decrease in accuracy). The RF algorithm was implemented using the R statistical software's *randomForest* built in package to predict the soybean grain yield (Liaw and Wiener, 2002).

2.3.3 Variable Importance Selection

Random forest calculates variable importance using the Gini index and the permutation variable importance measures (Boulesteix *et al.*, 2012). The permutation variable importance measure is defined as the variation between the OOB error from the data set acquired by random selection of the predictor variables and the OOB error from the original data set (Boulesteix *et al.*, 2012). The Gini index is a measure of variable importance used in a classification when growing trees in the random forest (Smyth, 2004). The permutation variable importance measure is the most preferred measure of importance as it assesses importance of variables using the mean decrease in accuracy in the OOB predictions as forests are being assembled (Boulesteix *et al.*, 2012). Permutation variable importance predicts the importance of a variable by determining how much prediction error rises when a variable is selected while others remain the same (Kuhn *et al.*, 2008, Fathima and Sheriff, 2012). For this study, the permutation variable importance was used to determine the combination of indices that were powerful than the others in predicting soybean grain yield. From

the ranking of the mean decrease in accuracy, the top 3 important combinations of indices were selected.

2.3.4 Accuracy Assessment

When using the random forest, research has shown that there is no need for a different test data for validation because the random forest uses an OOB error prediction built internally (Prasad *et al.*, 2006b, Adam, 2010, Adelabu, 2013, Adjorlolo, 2013, Karlson *et al.*, 2015). This is particularly remarkable in situations where data acquisition is highly dependent on oscillating weather conditions. The random forest computes the OOB error as a result of variance between the estimation made using the training data set and the OOB data set (Abdel-Rahman *et al.*, 2013, Belgiu and Drăguţ, 2016). OOB error produces an unbiased evaluation of the prediction accuracy of the model (Dye *et al.*, 2011). The coefficient of determination (R^2) and root mean square error (RMSE) were reported on the assessment of the accuracy of the random forest models. RMSE was calculated using the formula below:

$$RMSE = \sqrt{\frac{\sum(\hat{\mathbf{Y}}\mathbf{i} - \mathbf{Y})^{\mathbf{2}}}{\mathbf{n}}}$$

where Ŷ and Y are measured and predicted soybean grain yield respectively.

2.4 Results

2.4.1 Assessing the differences in soybean yields between study sits and fertiliser treatments

Exploratory statistics showed that soybean grain yield data does not significantly deviate away from a normal distribution for both sites (**Figure 2.3**) and thus meets the assumptions of ANOVA. Analysis of variance results showed that there were significant differences between the soybean grain yield in Free State and Mpumalanga provinces ($p \leq 0.05$). However, the results showed no significant differences in soybean grain yield between fertilizer treatments on the study sites ($p \geq 0.05$). The total soybean grain yield obtained in FS was 72816 g/m^2 with an average of 1011.3 g/m^2 per field while the total soybean grain yield in MP was 156060 g/m^2 with an average of 2167.5 g/m^2 per field. In total, the soybean grain yield of both sites was 228876 g/m^2 with an average of 1589.4 g/m^2.

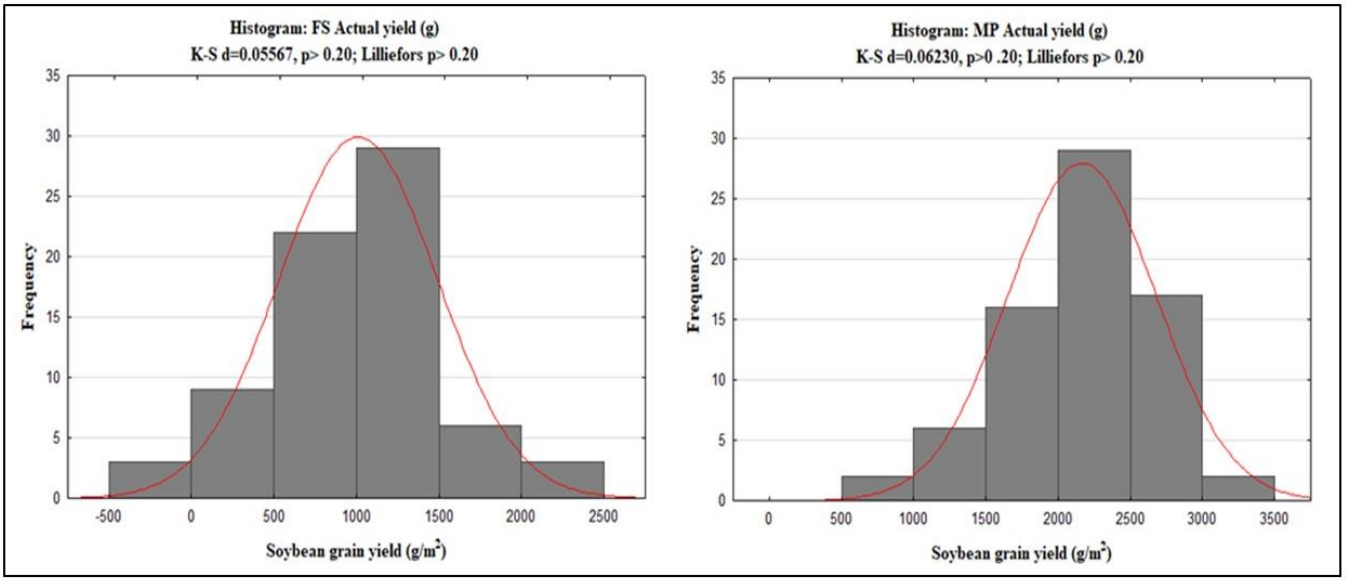


Figure 2.3: Descriptive statistics of soybean grain yields for FS (a) and MP (b) sites.

2.4.2 Narrow-band NDVI and SR relationship to soybean grain yield

Narrow-band NDVI and SR were computed for all probable two-band combinations in the spectral range 400 nm to 2399 nm. Spearman's correlation coefficients were applied to assess the relationships of the narrow-band NDVI and SR to soybean yields. The NDVI and SR obtained identical results of the correlations to the soybean grain yield (**Table 2.2** and **Table 2.3)**. The correlation coefficients (R) results obtained between NDVI/SR and soybean grain yield ranged from 0.00 to 0.68 shown in **Table 2.2** and **Table 2.3**.

Table 2.2: Top 20 narrow band NDVI indices (λ=30 nm) that produced the highest correlation coefficients with soybean grain yield.

Ranking	Wavelength (nm)	Wavelength (nm)	R-values	P-values
1	1806	2107	0.688	0.000
2	1806	2137	0.655	0.000
3	2377	2077	0.633	0.001
4	1806	2167	0.619	0.001
5	715	1536	0.618	0.001
6	1806	2317	0.617	0.001
7	1806	1476	0.616	0.001
8	2347	2107	0.613	0.002
9	1806	2287	0.605	0.002
10	475	2047	0.602	0.002
11	445	2077	0.602	0.002
12	715	1566	0.601	0.002

13	475	2077	0.601	0.002
14	715	1506	0.600	0.002
15	445	2107	0.598	0.002
16	475	2107	0.596	0.002
17	475	2017	0.595	0.002
18	445	2047	0.595	0.002
19	445	2017	0.588	0.006
20	715	1596	0.588	0.006

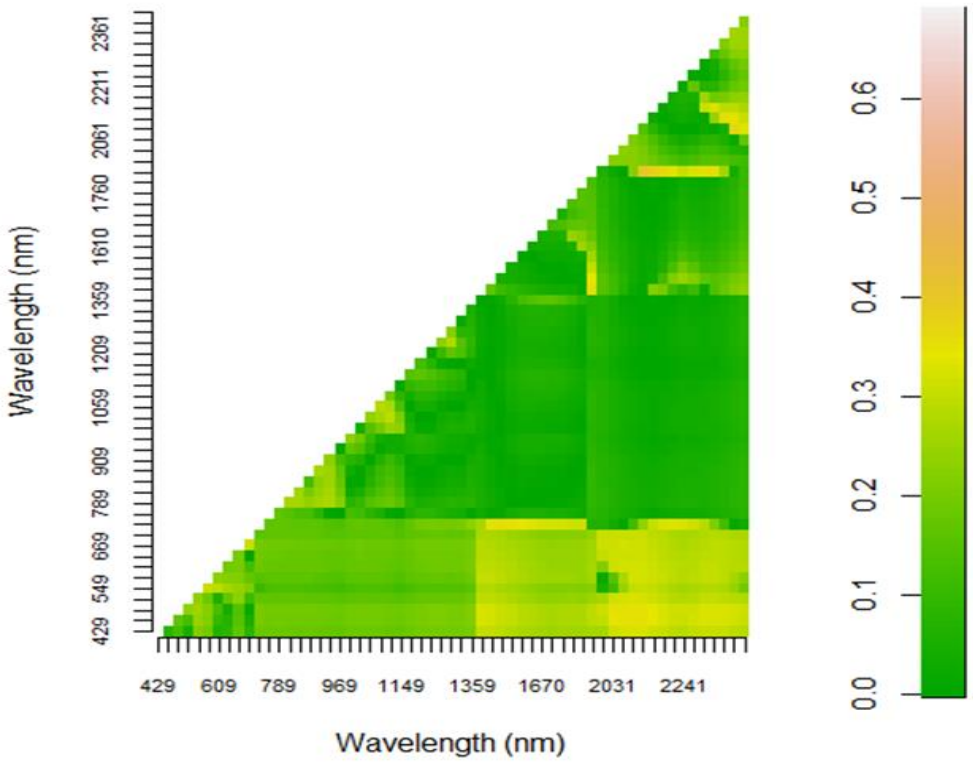


Figure 2.4: Heat map showing the correlation coefficients (R) between soybean grain yield and narrow band NDV acquired from all probable band combinations from the spectral range of 400 nm to 2399 nm.

Table 2.3: Top 20 narrow band SR indices (λ=30 nm) that produced the highest correlation coefficients with soybean grain yield.

Ranking	Wavelength (nm)	Wavelength (nm)	R-values	P-values
1	1806	2107	0.688	0.000
2	1806	2137	0.655	0.000
3	2377	2077	0.633	0.001
4	1806	2167	0.619	0.001
5	715	1536	0.618	0.001

6	1806	2317	0.617	0.001
7	1806	1476	0.616	0.001
8	2347	2107	0.613	0.002
9	1806	2287	0.605	0.002
10	475	2047	0.602	0.002
11	445	2077	0.602	0.002
12	715	1566	0.601	0.002
13	475	2077	0.601	0.002
14	715	1506	0.600	0.002
15	445	2107	0.598	0.002
16	475	2107	0.596	0.002
17	475	2017	0.595	0.002
18	445	2047	0.595	0.002
19	445	2017	0.588	0.006
20	715	1596	0.588	0.006

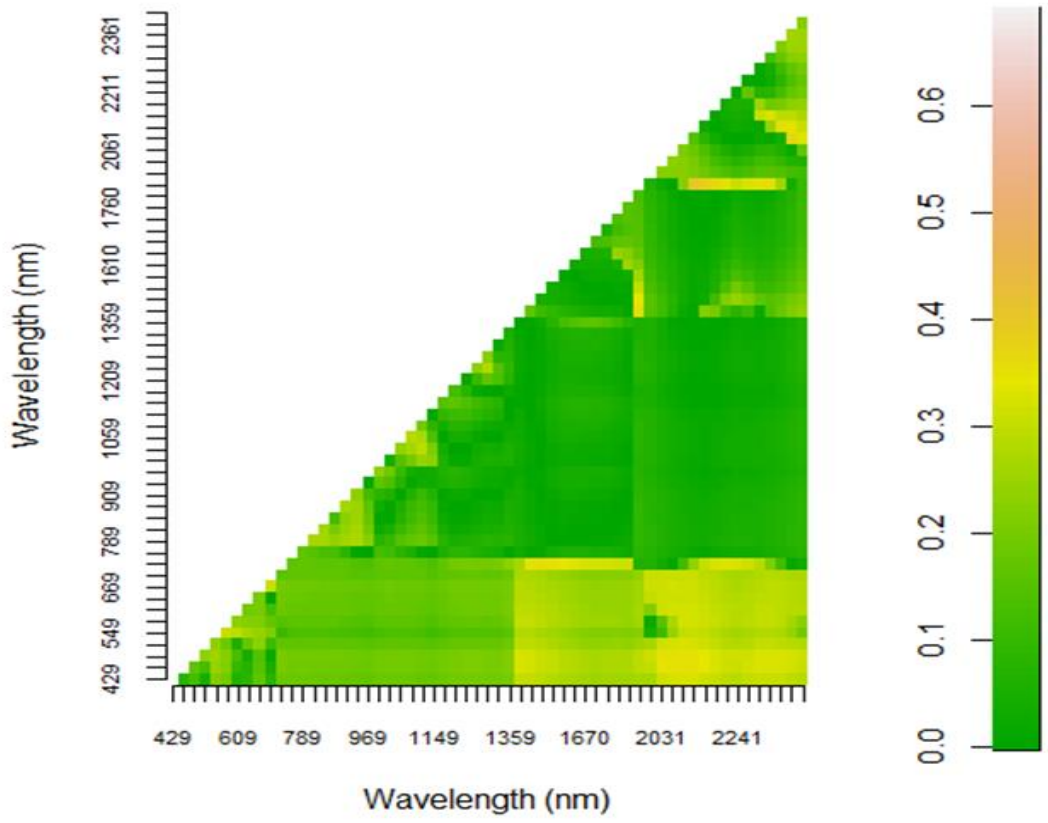


Figure 2.5: Heat map showing the correlation coefficients (R) between soybean grain yield and narrow band SR acquired from all probable band combinations from the spectral range of 400 nm to 2399 nm.

Figure 2.4 and Figure 2.5 depict a graphical presentation of the R-values for the relationship between soybean grain yield and NDVI and SR. These results show a moderate to strong relationship between NDVI/SR and the soybean grain yield (R-values from 0.588 to 0.688). In addition, the p-vales obtained for these results indicate that the relationships between soybean grain yield and the derived vegetation indices are significant as they are less than 0.05. Correlation coefficients of NDVI and SR were arranged in the order of the highest to the lowest and the top 20 R-values**.** The top 20 best NDVI/SR indices are situated in the blue (445 nm - 475 nm), red-edge (715 nm) and in the MIR regions (1506 nm – 2377 nm) of the electromagnetic spectrum (**Figure 2.4 and Figure 2.5)**.

2.4.3 Narrow-band EVI relationship to soybean grain yield

Narrow-band EVI was computed from all probable band combinations in the spectral range of 400 to 2399 nm of the electromagnetic spectrum. Spearman's correlation coefficients were calculated to assess the relationship between the EVI indices and the soybean grain yields. The correlation coefficient results of EVI indices ranged from 0.00 and 0.761. The relationship between soybean grain yield and the derived narrow-band EVI are significant as shown by the p-values less than 0.05 in **Table 2.4**. Correlation coefficients of the narrow-band EVI were ranked from the highest to the lowest and the top 20 best indices were selected and shown in **Table 2.4**. The best 20 EVIs are situated in the blue region (405 nm – 425 nm), red region (695 nm), red-edge ((705 nm- 735 nm) NIR (1245 nm) and the MIR (2357 nm– 2397 nm) regions of the electromagnetic spectrum.

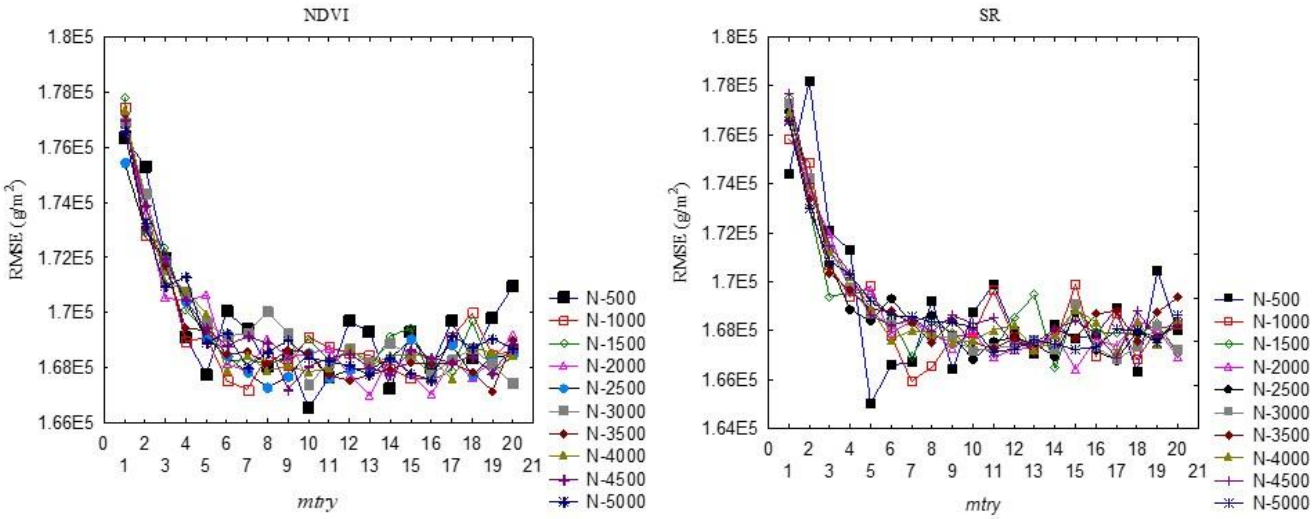
Table 2.4: Top 20 narrow-band EVI indices (λ= 10 nm) that produced the highest correlation coefficients with soybean grain yield

Ranking	Wavelength (nm)	Wavelength (nm)	Wavelength (nm)	R-values	P-values
1	2397	2357	705	0.761	0.00005
2	2387	2367	705	0.760	0.00005
3	2397	2367	705	0.757	0.00005
4	405	2357	705	0.757	0.00005
5	2387	2357	705	0.756	0.00005
6	2387	2357	695	0.752	0.00005
7	2397	2347	705	0.751	0.00006

8	405	2347	705	0.751	0.00006
9	415	2357	705	0.751	0.00006
10	415	2367	705	0.750	0.00007
11	415	2347	705	0.750	0.00007
12	2387	2347	705	0.749	0.00007
13	2397	2377	705	0.749	0.00007
14	405	2367	705	0.749	0.00007
15	2377	2357	695	0.748	0.00007
16	2377	2357	705	0.748	0.00007
17	425	2347	705	0.748	0.00007
18	425	2357	705	0.747	0.00007
19	735	1245	1325	0.746	0.00008
20	725	1245	1325	0.745	0.00008

2.4.4 Optimization of the random forest regression models

For the three indices (NDVI, SR and EVI), the ntree and mtry values were optimized using the training dataset to identify values that best predicted soybean grain yield. For each index, ntree values from 500 to 5000 were tested and mtry was tested from 1 to 20 (**Figure. 2.6**). The mtry and ntree values that produced the best RMSE were selected. According to the results (**Figure 2.6**), the best *mtry* for the NDVI and SR models were 10 and 5 and their ntree was 500 respectively. For EVI, the best mtry was 7 and the ntree was 1000.



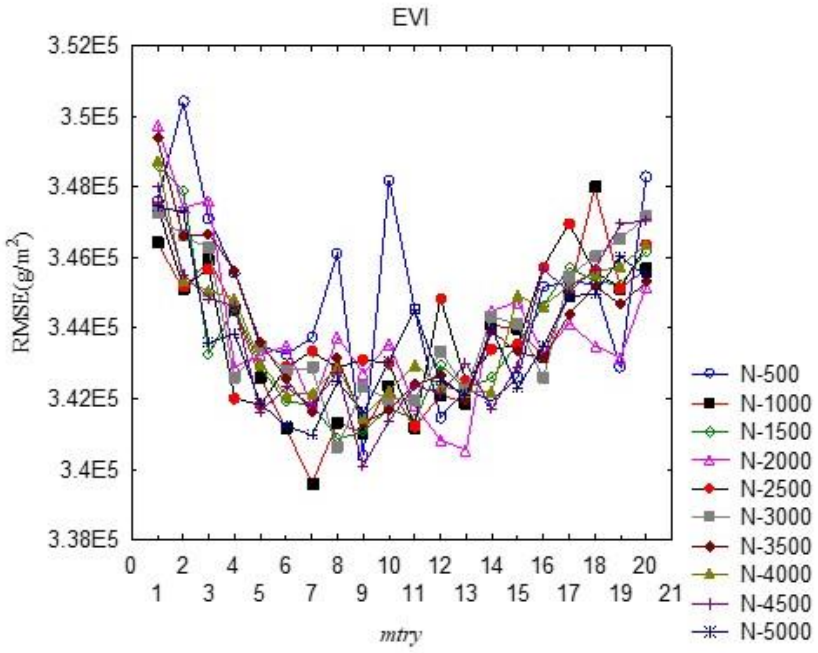


Figure 2.6: Optimization of random forest parameters (ntree (N) and mtry) using RMSE.

2.4.5 Variable importance of narrow-band indices in predicting soybean grain yield using the RF

From the best 20 selected indices that were highly correlated with the soybean grain yield, it was essential to categorize narrow-band indices of NDVI, SR and EVI that would highly perform when predicting soybean grain yield (g/m^2). The RF calculated variable importance using the mean decrease in accuracy to measure the ability of NDVI, SR and EVI to predicting soybean grain yield (g/m^2). The RF algorithm was capable of ranking the NDVI (**Figure 2.7a**), SR (**Figure 2.7b**) and EVI (**Figure 2.7c**) indices according to their importance in predicting soybean grain yield.

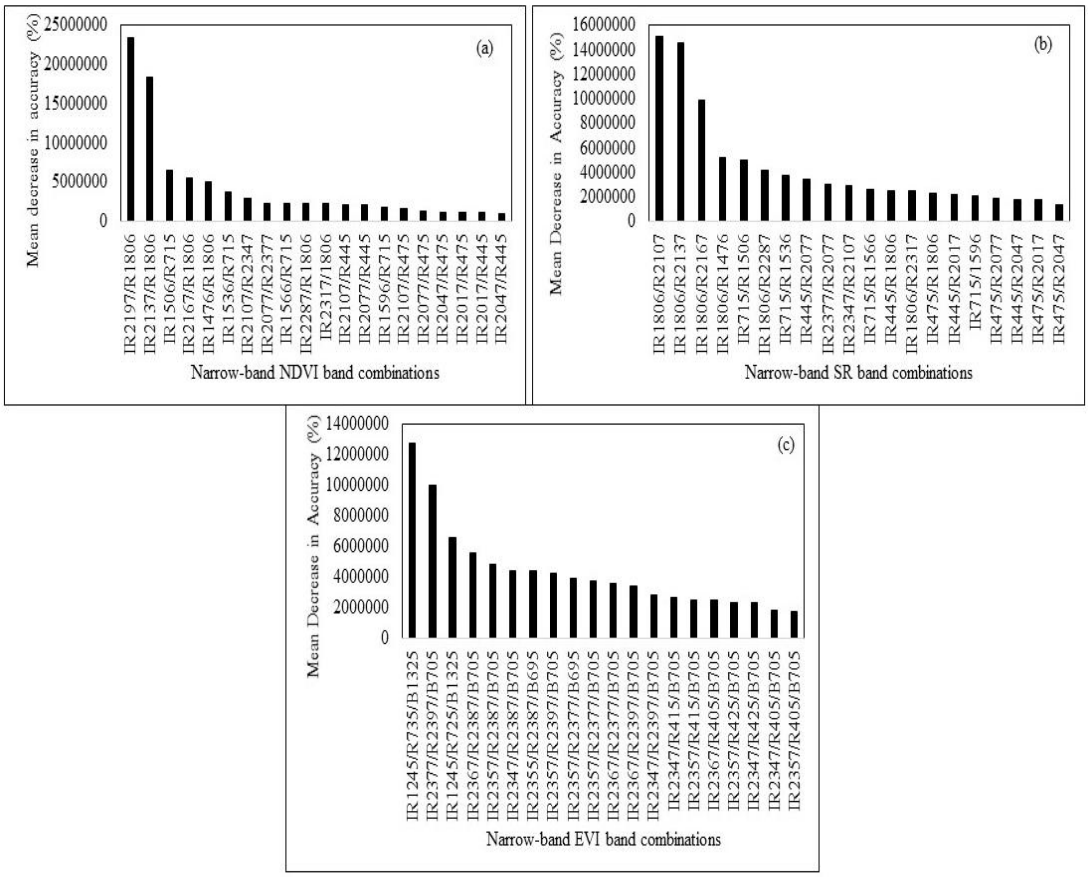


Figure 2.7: Mean Decrease in Accuracy (%) of NDVI (a), SR (b) and EVI (c) from the random forest algorithm. Important variables ranked are those with the highest mean decrease accuracy from the left of each graph.

Using the mean decrease in accuracy arrangement, top 3 wavelength combinations that had significant importance in predicting the soybean grain yield were selected. For NDVI, top 3 band combinations included: (i) 2197 nm and 1806 nm, (ii) 2137 nm and 1806 nm and (iii) 1506 nm and 715 nm. Similarly, SR top 3 important wavelength combinations includes (i) 1806 nm and 2107 nm, (ii) 1806 nm and 2137 nm and (iii) 1806 nm and 2167 nm. In addition, EVI top three significant wavelengths included (i) 1245 nm, 735 nm and 1325 nm, (ii) 2377 nm, 2397 nm and 705 nm and (iii) 1245 nm, 725 nm and 1325 nm

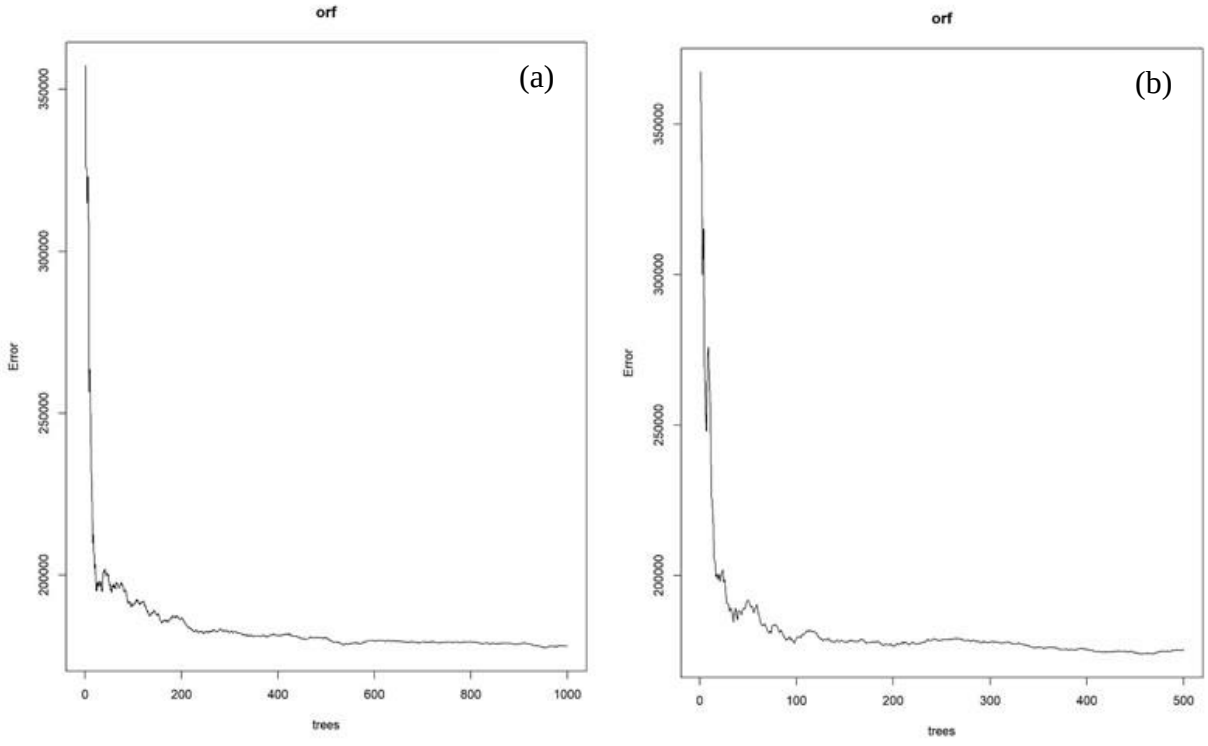
2.4.6 Accuracy assessment

Figure 2.8 shows the best *ntree* results of the RF models for NDVI (a), SR (b) and EVI (c). This indicates that for NDVI and SR, the models obtained accuracy at 500 trees and at 1000 trees for EVI. The coefficient of determination (R^2) and Root Mean Square Error (RMSE) were statistical measures that were used to evaluate the predictive performance and accuracy of the random forest

regression models (NDVI, SR and EVI). **Table 2.5,** shows the performance results of the random forest prediction models. The results show that SR obtained the highest R^2 of 0.843 with a RMSE of 422.84 (26.11% of the average soybean grain yield) compared to NDVI that obtained R^2=0.841 with an RMSE of 423.94 (26.04% of the average soybean grain yield) and EVI (R^2= 0.578) with RMSE of 615.94 (37.04% of the average soybean grain yield). These results suggest that SR can better predict soybean.

Table 2.5: Predictive performance of the NDVI, SR and EVI random forest prediction models using top 20 best indices

Narrow-band Vegetation Indices	Correlation between actual and predicted yield (R^2)	RMSE (g/m^2)
NDVI	0.841	422.84
SR	0.843	423.94
EVI	0.578	615.69



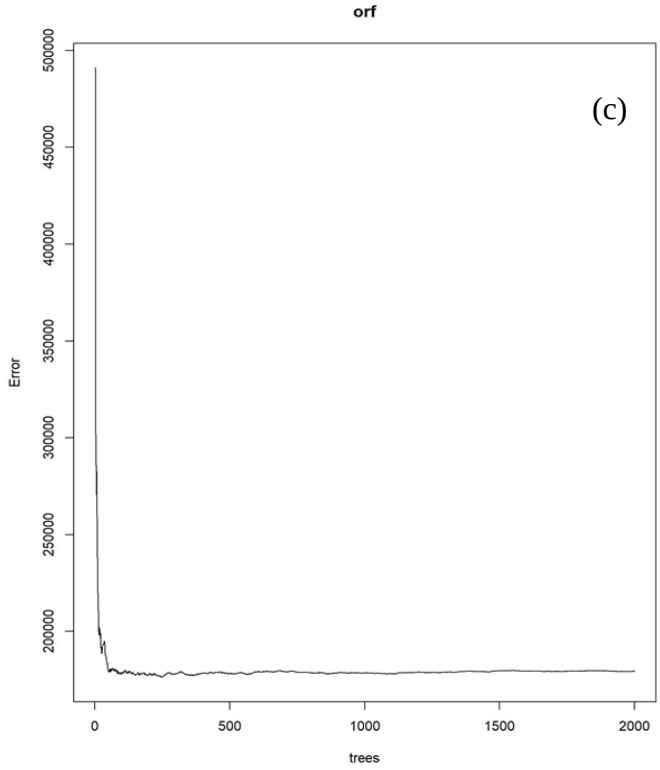


Figure 2.8: Random Forest models (NDVI (a), SR (b) and EVI (c)) showing sensitivity of ntree to the OOB error.

2.5 Discussion

The aim of the study was to evaluate the potential of narrow-band indices (NDVI, SR and EVI) in predicting soybean grain yield (g/m^2). Broadly, the results demonstrate that bands situated in the blue, red, red edge and MIR regions have a potential to predict soybean grain yield. The objectives were to assess the relationships of the narrow-band indices to soybean grain yield, identify suitable narrow-band indices that accurately predict soybean yield and to compare the accuracy of the prediction models. The study revealed that important bands in predicting soybean grain yield are not only bands in the NIR and red regions but also bands situated in the MIR region.

(i) Assessment of the relationships of narrow-band indices to soybean grain yield

The R-values obtained for NDVI (0.00-0.688), SR (0.00-0.688) and EVI (0.00-0.761) showed that different combinations of bands respond differently to variations in soybean grain yield. As shown in **Table 2.2**, **2.3** and **2.4,** strong correlations to the soybean grain yield did not only consist of combinations of bands in the red and NIR regions. Strongly correlated indices of NDVI, SR and EVI to soybean consisted of combinations of bands in the blue region (405 nm - 475 nm), red

region (695 nm), red edge (705-735 nm), NIR (1245 nm) and the MIR regions (1325 nm -2397 nm). These results are in agreement with those reported by Mutanga and Skidmore (2004), which suggested that information on vegetation biomass is not only limited to the red and NIR bands. As a result, NDVI, SR and EVI highest correlations mainly consisted of combinations of bands in the MIR (1300-2399 nm) and combinations of the blue (400-500 nm) bands and red-edge bands (700-729 nm). The MIR region is known to be sensitive to water content of leaves and has low reflectance (Kumar *et al.*, 2002). Similarly, wavelengths in the blue region are highly sensitive to chlorophyll *a* and *b* since plants absorb the violet-blue light for photosynthesis (Kumar *et al.*, 2002). Based on these results it is understandable that combinations of these bands would yield the highest correlation to the soybean grain yield. These results also concur with those reported by Darvishzadeh *et al.* (2006) and Mariotto *et al.* (2013). Darvishzadeh *et al.* (2006), showed that bands in the MIR had the strongest relationship to leaf area index (LAI) compared to the red and NIR bands. Mariotto *et al.* (2013), reported that about 74% of bands sensitive to biophysical properties were situated in the MIR (1051 to 2331 nm). Additionally, the red-edge band is characterised by high reflectance and is linked to differences in the chlorophyll content that is associated with biomass of vegetation (Kumar *et al.*, 2002, Mutanga and Skidmore, 2004). It is reasonable that combinations of wavelengths including the red-edge would obtain a strong relationship to soybean grain yield. Generally, these results provided more understanding of the relationship of the soybean grain yield and its significant wavelength regions. Furthermore, the results showed that important information on soybean yield is mostly contained in the MIR (1300 to 2399 nm) and indicate that narrow-bands have the potential to predict soybean grain yield.

(ii) Variable importance and assessment of the predictive performance of the NDVI, SR and EVI random forest models

In the top 20 selected indices that had a strong relationship to soybean grain yield, it was necessary to identify which of those were significant in the prediction of soybean grain yield. The random forest used the mean decrease in accuracy measures to identify combinations of bands that are most significant in the prediction of soybean grain yield. The results of the optimization of the random forest showed that 10, 5, and 7 indices (NDVI, SR and EVI) out of 20 indices (predictors) at 500 and 1000 ntrees were significant at predicting soybean grain yield. These results further demonstrated that accuracy of the prediction was obtained with a smaller number of trees (ntree = 500) compared to a larger number of trees (ntree = 1000). These results were validated by the differences in RMSE of 423.94 at 500 ntree compared to the RMSE = 615.69 at 1000 ntree. The

obtained results agree with those of Abdel-Rahman *et al.* (2013) who suggested that fewer number of trees (*ntree*) results in lower RMSE, which indicates better accuracy. The R^2 results of the NDVI, SR and EVI random forest models showed that SR yielded the highest R^2 in predicting soybean grain yield. These results indicate that, compared to NDVI and EVI, SR is a better index at predicting soybean grain yield. These findings are similar to those obtained by (Mutanga and Skidmore (2004)) who in their study concluded that SR was a better index at predicting biomass in dense canopies than NDVI. Higher performance of SR could be because of its high sensitivity to high biomass compared to NDVI, which saturates when faced with high biomass (Jackson and Huete, 1991, Xue and Su, 2017). Although the SR obtained the highest R^2, the NDVI obtained a lower RMSE of 422.84 compared to 423.94 for SR and 615.69 for EVI (RMSE=615.69). These findings indicate that NDVI has better accuracy at predicting soybean yield since a lower RMSE indicates better accuracy. In conclusion, these results suggest that although SR and NDVI can both accurately predict soybean yield, NDVI outperforms SR.

2.6 Conclusion

This study shows convincingly the success of narrow-band indices in predicting soybean grain yield. The results have shown that important narrow-bands in predicting soybean grain yield are not only combinations of bands situated in the red (695 nm) and the NIR (1245 nm) regions but are also combinations of bands found in the blue region (405 nm - 475 nm), red edge (705 nm - 735 nm) and the MIR regions (1325 nm -2397) nm. Furthermore, the SR index ($R^2 = 0.843$) proved to be a better index in predicting soybean grain yield compared to the NDVI ($R^2 = 0.841$) and EVI ($R^2 = 0.578$).

References

Abdel-Rahman, E. M., Ahmed, F. B. & Ismail, R. 2013. Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data. *International Journal of Remote Sensing,* 34**,** 712-728.

Abdel-Rahman, E. M., Mutanga, O., Odindi, J., Adam, E., Odindo, A. & Ismail, R. 2014. A comparison of partial least squares (PLS) and sparse PLS regressions for predicting yield of Swiss chard grown under different irrigation water sources using hyperspectral data. *Computers and Electronics in Agriculture,* 106**,** 11-19.

Adam, E. 2010. *The remote sensing of Papyrus vegetation (Cyperus papyrus L.) in swamp wetlands of South Africa.* Doctor of Philosophy in Environmental Sciences, University of KwaZulu-Natal.

Adam, E., Mutanga, O., Abdel-Rahman, E. M. & Ismail, R. 2014. Estimating standing biomass in papyrus (Cyperus papyrus L.) swamp: exploratory of in situ hyperspectral indices and random forest regression. *International Journal of Remote Sensing,* 35**,** 693-714.

Adam Equipment. 2017. *Adam Equipment Prodcucts - Weighing Scales and Equipment Manufacturer* [Online]. Available: https://www.adamequipment.co.za/products.

Adelabu, S. 2013. *The Remote Sensing of Insect Defoliation in Mopane Woodland.* Phd, University of KwaZulu-Natal.

Adelabu, S., Mutanga, O. & Cho, M. A. 2012. A review of remote sensing of insect defoliation and its implications for the detection and mapping of Imbrasia belina defoliation of Mopane Woodland. *The African Journal of Plant Science and Biotechnology,* 6**,** 1-13.

Adjorlolo, C. 2013. *REMOTE SENSING OF THE DISTRIBUTION AND QUALITY OF SUBTROPICAL C3 AND C4 GRASSES.* University of KwaZulu-Natal, Pietermaritzburg, South Africa.

Ahmad, I., Ghafoor, A., Bhatti, M. I. & Akhtar, I.-U. H. 2014. Satellite Remote Sensing and GIS based Crops Forecasting & Estimation System in Pakistan. *Crop monitoring for improved food security*.

Asd, A. S. D. 2005. Handheld spectroradiometer: user guide version 4.05. *Boulder: Analytical Spectral Devices Inc*.

Babar, M., Van Ginkel, M., Klatt, A., Prasad, B. & Reynolds, M. 2006. The potential of using spectral reflectance indices to estimate yield in wheat grown under reduced irrigation. *Euphytica,* 150**,** 155-172.

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

Board, J. E. & Kahlon, C. S. 2011. Soybean yield formation: what controls it and how it can be improved. *Soybean physiology and biochemistry.* InTech.

Boulesteix, A. L., Janitza, S., Kruppa, J. & König, I. R. 2012. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 2**,** 493-507.

Breiman, L. 2001. Random forests. *Machine learning,* 45**,** 5-32.

Cho, M. A., Skidmore, A., Corsi, F., Van Wieren, S. E. & Sobhan, I. 2007. Estimation of green grass/herb biomass from airborne hyperspectral imagery using spectral indices and partial least squares regression. *International Journal of Applied Earth Observation and Geoinformation,* 9**,** 414-424.

Darvishzadeh, R., Atzberger, C. & Skidmore, A. Hyperspectral vegetation indices for estimation of leaf area index. ISPRS Commission VII Mid-term Symposium" Remote Sensing: From Pixels to Processes", Enschede, Netherlands, 2006. 8-11.

Dell Inc 2015. Dell Statistica (data analysis software system). Version 13 ed.

Dye, M., Mutanga, O. & Ismail, R. 2011. Examining the utility of random forest and AISA Eagle hyperspectral image data to predict Pinus patula age in KwaZulu-Natal, South Africa. *Geocarto International,* 26**,** 275-289.

Esquerdo, J., Zullo Júnior, J. & Antunes, J. 2011. Use of NDVI/AVHRR time-series profiles for soybean crop monitoring in Brazil. *International Journal of Remote Sensing,* 32**,** 3711-3727.

Fao 2016. Crop Yield Forecasting: Methodological and Institutional Aspects. Food and Agriculture Organiszation of the United Nations Rome.

Fathima, A. S. & Sheriff, L. a. K. 2012. Exploring Support Vector Machines and Random Forests for the Prognostic Study of an Arboviral Disease. *International Journal of Computer Applications,* 57.

Fermont, A. & Benson, T. 2011. Estimating yield of food crops grown by smallholder farmers. *International Food Policy Research Institute, Washington DC***,** 1-68.

Huang, Y., Lee, M. A., Thomson, S. J. & Reddy, K. N. 2016. Ground-based hyperspectral remote sensing for weed management in crop production. *International Journal of Agricultural and Biological Engineering,* 9**,** 98-109.

Huete, A., Justice, C. & Liu, H. 1994. Development of vegetation and soil indices for MODIS-EOS. *Remote Sensing of Environment,* 49**,** 224-234.

Jackson, R. D. & Huete, A. R. 1991. Interpreting vegetation indices. *Preventive veterinary medicine,* 11**,** 185-200.

Jordan, C. F. 1969. Derivation of leaf-area index from quality of light on the forest floor. *Ecology,* 50**,** 663-666.

Karlson, M., Ostwald, M., Reese, H., Sanou, J., Tankoano, B. & Mattsson, E. 2015. Mapping tree canopy cover and aboveground biomass in Sudano-Sahelian woodlands using Landsat 8 and random forest. *Remote Sensing,* 7**,** 10017-10041.

Koatla, T. a. B. 2012. *Mainstreaming small-scale farmers in Qwaqwa, Free State Province, South Africa.* University of the Free State.

Kuhn, S., Egert, B., Neumann, S. & Steinbeck, C. 2008. Building blocks for automated elucidation of metabolites: Machine learning methods for NMR prediction. *BMC bioinformatics,* 9**,** 400.

Kumar, L., Schmidt, K., Dury, S. & Skidmore, A. 2002. Imaging spectrometry and vegetation science. *Imaging spectrometry.* Springer.

Liaw, A. & Wiener, M. 2002. Classification and regression by randomForest. *R news,* 2**,** 18-22.

Locke, C., Carbone, G., Filippi, A., Sadler, E., Gerwig, B. & Evans, D. Using remote sensing and modeling to measure crop biophysical variability. 5th International Conference on Precision Agriculture, 2000.

Lokupitiya, E., Lefsky, M. & Paustian, K. 2010. Use of AVHRR NDVI time series and ground-based surveys for estimating county-level crop biomass. *International Journal of Remote Sensing,* 31**,** 141-158.

Ma, B., Dwyer, L. M., Costa, C., Cober, E. R. & Morrison, M. J. 2001. Early prediction of soybean yield from canopy reflectance measurements. *Agronomy Journal,* 93**,** 1227-1234.

Mariotto, I., Thenkabail, P. S., Huete, A., Slonecker, E. T. & Platonov, A. 2013. Hyperspectral versus multispectral crop-productivity modeling and type discrimination for the HyspIRI mission. *Remote Sensing of Environment,* 139**,** 291-305.

Mashaba, Z., Chirima, G., Botai, J. O., Combrinck, L., Munghemezulu, C. & Dube, E. 2017. Forecasting winter wheat yields using MODIS NDVI data for the Central Free State region. *South African Journal of Science,* 113**,** 1-6.

Mourtzinis, S., Arriaga, F. J., Balkcom, K. S. & Ortiz, B. V. 2013. Corn grain and stover yield prediction at R1 growth stage. *Agronomy journal,* 105**,** 1045-1050.

Mukaka, M. M. 2012. A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal,* 24**,** 69-71.

Mutanga, O. & Skidmore, A. K. 2004. Narrow band vegetation indices overcome the saturation problem in biomass estimation. *International Journal of Remote Sensing,* 25**,** 3999-4014.

Noureldin, N., Aboelghar, M., Saudy, H. & Ali, A. 2013. Rice yield forecasting models using satellite imagery in Egypt. *The Egyptian Journal of Remote Sensing and Space Science,* 16**,** 125-131.

Palmer, D. S., O'boyle, N. M., Glen, R. C. & Mitchell, J. B. 2007. Random forest models to predict aqueous solubility. *Journal of chemical information and modeling,* 47**,** 150-158.

Powell, S. L., Cohen, W. B., Healey, S. P., Kennedy, R. E., Moisen, G. G., Pierce, K. B. & Ohmann, J. L. 2010. Quantification of live aboveground forest biomass dynamics with Landsat time-series and field inventory data: A comparison of empirical modeling approaches. *Remote Sensing of Environment,* 114**,** 1053-1068.

Prasad, A. M., Iverson, L. R. & Liaw, A. 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems,* 9**,** 181-199.

Rouse, J. 1974. Monitoring the vernal advancement of retrogradation of natural vegetation, NASA/GSFG, Type III. *Final Report,* 371.

Sakala, E., Fourie, F., Gomo, M. & Coetzee, H. 2017. Hydrogeological investigation of the Witbank, Ermelo and Highveld Coalfields: Implications for the subsurface transport and attenuation of acid mine drainage.

Shanahan, J. F., Schepers, J. S., Francis, D. D., Varvel, G. E., Wilhelm, W. W., Tringe, J. M., Schlemmer, M. R. & Major, D. J. 2001. Use of remote-sensing imagery to estimate corn grain yield. *Agronomy Journal,* 93**,** 583-589.

Sibanda, M., Mutanga, O. & Rouget, M. 2015. Examining the potential of Sentinel-2 MSI spectral resolution in quantifying above ground biomass across different fertilizer treatments. *ISPRS Journal of Photogrammetry and Remote Sensing,* 110**,** 55-65.

Sihlobo, W. & Kapuya, T. 2016. South Africa's soybean industry: A brief overview. [Accessed 15/02/2017].

Smyth, G. 2004. Statistical applications in genetics and molecular biology. *Linear models and empirical Bayes methods for assessing differential expression in microarray experiments*.

Teillet, P., Staenz, K. & William, D. 1997. Effects of spectral, spatial, and radiometric characteristics on remote sensing vegetation indices of forested regions. *Remote Sensing of Environment,* 61**,** 139-149.

Testa, S., Soudani, K., Boschetti, L. & Mondino, E. B. 2018. MODIS-derived EVI, NDVI and WDRVI time series to estimate phenological metrics in French deciduous forests. *International Journal of Applied Earth Observation and Geoinformation,* 64**,** 132-144.

Thenkabail, P. S., Smith, R. B. & De Pauw, E. 2000. Hyperspectral vegetation indices and their relationships with agricultural crop characteristics. *Remote sensing of Environment,* 71**,** 158-182.

Thenkabail, P. S., Smith, R. B. & De Pauw, E. 2002. Evaluation of narrowband and broadband vegetation indices for determining optimal hyperspectral wavebands for agricultural crop characterization. *Photogrammetric Engineering and Remote Sensing,* 68**,** 607-622.

Tucker, C. J. 1979. Red and photographic infrared linear combinations for monitoring vegetation. *Remote sensing of Environment,* 8**,** 127-150.

Van De Merwe, R., Van Biljon, A. & Hugo, A. Current and potential usage of soybean products as food in South Africa. 2013. Abstract.

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

Wang, Q., Nuske, S., Bergerman, M. & Singh, S. Automated crop yield estimation for apple orchards. Experimental Robotics, 2013. Springer, 745-758.

Weber, V., Araus, J., Cairns, J., Sanchez, C., Melchinger, A. & Orsini, E. 2012. Prediction of grain yield using reflectance spectra of canopy and leaves in maize plants grown under different water regimes. *Field Crops Research,* 128**,** 82-90.

Xue, J. & Su, B. 2017. Significant remote sensing vegetation indices: A review of developments and applications. *Journal of Sensors,* 2017.

Zhao, D., Reddy, K. R., Kakani, V. G., Read, J. J. & Koti, S. 2007. Canopy reflectance in cotton for growth assessment and lint yield prediction. *European Journal of Agronomy,* 26**,** 335-344.

Chapter 3

Determining the suitable growth stage to predict soybean (*Glycine max (L.) Merr*) grain yield using hyperspectral data

Abstract

Remote sensing methods can provide accurate and real-time crop status and crop statistics at various growth phases through spectral signatures at minimal costs. Spectral signatures measure the amount of radiation that interacts with the crops at various wavelengths of the electromagnetic spectrum. That measured spectral signatures can be related to crop yield because crops exhibit their growth, development, stress, yield potential and other biophysical attributes through their canopy status. This study aimed at determining the suitable growth stage to predict soybean grain yield using hyperspectral data. For this study, hyperspectral reflectance data was collected during the flowering, pod formation and seed filling stages of the soybean in the Mpumalanga and Free State experimental sites. Random forest regression algorithm was used to predict the soybean yields using the NDVI and SR vegetation indices for each growth stage. The results showed that the flowering stage was the suitable growth stage to predict soybean grain yield compared to the other growth stages. These findings were similar for both the NDVI (R^2=0.863) and the SR (R^2=0.865). The results also indicated that the SR performed better than the NDVI at all growth stages with better accuracy. Nevertheless, the findings of this study in overall indicate the possibility of predicting soybean grain yield during the flowering stage instead of monitoring the entire soybean cycle.

Keywords: Soybean yield, growth stages, hyperspectral data, NDVI, SR

3.1 Introduction

Soybean is an important agricultural crop enjoyed all over the world for its high protein, oil, minerals and vitamin content (Hartman *et al.*, 2011, Vlahović *et al.*, 2013). In Sub-Saharan Africa (SSA) and in particular, South Africa, soybean is grown in about 9.1 million households (Abate *et al.*, 2012). Production of soybean in South Africa is expected to have increased by year 2020 based on the current high demands (Bahta and Willemse, 2016). In line with the above, soybean is perceived as a vital and developing component in the South African economy (Dlamini *et al.*, 2014). South Africa currently imports about 90% of soybean products mainly from Argentina (Bahta and Willemse, 2016). The growing soybean interest in South Africa requires that soybean yield information be available on time to ensure informed planning.

Accurate pre-harvest yield predictions of crops such as soybean are important for prior and after harvest management such as grain transportation, and trade between countries concerning agricultural commodities (Noureldin *et al.*, 2013). In addition, pre-harvest yield predictions help evaluate expected yield and its market viability (Monteiro *et al.*, 2012). Pre-harvest yield statistics inform the valuing, imports of shortfalls and exports in case of surplus (Esfandiary *et al.*, 2009, Rajah *et al.*, 2017). In addition, early soybean yield predictions provide farmers with the knowledge of growth and development of the crops (Chen and Yang, 2005). Similarly, this information allows farmers time to make interventions in farming techniques and management to obtain better yields (Chen and Yang, 2005). Thus, the above suggest that accurate techniques to predict soybean grain yield during the growing season are required.

Commonly, yield predictions have been based on ground-based data, agricultural surveys and manual calculation of yield (Noureldin *et al.*, 2013). Crop yield predictions in South Africa are dependent on statistics acquired through manual field-surveys and also through surveys performed through telephones and or emails (FAO, 2016). Such techniques are often biased, expensive to carry out, and may be subject to immense mistakenness (FAO, 2016). Subsequently, data obtained through these techniques may be undependable and may delay decision-making processes. Agricultural produce is an important aspect of the country and thus, yield data from estimations need to be handled with highest care and computed with highest accuracy (FAO, 2016).

Remote sensing techniques offer accurate and real-time crop status and crop statistics at various growth phases through spectral signatures at minimal costs (Fernandez-Ordoñez and Soria-Ruiz, 2017). Spectral signatures measure the amount of radiation that interacts with the crops at various

wavelengths of the electromagnetic spectrum (Montesinos-López *et al.*, 2017). Measured spectral signatures can be related to crop yield because crops exhibit their growth, development, stress, yield potential and other biophysical attributes through their canopy status (Wiegand *et al.*, 1986). The spectral reflectance is regarded as an indirect determinant of crop yield (Raun *et al.*, 2001). Basnet *et al.* (2003), indicated that the time in which spectral reflectance is acquired could have an influence on the relationship between crop yield and spectral reflectance. Tagarakis and Ketterings (2017), hypothesized that the period of acquisition of remote sensed data may affect the accuracy of yield estimations. Based on the relationship between reflectance and crops, Raun *et al.* (2001) determined a strong correlation between NDVI and winter wheat yields during the early growing season than late in the growing season. Similarly, Spitkó *et al.* (2016) established that spectral reflectance measured during the flowering stage could be used to predict maize grain yield. These studies demonstrate the potential application of reflectance in predicting grain yield of other crops, as well as suggesting that spectral data can be useful information in making early predictions of soybean.

Soybean undergoes growth and development in two phases, the vegetative and the reproductive stages (McWilliams *et al.*, 1999). The reproductive stage is a critical phase in which the soybean utilises radiation to photosynthesise (Board and Kahlon, 2011). This stage is subdivided into four developmental stages, which are flowering (R1 and R2), pod formation (R3 and R4), seed development (R5 and R6) and maturity stages (R7 and R8) (McWilliams *et al.*, 1999). With soybean, the pod formation and seed filling stages (R3 to R6) seem to be the most important because these are the phases when the soybean forms pods and grains during photosynthesis (McWilliams *et al.*, 1999, Board and Kahlon, 2011). Pods and grains are important components, which have the capability to guide the yield outcome (Board and Kahlon, 2011). During these stages any environmental stress such as drought, temperature and precipitation variability can result in reduced yields (Pannar, 2006, Puteh *et al.*, 2013). Due to the high activity of soybean during the reproductive phase, it would be interesting to determine which of these developmental stages is suitable to predict soybean grain yield.

Often, researchers use spectral vegetation indices such as normalised difference vegetation index (NDVI) and simple ratio (SR) derived from multispectral datasets to determine the optimal time to predict crop yields (Chen and Yang, 2005, Mutanga *et al.*, 2013, Mashaba *et al.*, 2017). The NDVI and SR are amongst the primary vegetation indices developed to assess crop health and other crop attributes (Bannari *et al.*, 1995). These vegetation indices have been applied

successfully in predicting biomass of various crops such as maize, rice, and sugarcane (Ma *et al.*, 2001, Mutanga *et al.*, 2013, Noureldin *et al.*, 2013, Ngie and Ahmed, 2018). Both of these indices are calculated using the red and near infrared bands. In this study, the same indices obtained the highest accuracy in predicting soybean grain yield in the previous chapter. The difference between these indices is that the SR is highly sensitive to biomass above 50% coverage (Xue and Su, 2017). Whereas, the NDVI is less sensitive to high biomass crops (above 50% coverage) because at this point it saturates especially when computed from multispectral data (Bannari *et al.*, 1995). Due to this limitation, researchers have suggested the use of hyperspectral data to predict vegetation biomass and yield of several crops (Mutanga and Skidmore, 2004, Christenson *et al.*, 2016, Rajah *et al.*, 2017). The advantage of hyperspectral data is that it is manageable, flexible and has high temporal resolution, which are advantages in precision agriculture applications compared to other satellite products (Huang *et al.*, 2016). Furthermore, hyperspectral data has numerous spectral bands situated in the visible, near infrared and short wave infrared regions (Rajah *et al.*, 2017). The spectral bands in these regions correlate with biophysical characteristics of crops such as chlorophyll, plant cell structure and water content of the leaf (Rajah *et al.*, 2017, Menke, 2018).

Many studies have tested the capability of hyperspectral data in determining the optimal growth stage to predict biomass and crop yield (Gao *et al.*, 2012, Gutierrez *et al.*, 2012, Christenson *et al.*, 2016, Wang *et al.*, 2016). For example, Gao *et al.* (2012) used NDVI derived from ground spectral reflectance to predict grassland biomass at the peak of the growing season in China. Their study concluded that ground spectral models could be utilised as the basis for prediction of yields. Gutierrez *et al.* (2012), associated spectral reflectance indices (simple ratio (SR), ratio vegetation index (RVI), near infrared (NIR) and NDVI), with lint and cotton growth stages and lint yield to determine the optimum growth stage to predict lint yield. The study concluded that NDVI and other spectral indices explained 87% of distinction in cotton biomass. However, for lint yield, RVI (R^2=0.58), SR (R^2=0.56) and NIR (R^2=0.60) performed better than NDVI (R^2=0.47) as it reached saturation. Their study concluded that reflectance measurements could be used to estimate lint yield at peak bloom stage using SR, NIR, and RVI. Christenson *et al.* (2016), estimated soybean maturity and grain yield using canopy reflectance. That study sought to determine the growth stage that contributed most to yield prediction. The study did not obtain significant results in determining the optimal developmental stage to use canopy reflectance to predict soybean grain yield. It would be interesting to determine at which stage of the reproductive stage is it suitable to predict soybean grain yield using hyperspectral data.

Hyperspectral data in its nature is characterised by numerous narrow-bands, which are redundant (Abdel-Rahman *et al.*, 2014, Adjorlolo *et al.*, 2015). This then makes yield prediction difficult, as it is problematic to determine which bands are most valuable. Advanced statistical methods such as random forest algorithm have been suggested to be able to handle the high dimensionality and redundancy of hyperspectral data (Dye *et al.*, 2011, Abdel-Rahman *et al.*, 2013, Adjorlolo, 2013). The random forest is considered to be an accurate prediction technique when used in a regression (Chen and Ishwaran, 2012, Wang *et al.*, 2016). The aim and objective of this study was to determine the optimal growth stage to predict soybean grain yield using hyperspectral data. The NDVI and SR were used to predict the soybean as they proved to be the most useful indices for predicting soybean grain yield in the previous chapter. The second objective was to compare the predictive ability of NDVI and SR models regarding soybean yields.

3.2 Materials and Methods

3.2.1 Study area

The study was conducted on the same two experimental farms established in Mpumalanga (Ermelo) and Free State (Phuthaditjhaba) provinces as described in Chapter2. The Pedology division at the Agricultural Research Council (ISCW) planned the experiments in these farms. Climatic conditions for the study areas are characterised as warm summers and cold winters. The areas receive an average of 625 mm of precipitation dominantly during the summer season from October to March. Soil on these farms can be described as rich loam in Phuthaditjhaba (Koatla, 2012) whereas in Ermelo it can be described as low clay and sandy (Sakala *et al.*, 2017)

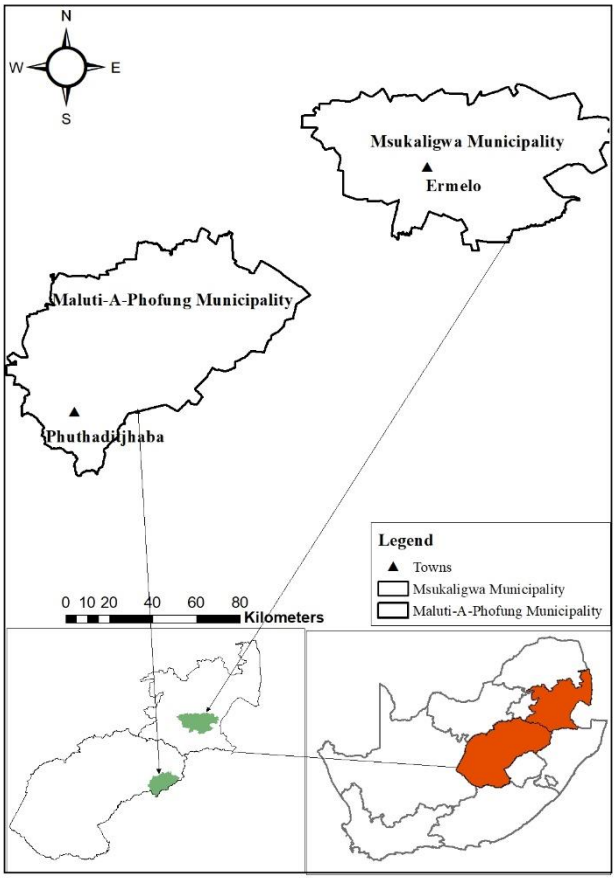


Figure 3.1: Locality map of Ermelo and Phuthaditjhaba in the provinces of South Africa.

3.2.2 Field experiment

A split plot Randomized Complete Block design was followed in the experiments on the two study locations as describe in chapter 2. Soybean was sown in 72 experimental plots with the size of 7 m length and 3 m width. In each plot, there were 7 rows of soybean that were 60 cm apart replicated three times. The soybean planted were three cultivars from Pannar seeds (PANN 1500R, PANN 1614R and PANN 1664R) that were evenly distributed amongst the 72 plots. The cultivars were planted from 13th to 15 December 2016 in Ermelo and from the 19th to 21 December 2016 in Phuthaditjhaba. Phosphorus fertilizer treatments (0 kg, 30 kg and 60 kg) were administered in the plots to determine if the effect of fertilizer would have effect on the yield. For irrigation, the soybean depended on rainwater.

3.2.3 Hyperspectral and soybean grain yield data acquisition

Spectral reflectance signatures were measured in March 2017 and April 2017. During the first visit, the soybean was in the pod formation stage in Mpumalanga site, while in the FS site it was in the flowering stage. The spectral reflectance measurements were recorded using an Analytical

Spectral Device (ASD) spectroradiometer (ASD, 2005) as described in chapter2. Later on the spectral measurements were averaged to get the mean spectral measurement for every plot.

At the end of the growing season, soybean grains were harvested in May 2017 in Ermelo and in June 2017 in Phuthaditjhaba. The obtained soybean grains were measured and using the LBK1 scale from ADAM Equipment (Adam Equipment, 2017) for every plot. The statistics of the total soybean grain yield that was measured are shown in **Table 3.1**.

Table 3.1: Statistics of measured soybean grain yield (g/m^2)

No. Of Samples	Min (g/m^2)	Max (g/m^2)	Mean (g/m^2)	Standard deviation (g/m^2)
144	202	3374	1589.4	734.4628

3.3 Data analysis

3.3.1 Analysis of hyperspectral data

Reflectance measurements of 448 bands situated from 350 to 399 nm, 1350 to 1450 nm, 1800 to 1950 nm and 2400 to 2500 nm were not included in the analysis due to atmospheric water absorption and noise in those regions (Abdel-Rahman *et al.*, 2014, Adam *et al.*, 2014). Thus, only 1702 narrow wavelengths situated between 400 nm and 2399 nm were used for analysis. The NDVI and SR were computed utilising the standard equations for these indices shown in **Table 3.2** to evaluate the suitable growth stage for predicting the soybean grain yield.

Table 3.2: Predictor variables used to predict soybean grain yield.

Growth stage	**Vegetation indices**	**Equation**	**Reference**
Flowering, Pod formation, Seed filling stages	NDVI	$NDVI = \frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2}$	(Rouse, 1974)
	SR	$SR = \frac{\lambda_1}{\lambda_2}$	(Jordan, 1969)

3.3.2 Statistical analysis

For each growth stage, the random forest regression was applied to predict soybean grain yield. The random forest is an algorithm in which groups of trees (*ntree*) are created based on the random

selection of samples in the training data (Breiman, 2001). In a regression, random forest allows trees to increase to the maximum size without pruning depending on the sample of the training data (Breiman, 2001). For each tree, the random forest then develops a randomized subcategory of predictors (*mtry*) to determine the best split at the node of the tree (Abdel-Rahman *et al.*, 2013). The random forest then averages the results of the total number of trees in order to attain the complete prediction (Prasad *et al.*, 2006b). To implement the random forest, two parameters that are *ntree* and *mtr*y need to be tuned (Abdel-Rahman *et al.*, 2013).

During the regression process, the random forest is able to determine variables that are important in the prediction using the built in permutation variable importance and the Gini index (Boulesteix *et al.*, 2012, Chen and Ishwaran, 2012). The Gini index importance measures node purity for node splitting, as a result it is mostly used for classification purposes (Chen and Ishwaran, 2012). On the other hand, permutation variable importance measure importance by determining the difference between OOB error resultant of the data attained via random selection of predictors and the OOB error attained from the initial dataset (Boulesteix *et al.*, 2012). Between these two measures, the permutation of variables is mostly applied measure in regression (Xi and Hemant, 2012). This is because, the permutation variable importance provides the prediction accuracy (Boulesteix *et al.*, 2012). In this study, permutation of variable importance was used to determine important indices in predicting soybean.

For assessing the accuracy of the model, the random forest utilised the out of bag (OOB) data that was not used in the training of the model (Prasad *et al.*, 2006b, Powell *et al.*, 2010). The random forest then calculates the OOB error to evaluate variation in estimation from the 2/3 of data utilised to train the model and 1/3 OOB data used to test the model (Abdel-Rahman *et al.*, 2013). OOB error is reliable measure of accuracy as it utilises data that was not used in the training of the model (Boulesteix *et al.*, 2012). In this way, it acts as a cross-validation measure. For this reason, researchers have indicated that there is no requirement for an independent dataset for validating the model (Prasad *et al.*, 2006b, Adjorlolo, 2013, Karlson *et al.*, 2015). The mean accuracy error (MAE), root mean square error (RMSE) and coefficient of determination (R^2) were utilised as ways to assess the predictive accuracy of the models. The random forest algorithm was executed using the *randomForest* built in package in R statistical software to predict the soybean grain yield (Liaw and Wiener, 2002).

3.4 Results

The results in **Table 3.3** show the prediction and accuracy of NDVI and SR vegetation indices calculated during the flowering, pod formation and seed filling growth stages of soybean. **Figure 3.2** is graphical presentation of the prediction results of soybean by both NDVI and SR. For NDVI, the results show that the highest prediction of soybean was obtained during the flowering stage (R^2=0.863) followed by the pod formation stage (R^2=0.856) and lastly the seed filling stage (R^2=0.771). Similarly, with the SR, the highest performing growth stage in predicting soybean appeared to be the flowering stage with an R^2= 0.865 followed by pod formation stage (R^2=0.857) and the lastly the seed filling stage (R^2=0.777).

Table 3.3: Performance of NDVI and SR in predicting soybean grain yield during flowering, pod formation, and seed filling stages.

Vegetation indices	**R^2**	**RMSE**	**MAE**	**P-value**
NDVI $_{\text{Flowering stage}}$	0.863	382.425	295.39	0.0000
NDVI $_{\text{pod formation stage}}$	0.856	394.88	309.80	0.0000
NDVI $_{\text{seed filling stage}}$	0.771	494.09	380.82	0.0000
SR $_{\text{flowering stage}}$	0.865	379.81	293.15	0.0000
SR $_{\text{pod formation stage}}$	0.857	394.286	308.81	0.0000
SR $_{\text{seed filling stage}}$	0.777	484.37	373.41	0.0000

The p-values in **Table 3.3** show that prediction results during the growth stages are all significant as they are less than 0.05. In overall, these results indicate that the most suitable growth stage to predict soybean grain yield was during the flowering stage as determined by both the NDVI (R^2 = 0.863) and SR (R^2 = 0.865). Furthermore, the R^2 results show that the SR performed better that the NDVI in predicting soybean grain yield at all the growing stages. Similarly, the RMSE and MAE also illustrate that the SR had better accuracy when predicting soybean grain yield during the all the growth stages than NDVI.

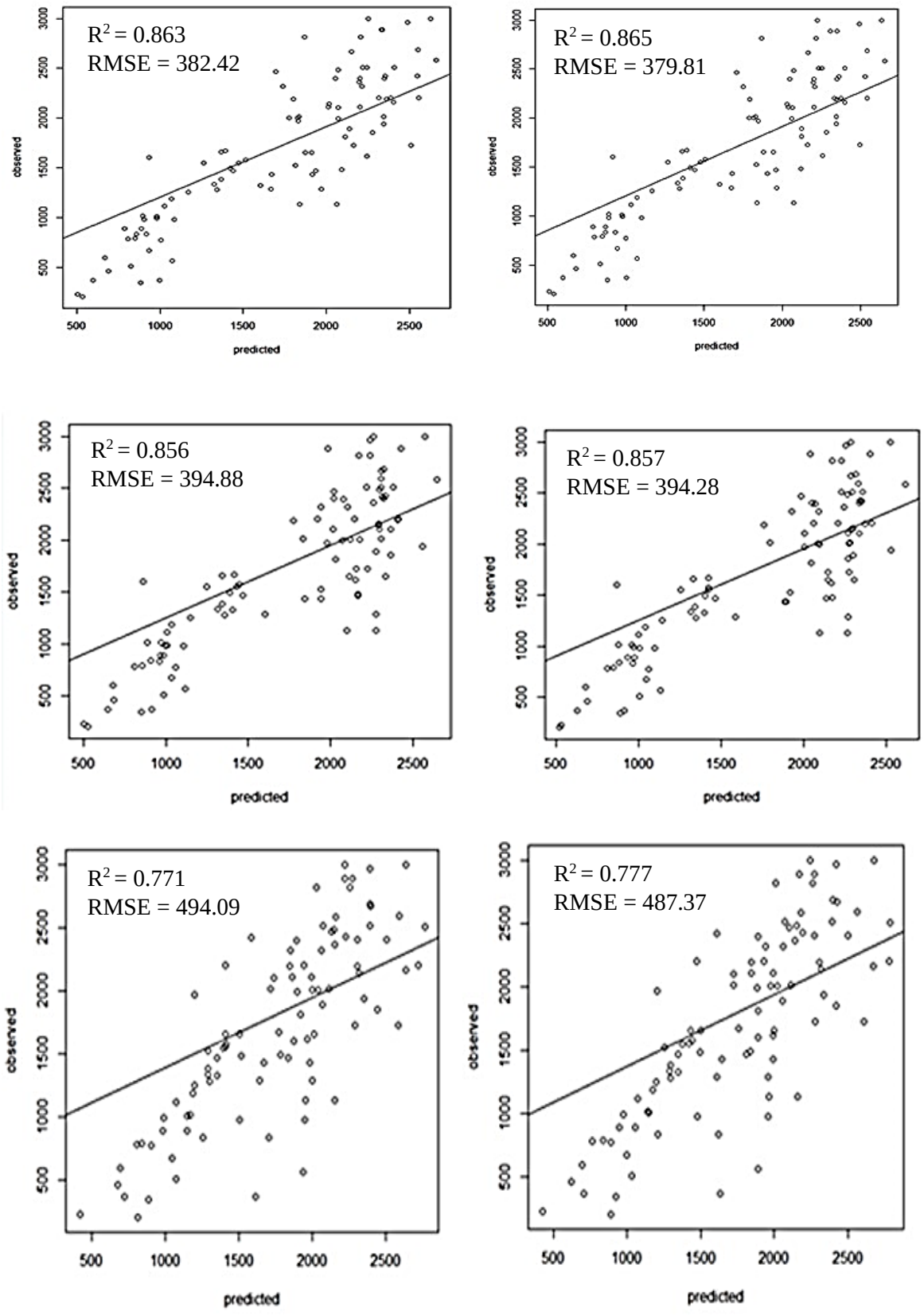


Figure 3.2: One on one relationship of predicted and observed soybean grain yields.

3.5 Discussion

Crop yield prediction before harvest is an important exercise that supports the planning and management of agricultural produce. This study sought to determine the suitable growth stage to predict soybean grain yield using remote sensing hyperspectral data before harvest. NDVI and SR vegetation indices were used to predict the soybean grain yield during flowering, pod formation and seed filling stages. The results showed that, it would be most suitable to predict soybean grain yield during the flowering stage as determined by both NDVI and SR with significant p-values ($p < 0.05$).

Prediction results in **Table 3.3** showed that the highest NDVI and SR R^2 values in predicting soybean grain yield were obtained during the flowering stage compared to other growth stages. These results indicate that the flowering growth stage is the suitable growth stage to predict soybean grain yield using hyperspectral data. The results of this study are similar to those obtained by Zhang *et al.* (1999) who determined a strong relationship between soybean grain yield and spectral measurements during the flowering stage and weaker correlations between early and late growth stages. Similarly, Fernandez-Ordoñez and Soria-Ruiz (2017) obtained effective results when predicting maize grain yield during the flowering stage. Also, Spitkó *et al.* (2016) obtained strong correlations between maize yield and measurements acquired during flowering stage than late in the growing season of maize. This is because high amount of solar radiation introduced to soybean at the flowering stage results in high soybean grain yield (Müller *et al.*, 2017). Due to this, studies have shown that radiation intensity during the late flowering and midpod formation stages is more important than during the vegetative and late reproductive stages (Mathew *et al.*, 2000). Similarly, other studies have indicated that the soybean results in poor yield once it is exposed to shading during the flowering period (Mathew *et al.*, 2000, Board and Kahlon, 2011). The findings of this study, however disagree with the results obtained by (Ma *et al.* (2001)) who suggested that the optimal stage to predict soybean grain yield was between R4 and R5. The differences in results of these studies could be based on differences in the types of soybean cultivars used in both studies, geographic locations of studies and differences in the instrument used to capture the canopy reflectance. Although, the flowering stage appeared to produce better predictions of soybean than other growth stages, it is important to note that the pod formation stage obtained results close to the flowering stage for both NDVI ($R^2 = 0.856$) and SR ($R^2 = 0.857$). The reason for this is that the flowering and pod formation stages of soybean overlap (McWilliams *et al.*, 1999). However, results showed differences in performance between the other growth stages

and the seed filling stage and these findings are in agreement with those reported by Spitkó *et al.* (2016) and Zhang *et al.* (1999). The reason for this is that the rate of photosynthetic activity decreases as crops approach maturity stage (Board and Kahlon, 2011). As a result, high reflectance can be observed in the visible region of the spectrum during the seed filling stage than other growth stages that have high absorption as already shown in **Figure 2.2, p18.**

Both NDVI and SR performed well in determining the best stage to predict soybean grain yield during the growth stages. These results are similar to those obtained by (Mutanga and Skidmore, 2004) who observed that SR was better at predicting biomass compared to NDVI and TVI. Better performance of the SR may be attributed to high correlation with high biomass vegetation of soybean (Xue and Su, 2017). As for NDVI, results show that it may have reached saturation level since soybean has high biomass apart from being influenced by other factors such as its high sensitivity to soil such as brightness, colour and atmospheric changes (Xue and Su, 2017). These might have caused NDVI to perform less than the SR since it is not sensitive to these factors.

3.6 Conclusion

The present study investigated the optimal growth stage to predict soybean grain yield using hyperspectral data acquired during the flowering, pod formation and seed filling stages. NDVI and SR vegetation indices were calculated using the random forest regression method and compared to yield data that were obtained from 3 experimental plots. The results showed that the most suitable growth stage to predict soybean grain yield was during the flowering stage when using both NDVI (R^2 = 0.863) and SR (R^2 = 0.865) with SR (RMSE = 379.81) performing better than NDVI (RMSE = 382.4) in terms of accuracy. Overall, the results of this study demonstrate that it is possible to use NDVI and SR to accurately predict soybean grain yield during the flowering growth stage instead of monitoring the soybean throughout its entire life cycle. However, it is recommended that further studies be conducted using different indices in order to validate the observed results in this study.

References

Abate, T., Alene, A. D., Bergvinson, D., Shiferaw, B., Silim, S., Orr, A. & Asfaw, S. 2012. Tropical Grain Legumes in Africa and South Asia. *Knowledge and Opportunities*.

Abdel-Rahman, E. M., Ahmed, F. B. & Ismail, R. 2013. Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data. *International Journal of Remote Sensing,* 34**,** 712-728.

Abdel-Rahman, E. M., Mutanga, O., Odindi, J., Adam, E., Odindo, A. & Ismail, R. 2014. A comparison of partial least squares (PLS) and sparse PLS regressions for predicting yield of Swiss chard grown under different irrigation water sources using hyperspectral data. *Computers and Electronics in Agriculture,* 106**,** 11-19.

Adam, E., Mutanga, O., Abdel-Rahman, E. M. & Ismail, R. 2014. Estimating standing biomass in papyrus (Cyperus papyrus L.) swamp: exploratory of in situ hyperspectral indices and random forest regression. *International Journal of Remote Sensing,* 35**,** 693-714.

Adam Equipment. 2017. *Adam Equipment Prodcucts - Weighing Scales and Equipment Manufacturer* [Online]. Available: https://www.adamequipment.co.za/products.

Adjorlolo, C. 2013. *REMOTE SENSING OF THE DISTRIBUTION AND QUALITY OF SUBTROPICAL C3 AND C4 GRASSES.* University of KwaZulu-Natal, Pietermaritzburg, South Africa.

Adjorlolo, C., Mutanga, O. & Cho, M. A. 2015. Predicting C3 and C4 grass nutrient variability using in situ canopy reflectance and partial least squares regression. *International Journal of Remote Sensing,* 36**,** 1743-1761.

Asd, A. S. D. 2005. Handheld spectroradiometer: user guide version 4.05. *Boulder: Analytical Spectral Devices Inc*.

Bahta, Y. T. & Willemse, J. 2016. The comparative advantage of South Africa soybean production. *OCL,* 23**,** A301.

Bannari, A., Morin, D., Bonn, F. & Huete, A. 1995. A review of vegetation indices. *Remote sensing reviews,* 13**,** 95-120.

Basnet, B. B., Apan, A., Kelly, R., Jensen, T., Strong, W. & Butler, D. Relating satellite imagery with grain protein content. Proceedings of the 2003 Spatial Sciences Institute Biennial Conference: Spatial Knowledge Without Boundaries (SSC2003), 2003. Spatial Sciences Institute, 1-11.

Board, J. E. & Kahlon, C. S. 2011. Soybean yield formation: what controls it and how it can be improved. *Soybean physiology and biochemistry.* InTech.

Boulesteix, A. L., Janitza, S., Kruppa, J. & König, I. R. 2012. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 2**,** 493-507.

Breiman, L. 2001. Random forests. *Machine learning,* 45**,** 5-32.

Chen, R.-K. & Yang, C.-M. 2005. Determining the optimal timing for using LAI and NDVI to predict rice yield. *J. Photogramm. Remote Sens,* 10**,** 239-254.

Chen, X. & Ishwaran, H. 2012. Random forests for genomic data analysis. *Genomics,* 99**,** 323-329.

Christenson, B. S., Schapaugh, W. T., An, N., Price, K. P., Prasad, V. & Fritz, A. K. 2016. Predicting Soybean Relative Maturity and Seed Yield Using Canopy Reflectance. *Crop Science,* 56**,** 625-643.

Dlamini, T. S., Tshabalala, P. & Mutengwa, T. 2014. Soybeans production in South Africa. *OCL,* 21**,** D207.

Dye, M., Mutanga, O. & Ismail, R. 2011. Examining the utility of random forest and AISA Eagle hyperspectral image data to predict Pinus patula age in KwaZulu-Natal, South Africa. *Geocarto International,* 26**,** 275-289.

Esfandiary, F., Aghaie, G. & Mehr, A. D. 2009. Wheat yield prediction through agro meteorological indices for Ardebil District. *World Academy of Science, Engineering and Technology,* 49**,** 32-35.

Fao 2016. Crop Yield Forecasting: Methodological and Institutional Aspects. Food and Agriculture Organiszation of the United Nations Rome.

Fernandez-Ordoñez, Y. M. & Soria-Ruiz, J. Maize crop yield estimation with remote sensing and empirical models. 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2017. IEEE, 3035-3038.

Gao, J.-X., Chen, Y.-M., Lü, S.-H., Feng, C.-Y., Chang, X.-L., Ye, S.-X. & Liu, J.-D. 2012. A ground spectral model for estimating biomass at the peak of the growing season in Hulunbeier grassland, Inner Mongolia, China. *International journal of remote sensing,* 33**,** 4029-4043.

Gutierrez, M., Norton, R., Thorp, K. R. & Wang, G. 2012. Association of spectral reflectance indices with plant growth and lint yield in upland cotton. *Crop science,* 52**,** 849-857.

Hartman, G. L., West, E. D. & Herman, T. K. 2011. Crops that feed the World 2. Soybean—worldwide production, use, and constraints caused by pathogens and pests. *Food Security,* 3**,** 5-17.

Huang, Y., Lee, M. A., Thomson, S. J. & Reddy, K. N. 2016. Ground-based hyperspectral remote sensing for weed management in crop production. *International Journal of Agricultural and Biological Engineering,* 9**,** 98-109.

Jordan, C. F. 1969. Derivation of leaf-area index from quality of light on the forest floor. *Ecology,* 50**,** 663-666.

Karlson, M., Ostwald, M., Reese, H., Sanou, J., Tankoano, B. & Mattsson, E. 2015. Mapping tree canopy cover and aboveground biomass in Sudano-Sahelian woodlands using Landsat 8 and random forest. *Remote Sensing,* 7**,** 10017-10041.

Koatla, T. a. B. 2012. *Mainstreaming small-scale farmers in Qwaqwa, Free State Province, South Africa.* University of the Free State.

Liaw, A. & Wiener, M. 2002. Classification and regression by randomForest. *R news,* 2**,** 18-22.

Ma, B., Dwyer, L. M., Costa, C., Cober, E. R. & Morrison, M. J. 2001. Early prediction of soybean yield from canopy reflectance measurements. *Agronomy Journal,* 93**,** 1227-1234.

Mashaba, Z., Chirima, G., Botai, J. O., Combrinck, L., Munghemezulu, C. & Dube, E. 2017. Forecasting winter wheat yields using MODIS NDVI data for the Central Free State region. *South African Journal of Science,* 113**,** 1-6.

Mathew, J. P., Herbert, S. J., Zhang, S., Rautenkranz, A. A. & Litchfield, G. V. 2000. Differential response of soybean yield components to the timing of light enrichment. *Agronomy Journal,* 92**,** 1156-1161.

Mcwilliams, D., Berglund, D. & Endres, G. 1999. Soybean growth and management quick guide. *North Dakota State University and University of Minnesota*.

Menke, E. J. 2018. *Using spectral reflectance in soybean breeding: evaluating genotypes for soybean sudden death disease resistance and grain yield.* Kansas State University.

Monteiro, P. F. C., Angulo Filho, R., Xavier, A. C. & Monteiro, R. O. C. 2012. Assessing biophysical variable parameters of bean crop with hyperspectral measurements. *Scientia Agricola,* 69**,** 87-94.

Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Los Campos, G., Alvarado, G., Suchismita, M., Rutkoski, J., González-Pérez, L. & Burgueño, J. 2017. Predicting grain yield using canopy hyperspectral reflectance in wheat breeding data. *Plant methods,* 13**,** 4.

Müller, M., Rakocevic, M., Caverzan, A. & Chavarria, G. 2017. Grain yield differences of soybean cultivars due to solar radiation interception. *American Journal of Plant Sciences,* 8**,** 2795.
Mutanga, O. & Skidmore, A. K. 2004. Narrow band vegetation indices overcome the saturation problem in biomass estimation. *International Journal of Remote Sensing,* 25**,** 3999-4014.
Mutanga, S., Van Schoor, C., Olorunju, P. L., Gonah, T. & Ramoelo, A. 2013. Determining the best optimum time for predicting sugarcane yield using hyper-temporal satellite imagery. *Advances in Remote Sensing,* 2**,** 269.
Ngie, A. & Ahmed, F. 2018. Estimation of Maize grain yield using multispectral satellite data sets (SPOT 5) and the random forest algorithm. *South African Journal of Geomatics,* 7**,** 11-30.
Noureldin, N., Aboelghar, M., Saudy, H. & Ali, A. 2013. Rice yield forecasting models using satellite imagery in Egypt. *The Egyptian Journal of Remote Sensing and Space Science,* 16**,** 125-131.
Pannar 2006. Soybeans Production Guide. Greytown, South Africa: Pannar Seed (Pty) Ltd.
Powell, S. L., Cohen, W. B., Healey, S. P., Kennedy, R. E., Moisen, G. G., Pierce, K. B. & Ohmann, J. L. 2010. Quantification of live aboveground forest biomass dynamics with Landsat time-series and field inventory data: A comparison of empirical modeling approaches. *Remote Sensing of Environment,* 114**,** 1053-1068.
Prasad, A. M., Iverson, L. R. & Liaw, A. 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems,* 9**,** 181-199.
Puteh, A. B., Thuzar, M., Mondal, M. M. A., Abdullah, A. & Halim, M. R. A. 2013. Soybean [Glycine max (L.) Merrill] seed yield response to high temperature stress during reproductive growth stages. *Australian Journal of Crop Science,* 7**,** 1472.
Rajah, P., Odindi, J., Abdel-Rahman, E. & Mutanga, O. 2017. Determining the optimal phenological stage for predicting common dry bean (Phaseolus vulgaris) yield using field spectroscopy. *South African Journal of Plant and Soil,* 34**,** 379-388.
Raun, W. R., Solie, J. B., Johnson, G. V., Stone, M. L., Lukina, E. V., Thomason, W. E. & Schepers, J. S. 2001. In-season prediction of potential grain yield in winter wheat using canopy reflectance. *Agronomy Journal,* 93**,** 131-138.
Rouse, J. 1974. Monitoring the vernal advancement of retrogradation of natural vegetation, NASA/GSFG, Type III. *Final Report,* 371.
Sakala, E., Fourie, F., Gomo, M. & Coetzee, H. 2017. Hydrogeological investigation of the Witbank, Ermelo and Highveld Coalfields: Implications for the subsurface transport and attenuation of acid mine drainage.
Spitkó, T., Nagy, Z., Zsubori, Z., Szőke, C., Berzy, T., Pintér, J. & Marton, C. 2016. Connection between normalized difference vegetation index and yield in maize. *Plant, Soil and Environment,* 62**,** 293-298.
Tagarakis, A. C. & Ketterings, Q. M. 2017. In-season estimation of corn yield potential using proximal sensing. *Agronomy Journal,* 109**,** 1323-1330.
Vlahović, B., Ilin, S. & Puškarić, A. 2013. Status and Perspectives of Soybean Production Worldwide and in the Republic of Serbia. *Economic Insights-Trends & Challenges,* 65.
Wang, L. A., Zhou, X., Zhu, X., Dong, Z. & Guo, W. 2016. Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *The Crop Journal,* 4**,** 212-219.
Wiegand, C. L., Richardson, A. J., Jackson, R. D., Pinter, P. J., Aase, J. K., Smika, D. E., Lautenschlager, L. F. & Mcmurtrey, J. 1986. Development of agrometeorological crop model inputs from remotely sensed information. *IEEE transactions on geoscience and remote sensing,* 24**,** 90-98.
Xue, J. & Su, B. 2017. Significant remote sensing vegetation indices: A review of developments and applications. *Journal of Sensors,* 2017.

Zhang, M., Hendley, P., Drost, D., O'neill, M. & Ustin, S. 1999. Corn and soybean yield indicators using remotely sensed vegetation index. *Precision Agriculture***,** 1475-1481.

Chapter 4

Assessing the ability of Sentinel-2 Multispectral Instrument (MSI) to estimate soybean (*Glycine max (L.) Merr.*) grain yield from resampled hyperspectral data

Abstract

Yield prediction is an important exercise for every country as pre-harvest and after harvest planning is based on that information. Crop monitoring using remote sensing technology was introduced in the field of agriculture to move away from conventional field survey methods that are subjective and expensive to implement. Although remote sensing techniques are objective and accurate, there are challenges experienced with high-resolution hyperspectral data and low-resolution multispectral data. High-resolution hyperspectral data is costly to obtain and expensive to process because of data redundancy. Through innovations in the field of remote sensing, high-resolution multispectral satellites such as Sentinel-2 MSI have been produced to overcome challenges between high-resolution hyperspectral and low-resolution multispectral data. As such, this study sought to assess the potential of Sentinel-2 MSI in predicting soybean grain yield. Because Landsat 8 OLI is also freely available and Worldview-2 has similar spectral configuration with Sentinel-2 MSI, it was therefore important to compare their performances. To achieve this, field hyperspectral data was resampled to Sentinel-2 MSI spectral bands. Sentinel-2 was compared to Landsat 8 OLI and WorldView-2 satellites. Spectral band sensitivity to soybean grain yield of the multispectral satellites was tested and the performances of Sentinel-2 MSI derived indices were compared to Landsat 8 and WorldView-2 derived indices. The results showed that bands highly sensitive to soybean for Sentinel-2 MSI included the blue, red and re-edge whereas for Landsat 8 and WorldView-2 included the red, blue and coastal blue bands. Sentinel-2 MSI derived indices (b8 and b7) predicted soybean grain yield better than Landsat 8 and WorldView-2 derived indices. In overall, the findings of this study show that Sentinel-2 MSI has the potential to predict soybean grain yield. These findings indicate that Sentinel-2 MSI could be utilised in monitoring and estimating soybean grain yield.

Keywords: soybean yield, spectral resampling, multispectral resolutions

4.1 Introduction

Soybean is amongst the six main legumes planted across the Sub-Saharan Africa (Abate *et al.*, 2012). In Africa, South Africa is among the leading producers of soybean along with Uganda and Nigeria (Abate *et al.*, 2012). The soybean undergoes growth and development over a period of five months (Smit, 2000). During growth and development, the soybean is highly susceptible to biotic factors such as pests, diseases and weeds and abiotic factors such as atmospheric influences and soil conditions that can affect yields (Board and Kahlon, 2011). Specifically, soybean is highly vulnerable to drought during flowering and pod formation stages and can result in premature death and thereby reducing the yields (Pannar, 2006). Given the growth dynamics of soybean, it is imperative to regularly observe and monitor the soybean using reliable methods such as remote sensing. Crop data in South Africa is presently obtained through field surveys (FAO, 2016) that are expensive to carry out. Data acquired through these methods are prone to subjectivity and inaccuracies (Noureldin *et al.*, 2013, Ngie and Ahmed, 2018). Crop growth statistics are important in the country they are used for various purposes such as yield predictions, logistic planning before harvest and after harvest. Thus, crop data needs to be accurate and acquired on time since yield predictions guide farmers and government to make informed decisions that have an impact on the pricing of agricultural products.

Traditional remote sensing satellites have demonstrated to be efficient tools in monitoring and predicting yields of agricultural crops such as rice, sugarcane, maize and soybean (Bappel *et al.*, 2005, Cicek *et al.*, 2010, Noureldin *et al.*, 2013, Ngie and Ahmed, 2018). Specifically for soybean, Prasad *et al.* (2006a) derived the normalized difference vegetation index (NDVI) from the advanced very high-resolution radiometer (AVHRR) to predict soybean yield. They obtained an R^2 of 0.86 prediction of soybean and recommended that their model could be enhanced by using higher spatial resolution data and with higher temporal resolution. Lokupitiya *et al.* (2010), utilised AVHRR NDVI time series and ground-based data surveys to predict country level biomass of crops such as corn, soybean and oats in Iowa, USA. The study obtained high correlation between crop biomass and the NDVI. However, they recommended utilization of satellite data with higher spatial, spectral and radiometric resolutions to obtain better results. Locke *et al.* (2000), computed Simple Ratio (SR), NDVI, Soil Adjusted Vegetation Index (SAVI) and Transformed SAVI (TSAVI) to examine soybean biophysical properties (leaf area index (LAI), yield and photosynthetically active radiation (PAR)). The study obtained strong correlation between

observed and predicted LAI, however they obtained poor correlation between observed and predicted yield. Conversely, they recommended use of satellite imagery with a higher spatial resolution than SPOT 4 in order to obtain higher correlation to the yield. The above studies demonstrate the challenges encountered with the use of traditional multispectral satellite data. The challenge with traditional satellites is that they have a low spectral resolution in which they only acquire four to six bands (Govender *et al.,* 2007). The fewer number of bands are a limiting factor for multispectral sensors as they lack sensitivity to vegetation biophysical characteristics (Govender *et al.*, 2007, Main *et al.*, 2008).

Innovations in remote sensing have produced the "new generation" multispectral sensors such as Sentinel-2 MSI that can overcome challenges encountered with the use of traditional satellite data. Sentinel-2 MSI is produced by the European Space Agency (ESA) for agricultural applications amongst other things (Ramoelo *et al.,* 2015, Martínez and Joel, 2017). Sentinel-2 MSI has an operational system, which produces data specifically for agricultural observations (Sen2 Agri) from Sentinel-2 (S2) L1C and Landsat 8 (L8) L1T time series (European Space Agency, 2015). This Sen2 Agri is useful for vegetation studies as it produces data such as the NDVI, leaf area index (LAI) and biophysical vegetation status indicator (European Space Agency, 2015). It is especially an advantage for third world countries since the data is freely available. The Sentinel-2 MSI is a combination of multispectral and hyperspectral characteristics exhibiting higher spatial, spectral and temporal resolutions (Clevers and Gitelson, 2013, Zheng *et al.*, 2018). In comparison to traditional satellites, Sentinel-2 MSI possesses 13 spectral bands situated in the visible (vis) and shortwave infrared regions (SWIR) (Sibanda *et al.*, 2015, Zheng *et al.*, 2018). Within the 13 spectral bands, there are three red-edge bands (705 nm, 740 nm and 783 nm) that provide useful information for predicting and observation of crop statuses (Zheng *et al.*, 2018).

Sentinel-2 MSI has shown competence in various vegetation and agricultural studies. For example, Zheng *et al.* (2018) used Sentinel-2 multispectral imagery to develop a new spectral index that can be used to detect yellow rust infection in winter wheat. Their study found that B4 (Red), B5 (Red-edge 1) and B7 (red-edge 3) were sensitive in detecting yellow rust in wheat. Sibanda *et al.* (2015), assessed the potential of Sentinel-2 MSI in quantifying above ground biomass for different fertilizer treatments and concluded that Sentinel-2 MSI red-edge (705 and 740 nm) bands have a potential of retrieving canopy chlorophyll and nitrogen content. Similarly, Clevers and Kooistra (2013) assessed the potential of Sentinel-2 MSI in predicting canopy chlorophyll content of potatoes. Their study found similar results to Sibanda *et al.* (2015) concerning the importance of

the red-edge bands (705 and 740 nm) sensitivity to chlorophyll content of potatoes. Ramoelo *et al.* (2015), used Sentinel-2 to assess rangeland quality and documented that the red-edge (705 nm) and the shortwave infrared bands (2190 and 1610 nm) were essential in estimating leaf nitrogen. Based on the mentioned studies, it would be interesting to investigate the ability of Sentinel-2 MSI on estimating soybean grain yield.

To evaluate Sentinel-2 MSI capabilities in predicting soybean grain yield it was important to compare it to other new resolution multispectral sensors such as WorldView-2 and Landsat 8 OLI. WorldView-2 is comparable to Sentinel-2 MSI since it has higher spectral configuration with an addition of the yellow, red-edge bands to the traditional four bands and a higher spatial resolution of 2 m (Huang *et al.*, 2017). WorldView-2 capabilities have been tested on monitoring rice nitrogen status (Huang *et al.*, 2017) and biomass of wetland vegetation (Mutanga *et al.*, 2012). Landsat 8 on the other hand contains 11 spectral bands with an addition of two shortwave (SWIR) and thermal infrared regions (Skakun *et al.*, 2017). Also, Landsat 8 OLI has been used to predict soybean and other crops such as maize (Sayago and Bocco, 2018). Given the similar and different spectral configurations of these multispectral resolutions, it would be interesting to test and compare their capabilities at predicting soybean grain yield.

Multispectral data have limitations because of the lack of higher spectral resolution and as such poses a limitation. Owing to that, other remote sensing data such as hyperspectral data have high resolution spectral data that have been used to assess biochemical characteristics of soybean (Thenkabail *et al.*, 2013). However, the challenge with hyperspectral data is that it is costly to obtain, comes in large volumes and it is expensive and difficult to process (Dye *et al.*, 2011). With the challenges that come with hyperspectral, it is evident that there is a need to utilise affordable data sources that can be also be used to exploit hyperspectral data through spectral resampling. The concept of spectral resampling involves the mapping of hyperspectral bands to the multispectral sensor bands in which the multispectral sensor will attain the higher spectral resolution of the hyperspectral sensor (Chakravortty and Subramaniam, 2014). This research sought to assess the ability of Sentinel-2 MSI to predict soybean grain yield by resampling hyperspectral data to multispectral bands. This objective was achieved by testing the sensitivity of all the bands of Sentinel-2 MSI, Landsat 8 OLI and WorldView-2 to soybean grain yield. The other objectives were; to compare the performance of NDVI and NDVI red-edge computed from Sentinel-2 MSI, Landsat 8 OLI and WorldView-2 in predicting soybean grain yield. To assess the

potential of Sentinel-2 MSI in predicting soybean grain yield, the results were compared to Landsat 8 OLI and WorldView-2.

4.2 Methodology

4.2.1 Study Area

This study was conducted on two experimental farms designed by the Agricultural Research Council, Pedology division as described in chapter 2. The experimental farms were situated in Mpumalanga and Free State provinces of South Africa. The exact locations of these farms are in Ermelo (26° 45'18" S and 30° 13'55" E) and Phuthaditjhaba (28°25'26"S and 28°56'12"E) (**Figure 4.1.**). The climatic conditions of the study sites can be described as cold winters and warm summers. The annual rainfall for the study areas is about an average of 625 mm yearly with most of the rainfall received from October to March (Summer). The soil type in Ermelo is described as low clay and sandy (Sakala *et al.*, 2017) whereas in Phuthaditjhaba it is described as rich loam type of soil (Koatla, 2012).

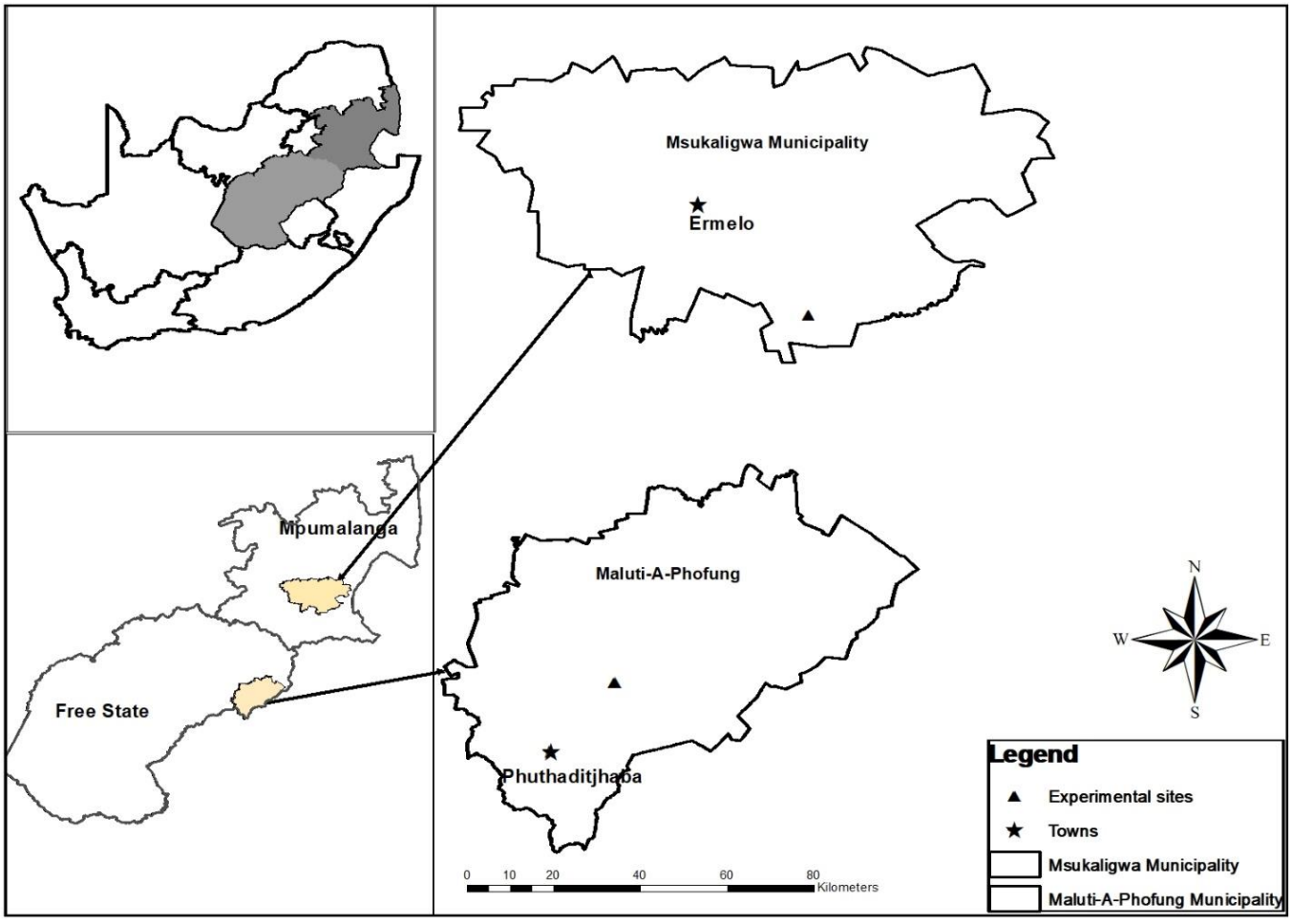


Figure 4.1: Map showing study area locations in South Africa.

4.2.2 Experimental design and setup

In the two study sites, the experiment was designed in a split plot Randomized Complete Block Design (RCBD). More details of the study site are presented in chapter 2. About 72 experimental plots with the size of 7 m length and 3 m were utilised. The experimental plots were comprised of 7 rows with 60 cm spacing in between them. In addition, the experiment had three replicates. In the plots, three soybean varieties from Pannar Seeds (PANN 1500 R, PANN 1614 R and PANN 1664 R) were planted in December 2016 in both study sites. Phosphorus fertiliser treatments of 0 kg, 30 kg and 60 kg were administered. The soybean relied on rainwater for irrigation.

4.2.3 Field canopy measurements and soybean grain yield

Field spectral measurements in the study sites were taken in March and April of 2017. At this time, the soybean biomass had become dense to such an extent that it covered the soil and could not have effect on the spectral reflectance. The spectral measurements were taken randomly plot by plot during the flowering, pod formation, and seed filling stages of the soybean plants. Spectral reflectance was measured using an Analytical Spectral Device (ASD) Field Spec®3 optical sensor (Analytical Spectral Devices, Inc., Boulder, CO, USA). The spectroradiometer measures wavelength starting from 350 to 2500 nm, measuring radiation at 1.4 nm bandwidths for the spectral region of 350-1000 nm and measures 2 nm intervals for the spectral region of 1001-2500 nm (ASD, 2005). The spectral measurements were taken from 10:00 am to 14:00 pm local time (GMT+2) under clear sky and sunny conditions in order to avoid any interference with the reflectance (Salisbury, 1998). For each plot, an optical cable attached to the spectroradiometer was used to take 5 spectral measurements in each plot held about 30 cm directly above the soybean canopy. A white reference calibration panel was used to stabilise any changes in the atmosphere and sun irradiance after every 10 to 15 minutes. The spectral measurements were then added together in order to get the average spectral measurements for every plot.

At the end of the growing season (May and June 2017), the soybean pods were gathered from 3 middle rows of every plot. The soybean pods were placed in a sack and crushed in order to get the soybean grains. Soybean grains were sieved from soybean dry biomass and were weighed utilising the LBK1 weighing scale from ADAM Equipment (Adam Equipment, 2017) to get soybean grain

yield of every plot. The soybean grains measurements of every plots for each study were added to get the total yield of the soybean of the sites.

4.3 Data analysis

4.3.1 Spectral Resampling

Before resampling the spectral data to multispectral resolutions, noise regions (350 to 399 nm, 1350 to 1450 nm, 1800 to 1950 nm and 2400 to 2500 nm) from the hyperspectral data were removed following a method defined by Abdel-Rahman *et al.* (2014) and (Adam *et al.*, 2014). The hyperspectral data was then resampled to Sentinel-2 MSI, Landsat 8 OLI and WorldView-2 spectral configuration shown in **Table 4.1** using R software following a method used by Sibanda *et al.* (2015) and Ramoelo *et al.* (2015). Some spectral bands contained by Sentinel-2 MSI are not all useful in the observation of vegetation (Shoko and Mutanga, 2017). Bands such as 1 (coastal), 9 (water vapour) and 10 (Cirrus) of Sentinel-2 MSI were removed and Landsat 8 OLI and WorldView-2 bands were all used. To assess the ability of these multispectral resolutions to predict soybean grain yield, the resampled spectral data was used in the computation of vegetation indices shown in **Table 4.2.** The selection of the indices was based on their application in prediction of crop yield (Mutanga *et al.*, 2012).

Table 4.1. Spectral description of Sentinel-2 MSI, WorldView-2 and Landsat 8 OLI sensors.

Satellite sensor	Spectral bands	Band centre (nm)	Wavelength range (nm)	Band width (nm)	Spatial Resolution (m)
Sentinel-2 MSI	B1	443	421-457	20	60
	B2	490	439-535	65	10
	B3	560	537-582	35	10
	B4	665	646-685	30	10
	B5	705	694-714	15	20
	B6	740	731-749	15	20
	B7	783	768-796	20	20
	B8	842	767-908	115	10
	B8a	865	849-881	20	20
	B9	945	931-958	20	60

	B10	1375	1338-1349	30	60
	B11	1610	1539-1681	90	20
	B12	2190	2072-2312	180	20
WorldView-2	Coastal	425	400-450	50	1.85
	Blue	480	450-510	60	1.85
	Green	545	510-580	70	1.85
	Yellow	605	585-625	40	1.85
	Red	660	630-690	60	1.85
	Red-Edge	725	705-745	40	1.85
	Near-IR 1	832	770-895	125	1.85
	Near-IR 2	950	860-1040	180	1.85
Landsat-8 OLI	B1	443	435-451	16	30
	B2	482	452-512	60	30
	B3	561	533-590	57	30
	B4	655	636-673	37	30
	B5	865	851-879	28	30
	B6	1609	1566-1651	85	30
	B7	2201	2107-2294	187	30
	B8	590	503-676	172	15
	B9	1373	1363-1384	20	30
	B10	1089	1060-1119	590	100
	B11	12005	1150-1251	1010	100

Table 4.2. Predictor variables utilised in predicting soybean grain yield

Variable name	Satellite	Spectral Bands / Index
Raw bands	Sentinel-2 MSI	b2, b3, b4, b5, b6, b7, b8, b8a, b11 and b12
	Landsat 8 OLI	b1, b2, b3, b4, b5, b6, b7, b10 and b11
	WorldView-2	

		Coastal blue, blue, green, yellow, red, red-edge, NIR1 and NIR2
Vegetation indices	Sentinel-2 MSI Landsat 8 OLI WorldView-2	NDVI, NDVIred-edge NDVI, NDVI, NDVIred-edge

4.4 Statistical Analysis

To assess the ability of Sentinel-2 MSI in predicting soybean grain yield in comparison to Landsat 8 OLI and WorldView-2, the random forest regression algorithm was utilised. The random forest was specifically used to assess the ability of raw multispectral bands and vegetation indices to predict soybean grain yield and to compare the performance of the multispectral sensors. Random forest algorithm is an ensemble of machine learning methods that uses bootstrap aggregation in which groups of trees (*ntree*) are created based on random bootstrap samples selected from the training data set with replacement (Breiman, 2001). The random forest is easy to implement as it requires two parameters i.e. *ntree* and *mtry* to be optimized (Breiman, 2001). Random forest obtains results by averaging the total number of trees to get the overall prediction (Prasad *et al.*, 2006b).

To assess the response of raw bands and vegetation indices in predicting soybean grain yield, the random forest uses a built in technique that is permutation of variables (also known as the mean decrease in accuracy) that can be used to assess importance of variables (Boulesteix *et al.*, 2012). Permutation of variables technique is concerned with the prediction accuracy or predictive ability predictor variables (Boulesteix *et al.*, 2012, Janitza *et al.*, 2015). Important predictor variables are determined by calculating the change in percentage of the root mean squared error (RMSE) when the OOB data are permuted for each variable while other variables remain the same (Janitza *et al.*, 2015). Importance of a predictor variable is determined by large changes that occur when a variable is permuted.

For validation, the random forest used the OOB data (1/3) that was not included in the training of the model (Abdel-Rahman *et al.*, 2013). From the OOB data and the 2/3 data used to train the model, OOB error is computed to test for the difference between predictions (Abdel-Rahman *et al.*, 2013). The OOB error is regarded as a dependable measure of the predictive accuracy since it uses OOB data was not included in the development of the model (Boulesteix *et al.*, 2012). Since

the OOB error acts as a cross-validation measure, researchers note that there is no need for a separate data for model validation. The root mean square error (RMSE) and coefficient of determination (R^2) were used as measures to evaluate the accuracy of the regression models. The random forest regression model was implemented using STATISTICA software (Dell Inc, 2015) using random forest module embedded in the software.

4.5 Results

4.5.1 Quantified soybean grain yield (g/m^2)

Prior to performing any statistical analysis, descriptive statistics of soybean grain yield were produced, and the outcome is shown in **Table 4.3**. The total soybean grain yield of the two experimental sites ranged between 202 g/m^2 and 3374 g/m^2.

Table 4.3: Descriptive statistics of measured soybean grain yield (g/m^2)

No. Of Samples	Min (g/m^2)	Max (g/m^2)	Mean (g/m^2)	Standard deviation (g/m^2)
144	202	3374	1623.2	734.4628

4.5.2 Multispectral bands sensitivity to soybean grain yield

Figure 4.2 shows the performance of single multispectral bands in predicting soybean grain yield. Based on the variable importance produced by the random forest, b4 (red) spectral band was the most important band for Sentinel-2 MSI followed by b5 (red-edge) and b2 (blue) spectral bands. Similarly, for Landsat 8 OLI, the most important bands in predicting soybean grain were the red band, blue and coastal blue bands. Also, for WorldView-2 the red, blue and coastal blue bands were the most important bands in comparison to others.

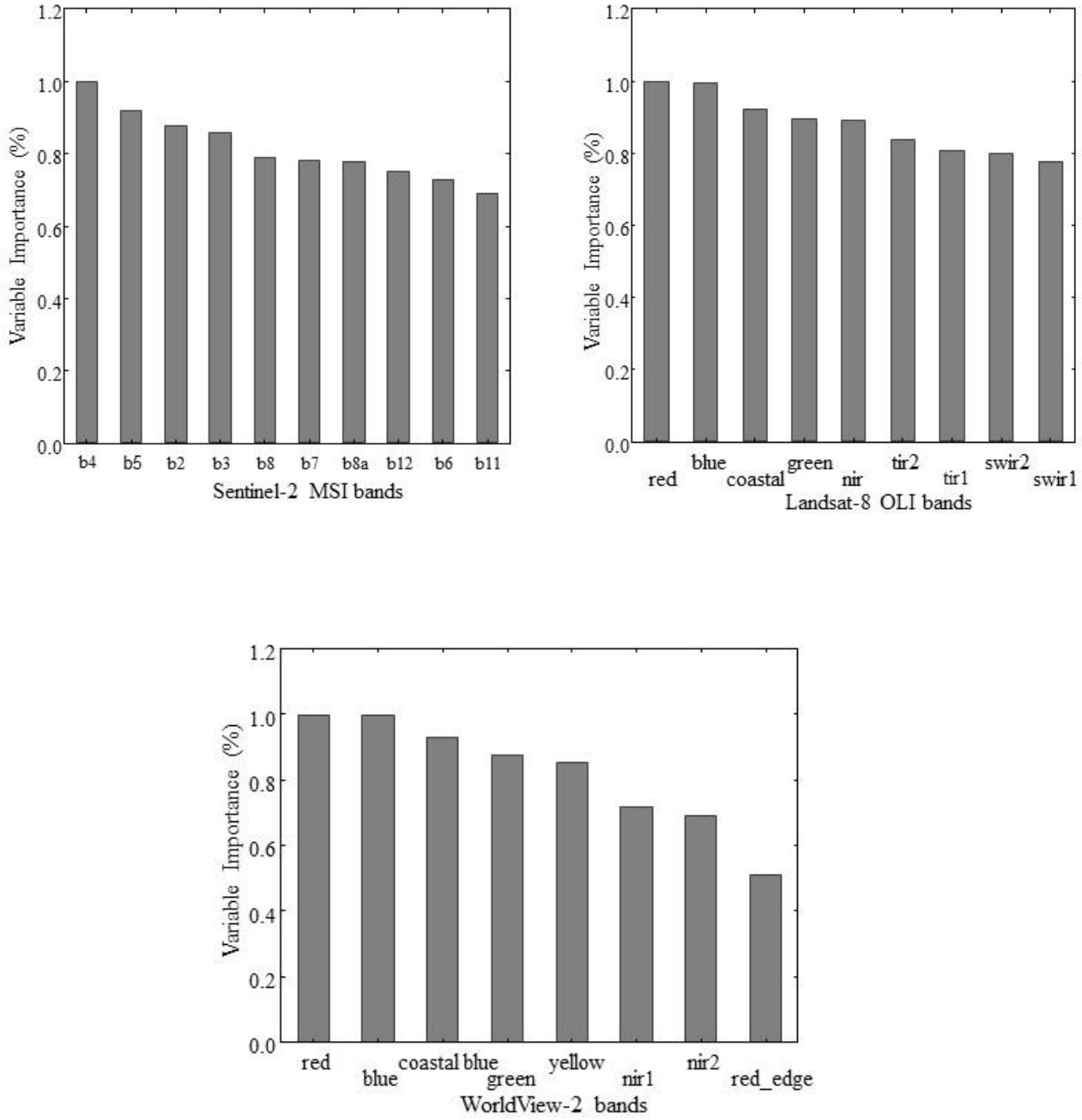


Figure 4.2: Important variables (bands) in predicting soybean grain yield using Sentinel-2 MSI, Landsat 8 OLI and WorldView-2.

4.5.3 Important vegetation indices in predicting soybean grain yield

Results of the variable importance of vegetation indices of Sentinel-2 MSI, Landsat 8 OLI and WorldView-2 in predicting soybean grain yield are shown in **Figure 4.3.** The most important band combination in predicting soybean for Sentinel-2 MSI was the b8 (NIR 843) and b7 (red-edge 783) combination. For Landsat 8 OLI, the NIR and red bands were the most important in predicting

soybean whereas for WorldView-2 the NIR2 and the red band combination showed highest importance.

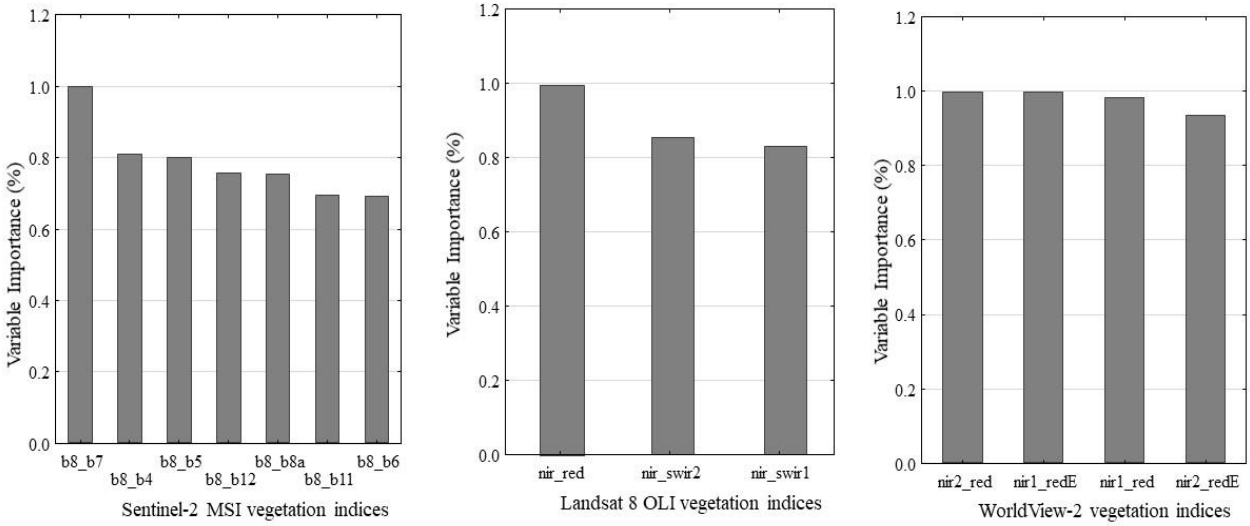


Figure 4.3: Important variables (Indices) in predicting soybean grain yield

4.5.4 Comparison of Sentinel-2 MSI, Landsat 8 OLI and WorldView-2 in predicting soybean grain yield

Figure 4.4 shows the comparison of Sentinel-2 MSI, Landsat 8 OLI and WorldView-2 vegetation indices in predicting soybean grain yield. The results show that the most important vegetation indices in predicting soybean grain yield is a combination of b8 and b7 bands of Sentinel-2 MSI, followed by the NIR and red bands of Landsat 8 OLI, and lastly a combination of bands b8 and b4 of Sentinel-2 MSI. The results also show that WorldView-2 vegetation indices performed poorly compared to Sentinel-2 MSI and Landsat 8 OLI vegetation indices; however, the highest performing vegetation index of WorldView-2 is a combination of NIR1 and the red band. **Table 4.4** illustrates the prediction performance of Sentinel-2 MSI, Landsat 8 OLI and WorldView-2 vegetation indices in predicting soybean grain yield. The highest performing index of Sentinel-2 MSI obtained an R^2= 0.208 with a RMSE of 658.7. Similarly, Landsat 8 OLI optimal index in predicting soybean grain yield yielded an R^2= 0.173 with a RMSE = 633.2. WorldView-2 highest performing index obtained an R^2= 0.172 with a RMSE = 635.8. Although Sentinel-2 MSI performed better compared to other sensors based on the R^2 values, however, the p values indicate that, results obtained for Landsat 8 OLI and WorldView-2 are all significant as they all obtained p = 0.0000. In addition, the RMSE values of these sensors show that Sentinel-2 MSI yielded the highest values compared to Landsat 8 and WorldView-2. **Figure 4.5** shows the one on one relationship between actual and predicted yield using the highest performing vegetation indices.

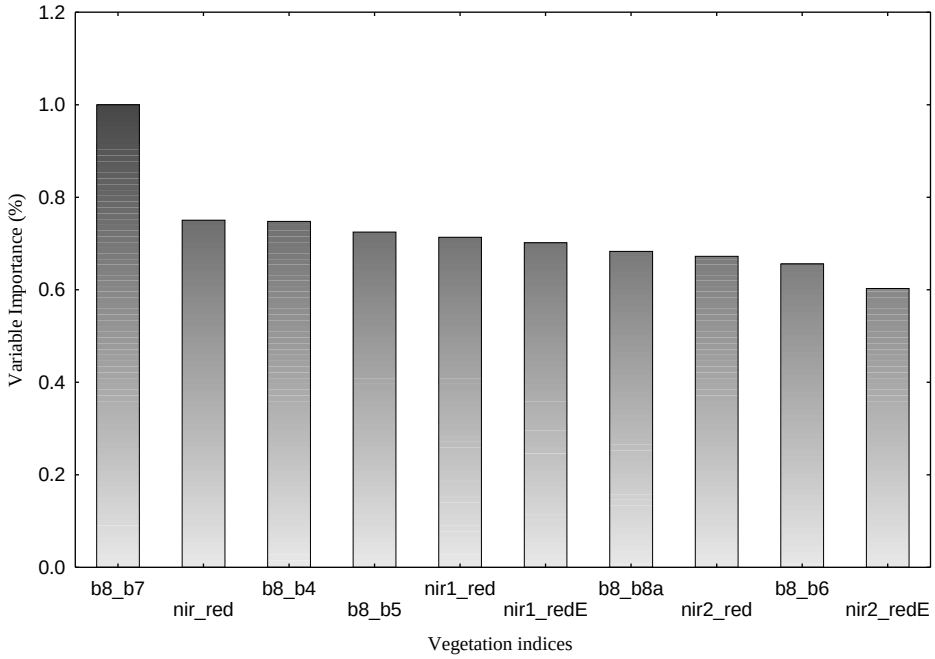


Figure 4.4: Comparison of Sentinel-2 MSI, Landsat 8 OLI and WorldView-2 vegetation indices in predicting soybean grain yield.

Table 4.4: Performance of Sentinel-2 MSI, Landsat 8 OLI and WorldView-2 in predicting soybean grain yield.

Satellite sensor	**R^2**	**RMSE**	**P-value**
Sentinel-2 MSI	0.208	658.7	0.0000
Landsat 8 OLI	0.173	633.2	0.0000
WorldView-2	0.172	635.8	0.0000

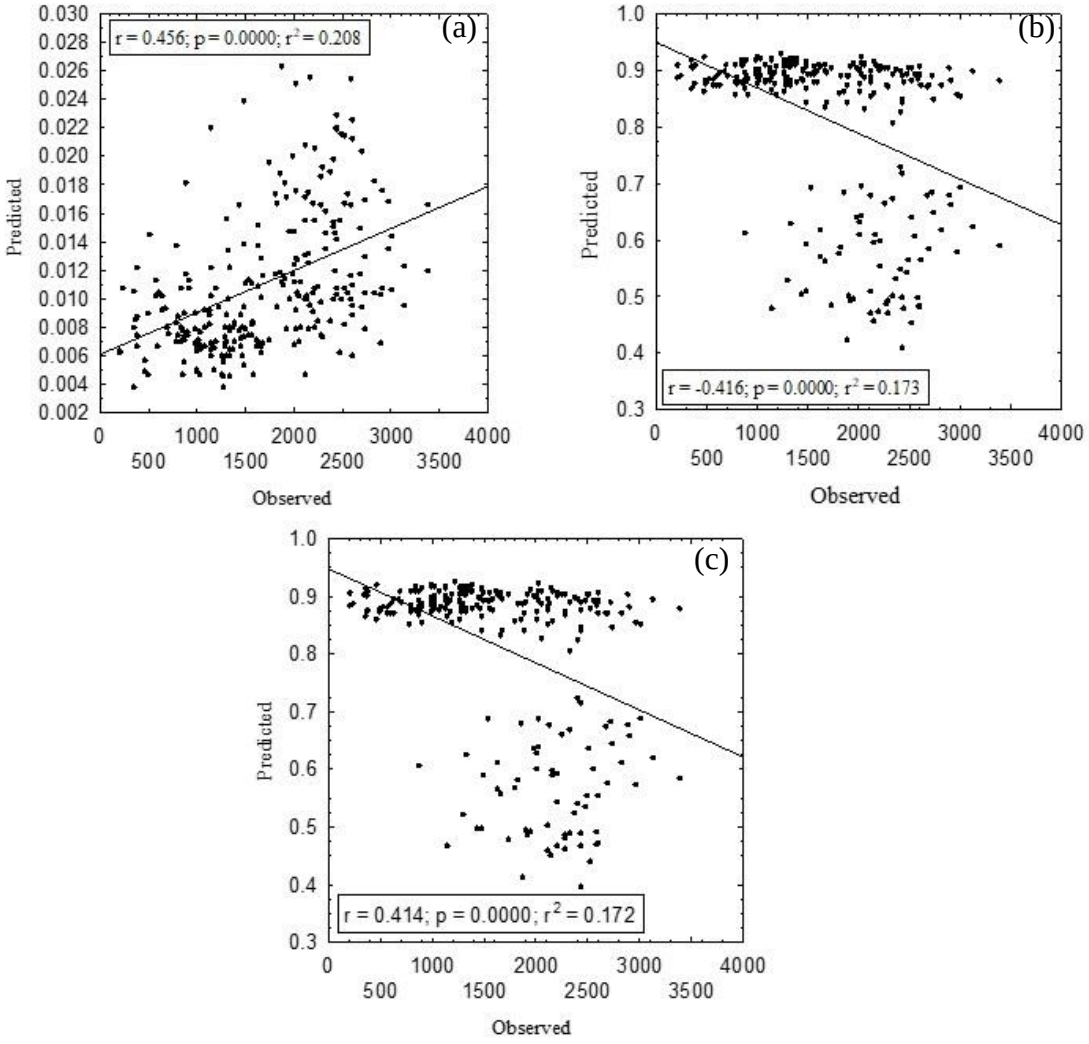


Figure 4.5: One on one relationship between predicted and observed soybean grain yield (a) Sentinel-2 MSI (b8 and b7), (b) Landsat 8 OLI (nir and red) and WorldView-2 (nir1 and red)

4.6 Discussion

Use of accurate, timely and affordable remote sensing data is crucial in the agricultural industry as it can assist in making crop statistics easier to acquire on time. Advancements in the field of remote sensing have produced high-resolution multispectral satellites such as Sentinel-2 MSI to fill the gap between expensive high-resolution hyperspectral data and low-resolution multispectral data. In this study, the aim was to assess the potential of Sentinel-2 MSI to predict soybean grain yield by resampling hyperspectral data. In assessing its potential, Sentinel-2 MSI spectral bands and vegetation indices were tested their sensitivity to soybean grain yield and its performance had to be compared to other multispectral satellites such as Landsat 8 OLI and WorldView-2.

The results in **Figure 4.2** show that highly sensitive bands to soybean grain yield for Sentinel-2 MSI included b4, b5 and b2 (red, red-edge and blue bands respectively). Similarly, for Landsat 8 OLI and WorldView-2 highly sensitive bands included the red, blue and coastal blue bands. The results obtained from these multispectral sensors are similar, however differing with red-edge band from Sentinel-2 MSI and coastal blue band from Landsat 8 OLI and WorldView-2. The selected multispectral bands are bands that are sensitive biophysical properties of crops such as chlorophyll (Kumar *et al.*, 2002). These findings are similar to Immitzer *et al.* (2016) who obtained the red and red-edge bands amongst the important bands in crop type classification using Sentinel-2 MSI data and Ngubane (2014) who identified the coastal blue band amongst important bands in classifying Braken fern. It is comprehensible that the red, blue and coastal blue bands were highly sensitive to the soybean grain yield. This is because crop/plants absorb the red and blue radiation for photosynthesis (Kumar *et al.*, 2002). Similarly, the coastal blue band is absorbed by chlorophyll in healthy crops (White Paper, 2010, Ngubane, 2014). In addition, the red-edge (b5 (705 nm)) has been reported to have significance in chlorophyll estimation in other studies (Clevers and Kooistra, 2013, Ramoelo *et al.*, 2015). It is interesting to note that, although Sentinel-2 MSI contains four red-edge bands (b5, b6, b7, b8a); however, only one of these bands was highly sensitive to soybean grain yield. The importance of the red-edge band (b5 (705 nm)) may have been due to the influence of the high absorption of the red band (b4 (665 nm)) since it is situated at the transition position from the red band to the NIR band (Mutanga and Skidmore, 2004). Other red-edge bands get more influence from the high reflectance of the NIR region as they are closer to the NIR in position (Mutanga and Skidmore, 2004).

Similarly, **Figure 4.3** shows that important vegetation indices for predicting soybean grain yield for Sentinel-2 MSI were a combination of b8 (NIR (843 nm)) and b7 (red-edge (783 nm)), for Landsat 8 OLI, NIR and red bands while for WorldView-2 it was a combination of NIR2 and red bands. The high performance of Sentinel-2 MSI can be associated with the involvement of the red-edge band as opposed to the red band as is the case with Landsat 8 OLI and WorldView-2. This finding is confirms the findings of Adelabu (2013) that the importance of the red-edge band comes about when it is combined with normal bands. **Figure 4.5 and Table 4.4** show the prediction performance of Sentinel-2 MSI, Landsat 8 OLI and WorldView-2 multispectral sensors produced by the random forest regression method. Prediction results show that all the models derived from these sensors can significantly predict soybean as is shown by the p-values that are less than 0.05. However, when comparing the models, the results showed that Sentinel-2 MSI (R^2 =0.208)

performed better that Landsat 8 OLI (R^2 =0.173) and WorldView-2 (R^2 =0.172). Better performance of Sentinel-2 MSI (p = 0.000) can be attributed to inclusion of the red-edge band (b7 (783)). The red-edge band contains a combination of attributes of the red and the NIR bands and therefore makes more sensitive to chlorophyll content. These results are similar to those of Sharma *et al.* (2015) and (Pathak (2016)) who discovered that the red-edge NDVI predicted crop yield better than the red NDVI. Although Sentinel-2 MSI predicted soybean yield better than other sensors, however, the RMSE shows that Sentinel-2 MSI yielded the highest RMSE. This means that Sentinel-2 MSI has higher prediction error than Landsat 8 and WorldView-2.

4.7 Conclusion

The present study aim was to assess the potential of the new generation Sentinel-2 MSI sensor in predicting soybean grain yield. The findings showed that important bands in predicting soybean grain yield included the red, red-edge and blue bands for Sentinel-2 MSI while for Landsat 8 OLI and WorldView-2 included the red, blue and coastal blue bands. Sentinel-2 MSI derived indices predicted soybean grain better than Landsat 8 OLI and WorldView-2 vegetation indices. The better performance of Sentinel-2 MSI can be associated with the involvement of the red-edge band. In overall, the study showed that Sentinel-2 MSI has the potential to predict soybean grain yield. Based on the results, it can be concluded that Sentinel-2 MSI could be useful in the constant monitoring and prediction of yield. However, there is a need for further research to be conducted to test the performance of Sentinel-2 MSI at an imagery level.

References

Abate, T., Alene, A. D., Bergvinson, D., Shiferaw, B., Silim, S., Orr, A. & Asfaw, S. 2012. Tropical Grain Legumes in Africa and South Asia. *Knowledge and Opportunities*.

Abdel-Rahman, E. M., Ahmed, F. B. & Ismail, R. 2013. Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data. *International Journal of Remote Sensing,* 34**,** 712-728.

Abdel-Rahman, E. M., Mutanga, O., Odindi, J., Adam, E., Odindo, A. & Ismail, R. 2014. A comparison of partial least squares (PLS) and sparse PLS regressions for predicting yield of Swiss chard grown under different irrigation water sources using hyperspectral data. *Computers and Electronics in Agriculture,* 106**,** 11-19.

Adam, E., Mutanga, O., Abdel-Rahman, E. M. & Ismail, R. 2014. Estimating standing biomass in papyrus (Cyperus papyrus L.) swamp: exploratory of in situ hyperspectral indices and random forest regression. *International Journal of Remote Sensing,* 35**,** 693-714.

Adam Equipment. 2017. *Adam Equipment Prodcucts - Weighing Scales and Equipment Manufacturer* [Online]. Available: https://www.adamequipment.co.za/products.

Adelabu, S. 2013. *The Remote Sensing of Insect Defoliation in Mopane Woodland.* Phd, University of KwaZulu-Natal.

Asd, A. S. D. 2005. Handheld spectroradiometer: user guide version 4.05. *Boulder: Analytical Spectral Devices Inc*.

Bappel, E., Bégué, A., Martiné, J.-F., Pellegrino, A. & Siegmund, B. Assimilation of a biophysical parameter estimated by remote sensing using SPOT 4 and 5 data into a sugarcane yield forecasting model. Proc. ISSCT, 2005.

Board, J. E. & Kahlon, C. S. 2011. Soybean yield formation: what controls it and how it can be improved. *Soybean physiology and biochemistry.* InTech.

Boulesteix, A. L., Janitza, S., Kruppa, J. & König, I. R. 2012. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 2**,** 493-507.

Breiman, L. 2001. Random forests. *Machine learning,* 45**,** 5-32.

Chakravortty, S. & Subramaniam, P. 2014. Fusion of Hyperspectral and Multispectral Image Data for Enhancement of Spectral and Spatial Resolution. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences,* 40**,** 1099.

Cicek, H., Sunohara, M., Wilkes, G., Mcnairn, H., Pick, F., Topp, E. & Lapen, D. 2010. Using vegetation indices from satellite remote sensing to assess corn and soybean response to controlled tile drainage. *Agricultural Water Management,* 98**,** 261-270.

Clevers, J. & Kooistra, L. Retrieving canopy chlorophyll content of potato crops using Sentinel-2 bands. Proceedings ESA Living Planet Symposium, Edinburgh, United Kingdom, 09-13 September, 2013, 2013. 8-8.

Clevers, J. G. & Gitelson, A. A. 2013. Remote estimation of crop and grass chlorophyll and nitrogen content using red-edge bands on Sentinel-2 and-3. *International Journal of Applied Earth Observation and Geoinformation,* 23**,** 344-351.

Dell Inc 2015. Dell Statistica (data analysis software system). Version 13 ed.

Dye, M., Mutanga, O. & Ismail, R. 2011. Examining the utility of random forest and AISA Eagle hyperspectral image data to predict Pinus patula age in KwaZulu-Natal, South Africa. *Geocarto International,* 26**,** 275-289.

European Space Agency, E. 2015. SENTINEL-2 User HandBook.

Fao 2016. Crop Yield Forecasting: Methodological and Institutional Aspects. Food and Agriculture Organiszation of the United Nations Rome.

Govender, M., Chetty, K. & Bulcock, H. 2007. A review of hyperspectral remote sensing and its application in vegetation and water resource studies. *Water Sa,* 33**,** 145-151.

Huang, S., Miao, Y., Yuan, F., Gnyp, M. L., Yao, Y., Cao, Q., Wang, H., Lenz-Wiedemann, V. I. & Bareth, G. 2017. Potential of RapidEye and WorldView-2 satellite data for improving rice nitrogen status monitoring at different growth stages. *Remote Sensing,* 9**,** 227.

Immitzer, M., Vuolo, F. & Atzberger, C. 2016. First experience with Sentinel-2 data for crop and tree species classifications in central Europe. *Remote Sensing,* 8**,** 166.

Janitza, S., Celik, E. & Boulesteix, A.-L. 2015. A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification***,** 1-31.

Koatla, T. a. B. 2012. *Mainstreaming small-scale farmers in Qwaqwa, Free State Province, South Africa.* University of the Free State.

Kumar, L., Schmidt, K., Dury, S. & Skidmore, A. 2002. Imaging spectrometry and vegetation science. *Imaging spectrometry.* Springer.

Locke, C., Carbone, G., Filippi, A., Sadler, E., Gerwig, B. & Evans, D. Using remote sensing and modeling to measure crop biophysical variability. 5th International Conference on Precision Agriculture, 2000.

Lokupitiya, E., Lefsky, M. & Paustian, K. 2010. Use of AVHRR NDVI time series and ground-based surveys for estimating county-level crop biomass. *International Journal of Remote Sensing,* 31**,** 141-158.

Main, R., Cho, M. A., Van Aardt, J. & Majeke, B. 2008. Comparison between sensors with different spectral resolutions, relative to the sumbandila satellite, for assessing site quality differences, in eucalyptus grandis plantations.

Martínez, M. & Joel, L. 2017. Relationship between crop nutritional status, spectral measurements and Sentinel 2 images. *Agronomía Colombiana,* 35**,** 205-215.

Mutanga, O., Adam, E. & Cho, M. A. 2012. High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *International Journal of Applied Earth Observation and Geoinformation,* 18**,** 399-406.

Mutanga, O. & Skidmore, A. K. 2004. Narrow band vegetation indices overcome the saturation problem in biomass estimation. *International Journal of Remote Sensing,* 25**,** 3999-4014.

Ngie, A. & Ahmed, F. 2018. Estimation of Maize grain yield using multispectral satellite data sets (SPOT 5) and the random forest algorithm. *South African Journal of Geomatics,* 7**,** 11-30.

Ngubane, Z. C. 2014. *Evaluating the Potential of WorldView-2's Strategically Located Bands in Mapping the Bracken Fern (Pteridium Aquilinum (L.) Kuhn).* Citeseer.

Noureldin, N., Aboelghar, M., Saudy, H. & Ali, A. 2013. Rice yield forecasting models using satellite imagery in Egypt. *The Egyptian Journal of Remote Sensing and Space Science,* 16**,** 125-131.

Pannar 2006. Soybeans Production Guide. Greytown, South Africa: Pannar Seed (Pty) Ltd.

Pathak, R. 2016. *Use of Digital Imagery to Evaluate the Relationship Between NDVI and Crop Production Field Data at Stutsman County, North Dakota.* North Dakota State University.

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

Prasad, A. M., Iverson, L. R. & Liaw, A. 2006b. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems,* 9**,** 181-199.

Ramoelo, A., Cho, M., Mathieu, R. & Skidmore, A. K. 2015. Potential of Sentinel-2 spectral configuration to assess rangeland quality. *Journal of Applied Remote Sensing,* 9**,** 094096.

Sakala, E., Fourie, F., Gomo, M. & Coetzee, H. 2017. Hydrogeological investigation of the Witbank, Ermelo and Highveld Coalfields: Implications for the subsurface transport and attenuation of acid mine drainage.

Salisbury, J. W. 1998. Spectral measurements field guide. EARTH SATELLITE CORP CHEVY CHASE MD.

Sayago, S. & Bocco, M. 2018. Crop yield estimation using satellite images: comparison of linear and non-linear models. *AgriScientia,* 1**,** 1-9.

Sharma, L. K., Bu, H., Denton, A. & Franzen, D. W. 2015. Active-optical sensors using red NDVI compared to red edge NDVI for prediction of corn grain yield in North Dakota, USA. *Sensors,* 15**,** 27832-27853.

Shoko, C. & Mutanga, O. 2017. Seasonal discrimination of C3 and C4 grasses functional types: An evaluation of the prospects of varying spectral configurations of new generation sensors. *International Journal of Applied Earth Observation and Geoinformation,* 62**,** 47-55.

Sibanda, M., Mutanga, O. & Rouget, M. 2015. Examining the potential of Sentinel-2 MSI spectral resolution in quantifying above ground biomass across different fertilizer treatments. *ISPRS Journal of Photogrammetry and Remote Sensing,* 110**,** 55-65.

Skakun, S., Vermote, E., Roger, J.-C. & Franch, B. 2017. Combined use of Landsat-8 and Sentinel-2A images for winter crop mapping and winter wheat yield assessment at regional scale. *AIMS geosciences,* 3**,** 163.

Smit, M. A. 2000. *Your Guide to successful soybean production,* Potchefstroom, Agricultural Research Council Crop Institute.

Thenkabail, P. S., Mariotto, I., Gumma, M. K., Middleton, E. M., Landis, D. R. & Huemmrich, K. F. 2013. Selection of hyperspectral narrowbands (HNBs) and composition of hyperspectral twoband vegetation indices (HVIs) for biophysical characterization and discrimination of crop types using field reflectance and Hyperion/EO-1 data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* 6**,** 427-439.

White Paper 2010. THE BENEFITS OF THE EIGHT SPECTRAL BANDS OF WORLDVIEW-2.

Zheng, Q., Huang, W., Cui, X., Shi, Y. & Liu, L. 2018. New Spectral Index for Detecting Wheat Yellow Rust Using Sentinel-2 Multispectral Imagery. *Sensors,* 18**,** 868.

Chapter 5

Research Synthesis: A review of objectives and conclusions

5.1 Introduction

Accurate yield predictions play a vital role in guiding market prices of agricultural products, enhance the management of food prior and after harvest, imports and exports and trade between countries, (Noureldin *et al.*, 2013). Thus, acquisition of crop statistics through accurate and reliable methods is crucial for farmers, policy-makers, and government. The accuracy of crop yield predictions is affected by subjective and at times inaccurate crop statistics generated through manual field survey methods. This is because manual field survey methods are labour intensive, costly to carry out and take a lot of time to execute (Adam, 2010, Dube, 2015). Consequently, this indicates the requirement for accurate, objective, cost effective and reliable techniques that can be used to capture crop statistics to derive yield information. Remote sensing is seen as the most objective method that can continuously provide real-time crop information, which can be used to derive precise yield predictions. Multispectral remote sensing data have been widely adopted for predicting biomass and crop yield. However, the challenge is that multispectral data have a propensity to saturate when predicting vegetation with high biomass (Mutanga and Skidmore, 2004, Mutanga *et al.*, 2012). In that case, researchers have promoted the use of high-resolution hyperspectral data (Mutanga and Skidmore, 2004). This is because hyperspectral data provides narrow, contiguous spectral bands that can subdue challenges of multispectral data and can accurately predict crop yield (Thenkabail *et al.*, 2000). Specifically, this study was interesting to carry out since few studies have used hyperspectral data to predict soybean grain yield. At the end, the findings of this study showed the rich information that is contained by hyperspectral data can be used to predict soybean grain yield. Similarly, the soybean undergoes growth and development over two stages with different phases. During this time, the behaviour of the soybean plant in relation to radiation is different. Due to this, it was important to investigate as to which growth stage is most suitable to predict soybean grain yield. Remote sensing provides different sensors that can provide useful data for crop estimation. For this study, it was necessary to test multispectral sensors such as Sentinel-2 MSI and Landsat 8 OLI that are freely available as they will be useful in South Africa due to limited resources. As for WorldView-2, its use in this study was to evaluate the performance of Sentinel-2 MSI when compared to a sensor that has similar spectral configuration. Chapter 2 compared the highly recommended NDVI, SR and EVI narrow-

band indices (Mutanga and Skidmore, 2004, Adam *et al.*, 2014) derived from hyperspectral data in predicting soybean grain yield. Findings in that chapter showed better performance of the narrow-band SR and NDVI vegetation indices compared to EVI vegetation indices in predicting soybean grain yield. These indices were of band combinations situated in the visible, NIR, red-edge and the middle infrared regions. After determining the potential of these narrow-band indices in predicting soybean grain yield, it was crucial to determine the suitable growth stage to predict soybean in chapter 3 using the best indices (NDVI and SR) as revealed in chapter 2. Results in that chapter showed that the flowering stage is most suitable to predict soybean grain yield. Similarly, in chapter 4, the SR and NDVI were computed from Sentinel-2 MSI, Landsat 8 OLI and WorldView-2 multispectral data that was resampled from hyperspectral data. The interest in executing chapter 4 was brought about the costs that are associated with acquisition and processing that is associated with hyperspectral data (Mutanga *et al.*, 2012). This is especially important because Sentinel-2 MSI and Landsat 8 OLI are open source systems which provide NDVI and LAI data that is freely available that can benefit resource-limited countries such as South Africa. Findings in that study showed that Sentinel-2 MSI better predicted soybean compared to Landsat 8 OLI and WorldView-2. In overall, this study explored the utility of remote sensing data in predicting soybean grain yield.

For this chapter, the aim and objectives founded in the introduction section are reviewed in contrast to the findings.

5.2 Objectives reviewed

5.2.1 Evaluating the potential of narrow-band indices to predict soybean (*Glycine max (L.) Merr*) grain yield

Vegetation indices derived from multispectral sensors proved to have saturation problems when predicting vegetation and yield of crops with high biomass. In chapter 2, narrow-band indices i.e. NDVI, SR and EVI were calculated from hyperspectral data and compared in predicting soybean grain yield. Precisely, the study examined the relationship between derived narrow-band indices to soybean grain yield and compared the performance of narrow-band indices in predicting soybean grain yield using the random forest regression algorithm. The findings in chapter 2 showed that narrow band indices have the potential to predict soybean grain yield. This was shown by the good performance of the SR ($R^2 = 0.843$) ($p = 0.000003$) and the NDVI ($R^2 = 0.841$) ($p= 0.000003$) compared to EVI ($R^2 = 0.578$) ($p = 0.007$) in predicting soybean. These results are

consistent with those obtained by Mutanga and Skidmore (2004) who obtained that SR better predicted biomass compared to NDVI and TVI. Better performance of SR can be attributed to its high sensitivity to high biomass (Bannari *et al.*, 1995). However, these results differ from those obtained by Locke *et al.* (2000) who used SPOT 4 imagery to assess soybean yield and LAI. Their study suggested that NDVI performed better than SR, SAVI and TSAVI. The differences in results for these studies could be based on the different datasets from which the vegetation indices were derived from. For this study, the reason for less performance of NDVI to SR could be the effects of environmental factors that could have affected the spectral reflectance. Although the SR better performed NDVI, the difference is insignificant, which might entail that NDVI is also capable of predicting soybean grain yield. In addition, in terms of accuracy, the NDVI had better accuracy than SR did, however; the p-values indicate that both these results are significant. This means that both NDVI and SR can be used to predict soybean grain yield. As for EVI, the findings of this study were inconsistent with those obtained by Adam *et al.* (2014) who obtained that EVI performed better than NDVI. The inconsistency with those results may be based on the differences in the type of vegetation that were predicted in both studies. This study also confirmed the findings of other studies that indicated that it is not only the red and NIR bands that contain useful information in predicting biomass and crop yield (Mutanga and Skidmore, 2004, Cho *et al.*, 2007, Adam *et al.*, 2014). For soybean, the study demonstrated through high correlations that were observed that combinations of spectral bands located in the blue (405 nm - 475 nm), red region (695 nm), red-edge (705 nm -725 nm), NIR (1245 nm) and the MIR regions (1325 nm – 2397 nm) can provide useful information in estimating soybean grain yield. In conclusion, the results in this chapter provide insight that information contained in the above mentioned regions is useful in predicting soybean grain yield. Additionally, based on the performance of the SR, it suggests that SR can be used to predict soybean grain yield. These findings demonstrate that narrow-band indices have the potential to predict soybean grain yield.

5.2.2 Determining the suitable growth stage to predict soybean (*Glycine max (L.) Merr*) grain yield using hyperspectral data

Raun *et al.* (2005), in their study highlighted the importance of timing for collecting remote sensing data as a key aspect in assessing prospective crop yield. Correspondingly, research has shown that spectral measurements can be related to crop yield because crops reveal their growth, development, stress and yield potential through their canopy status (Wiegand *et al.*, 1986). This study sought to determine the optimal growth stage to estimate soybean grain yield using hyperspectral data.

NDVI and SR were used for this study since they obtained good results in predicting soybean grain yield in chapter 2, hence, that motivated for their use in determining the optimal growth stage to predict soybean yield. For this study, both NDVI ($R^2 = 0.863$) ($p = 0.0000$) and SR ($R^2 = 0.865$) ($p = 0.0000$) showed that the optimal growth stage to predict soybean grain yield is during the flowering stage. This is because during the flowering stage, the soybean plant absorbs high quantities of radiation for photosynthesis. These results are similar to those obtained by Fernandez-Ordoñez and Soria-Ruiz (2017) who concluded that the flowering stage was best to predict maize grain yield. The comparisons between these studies can be associated with the positive correlation between grains and radiation interception during growth and development (Andrade, 1995). However, these findings contradict the findings of Christenson *et al.* (2016) who in their study concluded that no growth stage provided significant information for predicting soybean yield. The differences in results of these studies could have resulted from the differences in the number of growing seasons in which the spectral reflectance was collected. For the current study, spectral reflectance was collected over one seasons from flowering, pod formation and seed filling stages. Whereas Christenson *et al.* (2016) in their study collected their spectral reflectance over two seasons from the pod formation stage (R3) in 2011 and from flowering stage (R2) in 2012. The differences in results of these studies indicate that there is need to further this study and acquire spectral data for another season in order to be able to fully compare with other studies and have a conclusive conclusion. In overall, the findings of this study provided the understanding of how the soybean behaves during the growth stages in relation to its ability to absorb radiation. However, there is room for similar studies to be conducted.

5.2.3 Assessing the ability of Sentinel-2 Multispectral Instrument (MSI) to estimate soybean (*Glycine max (L.) Merr.*) Grain yield from resampled hyperspectral data

High cost linked to the acquisition and processing of hyperspectral data resulted in the evaluating high-resolution multispectral datasets that can be used to predict soybean grain yield. Specifically, these high-resolution multispectral datasets can be advantageous to resource-restricted countries such as South Africa. This study aimed at evaluating the potential of Sentinel-2 MSI in predicting soybean grain yield by resampling hyperspectral data. Sentinel-2 MSI performance was compared to Landsat 8 OLI and WorldView-2. Specifically, the study tested the sensitivity of spectral bands and vegetation indices derived from the aforementioned sensors. Findings proved that Sentinel-2 MSI has the potential to predict soybean grain yield. These findings were validated by the significant results that were obtained (Sentinel-2 MSI; $R^2 = 0.208$, $p = 0.0000$) in predicting

soybean grain yield. This study also indicated that Sentinel-2 MSI (R^2 = 0.208) performed better than WorldView-2 (R^2 = 0.172) and Landsat 8 OLI (R^2 = 0.173) with sensitive bands including the red, blue, red-edge and coastal blue bands. These findings are similar to those obtained by (Al-Gaadi *et al.* (2016)) who obtained that Sentinel-2 MSI performed better than Landsat 8 OLI in predicting potato yield. Results obtained suggest that Sentinel-2 MSI could be potentially used in prediction yield. Although the findings of these studies are similar, however, they obtained higher prediction values than the current study. The differences could be because of simulated hyperspectral data whereas the other study used actual Sentinel-2 MSI imagery. This then shows a gap for future research to find out if better results can be obtained if soybean grain yield was predicted from Sentinel-2 imagery.

5.3 Conclusion

The aim of this study was to examine the use of hyperspectral remote sensing data in predicting soybean grain yield. The findings of this study confirmed the ability of hyperspectral data in predicting soybean grain yield. This conclusion is established upon the previously mentioned objectives of the study. Firstly, when narrow-band indices were used to predict soybean grain yield, the SR and NDVI accounted for higher accuracy than EVI. Similarly, the results displayed that spectral bands in the red-edge and MIR also provide useful information in predicting soybean grain yield other than the red and NIR bands. Secondly, the findings of this study showed that the flowering stage was suitable to predict soybean grain yield than the pod and seed filling stages. Lastly, the study also demonstrated the ability of new generation Sentinel-2 multispectral sensors in predicting soybean grain yield. In overall, the study has demonstrated the use of remote sensing data in predicting soybean grain yield. The findings in this study indicate that remote sensing technology can play a significant role in monitoring crop conditions and in crop estimations.

5.4 Recommendations

- This study in chapter 2 showed that other spectral bands other than NIR and red bands also are good at predicting soybean grain yield. Hence, it would be good to exploit the information found in these other regions such as the red-edge and the middle infrared regions from hyperspectral data or multispectral datasets that contain these spectral bands.
- This study tried to determine the optimal growth stage to predict soybean grain yield from spectral data collected over one season. The study produced promising results; however, it

would be good to conduct similar studies using spectral data acquired over a period more than one season in order to verify the conclusiveness of the results obtained in this study.

- Hyperspectral data resampled to Sentinel-2 MSI showed great potential in predicting soybean grain yield. However, this objective was carried out from simulated data; therefore, there is still need to upscale this study.

References

Adam, E. 2010. *The remote sensing of Papyrus vegetation (Cyperus papyrus L.) in swamp wetlands of South Africa.* Doctor of Philosophy in Environmental Sciences, University of KwaZulu-Natal.

Adam, E., Mutanga, O., Abdel-Rahman, E. M. & Ismail, R. 2014. Estimating standing biomass in papyrus (Cyperus papyrus L.) swamp: exploratory of in situ hyperspectral indices and random forest regression. *International Journal of Remote Sensing,* 35**,** 693-714.

Al-Gaadi, K. A., Hassaballa, A. A., Tola, E., Kayad, A. G., Madugundu, R., Alblewi, B. & Assiri, F. 2016. Prediction of potato crop yield using precision agriculture techniques. *PloS one,* 11**,** e0162219.

Andrade, F. H. 1995. Analysis of growth and yield of maize, sunflower and soybean grown at Balcarce, Argentina. *Field Crops Research,* 41**,** 1-12.

Bannari, A., Morin, D., Bonn, F. & Huete, A. 1995. A review of vegetation indices. *Remote sensing reviews,* 13**,** 95-120.

Cho, M. A., Skidmore, A., Corsi, F., Van Wieren, S. E. & Sobhan, I. 2007. Estimation of green grass/herb biomass from airborne hyperspectral imagery using spectral indices and partial least squares regression. *International Journal of Applied Earth Observation and Geoinformation,* 9**,** 414-424.

Christenson, B. S., Schapaugh, W. T., An, N., Price, K. P., Prasad, V. & Fritz, A. K. 2016. Predicting Soybean Relative Maturity and Seed Yield Using Canopy Reflectance. *Crop Science,* 56**,** 625-643.

Dube, T. 2015. *Optical remote sensing of aboveground forest biomass and carbon stocks in resource-constrained African environments.*

Fernandez-Ordoñez, Y. M. & Soria-Ruiz, J. Maize crop yield estimation with remote sensing and empirical models. 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2017. IEEE, 3035-3038.

Locke, C., Carbone, G., Filippi, A., Sadler, E., Gerwig, B. & Evans, D. Using remote sensing and modeling to measure crop biophysical variability. 5th International Conference on Precision Agriculture, 2000.

Mutanga, O., Adam, E. & Cho, M. A. 2012. High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *International Journal of Applied Earth Observation and Geoinformation,* 18**,** 399-406.

Mutanga, O. & Skidmore, A. K. 2004. Narrow band vegetation indices overcome the saturation problem in biomass estimation. *International Journal of Remote Sensing,* 25**,** 3999-4014.

Noureldin, N., Aboelghar, M., Saudy, H. & Ali, A. 2013. Rice yield forecasting models using satellite imagery in Egypt. *The Egyptian Journal of Remote Sensing and Space Science,* 16**,** 125-131.

Raun, W., Solie, J., Martin, K., Freeman, K., Stone, M., Johnson, G. & Mullen, R. 2005. Growth stage, development, and spatial variability in corn evaluated using optical sensor readings. *Journal of plant nutrition,* 28**,** 173-182.

Thenkabail, P. S., Smith, R. B. & De Pauw, E. 2000. Hyperspectral vegetation indices and their relationships with agricultural crop characteristics. *Remote sensing of Environment,* 71**,** 158-182.

Wiegand, C. L., Richardson, A. J., Jackson, R. D., Pinter, P. J., Aase, J. K., Smika, D. E., Lautenschlager, L. F. & Mcmurtrey, J. 1986. Development of agrometeorological crop model inputs from remotely sensed information. *IEEE transactions on geoscience and remote sensing,* 24**,** 90-98.