Modelling leaf area index in a tropical grassland using multi-temporal hyperspectral data

By

Zolo Zime Zinu Serge Kiala

213569675

A thesis submitted in fulfilment for the degree of Master of Science in Environmental Science in the School of Agricultural, Earth and Environmental Sciences, University of KwaZulu-Natal

Pietermaritzburg

Supervisor: Prof. O. Mutanga

Co-supervisor: Dr J.Odindi

March 2016

Abstract

Leaf area index (LAI) is a critical parameter in assessing vegetation status and health of tropical grasslands. Synoptic and dynamic LAI estimates are particularly useful in monitoring changes in ground biomass, hence a basis for sustainable rangeland stewardship. Due to the huge information they provide, hyperspectral remotely sensed data in concert with multivariate regression techniques offer unique opportunities to accurately model LAI in tropical grasslands. This study was a two-step process. Firstly, interval partial least square regression (iPLSR) in forward mode was compared to partial least square regression (PLSR) in estimating LAI using in-situ canopy hyperspectral data at three sampling periods (onset, mid, end) in summer. iPLSR, which is a variant of PLSR, was implemented to reduce all available wavebands used in PLSR to 40 optimal wavebands. Then, optimal bands selected by iPLSR were used to compare PLSR and support vector regression (SVR). The performance of the three regression techniques was determined using root mean square error (RMSE) and coefficients of determination (R^2) based on the predicted and the measured variables. Results show that iPLSR outperformed PLSR for all the sampling periods. iPLSR models could explain LAI variation with R_p^2 values ranging from 0.809 to 0.933 and low RMSEP values from 0.211 to 0.603 m^2m^{-2}, while PLSR models yielded R_p^2 and RMSEP values ranging from 0.364 to 0.649 and from 0.542 to 0.694 m^2m^{-2}, respectively. The best periods for estimating LAI were at beginning and end of summer (R^2_p = 0.882 and RMSEP = 0.299 m^2m^{-2}; R^2_p = 0.890 and RMSEP = 0.211 m^2m^{-2} respectively). Pooling data sets from the three assessed periods yielded the highest prediction error (RMSEP=0.603). PLSR outperformed SVR at the beginning and end of summer in generating optimal wavebands. PLSR models could explain 86.5 % and 85.1 % in LAI variance with RMSEP values of 0.263 m^2m^{-2} and 0.204 m^2m^{-2}, respectively. The SVR models could explain 85.8 % and 83.2 % in LAI variance with RMSEP values of 0.287 m^2m^{-2} and 0.218 m^2m^{-2}, respectively. However, at mid-summer, SVR models yielded higher accuracies (R_p^2 = 0.902 and RMSEP= 0.371 m^2m^{-2}) than PLSR models (R_p^2 = 0.886 and RMSEP = 0.379 m^2m^{-2}). Similarly, for pooled dataset, SVR models were slightly more accurate (R_p^2= 0.74 and RMSEP = 0.578 m^2m^{-2}) than PLSR models (R_p^2 = 0.732 and RMSEP= 0.58 m^2m^{-2}). Variable Importance in the Projection (VIP) analysis of optimal bands show that the most influential bands were located in the near infrared (NIR) and shortwave (SWIR) regions of the electromagnetic spectrum. The superior performance of iPLSR over PLSR confirmed the fact that the reduction of data dimensionality to optimal wavebands improves model accuracies. The superiority of PLSR over SVR at early and late summer could be attributed to

its ability to quantify linear relationships in dataset and to be less sensitive to background reflectance at low canopy cover. However, the outperformance of SVR over PLSR at mid-summer and for pooled dataset may be explained by its ability to deal with non-linearity, observed in high dense canopy, when saturation in reflectance sets in. Findings in this study provide a practical insight on the potential of mapping LAI on heterogeneous grasslands at a regional scale using air or space-borne sensors.

Preface

This study was undertaken in the School of Agricultural, Earth and Environmental Sciences, University of KwaZulu-Natal, Pietermaritzburg, South Africa, from July 2014 to December 2015, under the supervision of Prof Onisimo Mutanga and Dr. John Odindi to fulfil the requirements of Master in Science.

I declare that the current work represents my own ideas and has never been submitted to any other academic institutions. Acknowledgement has been duly made for statements originating from other authors.

ZZZS Kiala Signed: _____ Date: _____

1. Prof Onisimo Mutanga (Supervisor) Signed: _____ Date: _____

2. Dr. John Odindi (Co-superviosr) Signed: _____ Date: _____

Plagiarism declaration

I ZZZS Kiala, declare that:

1. The research reported in this thesis, except where otherwise indicated is my original research.
2. This thesis has not been submitted for any degree or examination at any other institution.
3. This thesis does not contain other person's data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.
4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted:
    a. Their words have been re-written and the general information attributed to them has been referenced.
    b. Where their exact words have been used, their writing has been placed in italics inside quotation marks and referenced
5. This thesis does not contain text, graphics or tables copied and pasted from the internet, unless specifically acknowledged, and the source being detailed in the thesis and in the references section.

Signed: _____

Acknowledgement

The completion of this study was made possible with the contribution of many helping hands that I would like to acknowledge.

To my supervisors, Prof. Onisimo Mutanga and Dr. John Odindi, thank you for your mentorship and for trusting in me. You taught me to be an independent scientist and a critical thinker. I would also like to thank you for your financial support, which was a great source of motivation to work harder. I appreciate the valuable contribution by Dr Kabir Peerbhay on manuscripts emerging from this research. Special thanks to Timothy Dube and Mbulisi Sibanda for their constant willingness to proof read my work.

I would like to extend my gratitude to my cousin, Ognelet Marie Claude for his steadfast financial support throughout my masters program. His dream to see me pursuing a master degree is at last fulfilled. Many thanks to my dear parents and relatives for their moral support and love. Parting from them was heart-breaking, but they never got tired to follow up on my studies. My family in Christ cannot be forgotten; I am particularly thankful to Bro John Miteo and his family for their prayers and encouragement.

I deeply acknowledge the assistance of fellow students during the field work, namely Mfundiso Cele, Nokwanda Mthethwa, Ndoni Mcunu, Reneilwe Maake, Kgaugelo mogano, Dionne Farai and Bangani Dube. Find in this work the results of your labour and sacrifice.

Finally, I would like to thank the staff of Geography Department, especially Brice Gijsbertsen, Victor Bangamwabo and Prof Trevor Hill for the scientific and technical support.

Table of contents

Chapter 3 Comparison between PLSR and SVR in estimating LAI using optimal bands

Chapter 4 Synthesis

List of tables

List of figures

Chapter 1

General introduction

1.1 Introduction

Grasslands of Southern Africa are of significant subsistence, commercial and ecological value. They provide forage for livestock and wildlife and valuable goods and services that include fuel wood, edible herbs and insects for humans (Chen et al., 2009; Shackleton et al., 2002). Quantitative (e.g. biomass and Leaf Area Index) and qualitative (e.g. nitrogen and phosphorus) variables of grasses, which are spatially and temporally dynamic, determine sustainable range production and ensure the continuous supply of goods and services (Shen et al., 2014; Si et al., 2012). For example, for continued sustainable grazing, a minimum of grass quantity is required (Adler et al., 2001). However, land degradation due to overgrazing, have been attributed to poor management of grazing lands (Ramoelo et al., 2013). According to Snyman (1999), 66 % of rangelands have undergone a moderate to serious land degradation in South Africa. Therefore, it is crucial to understand the spatial and temporal patterns of grass quality and quantity in order to sustainably manage both subsistence and communal rangelands.

Leaf Area Index (LAI) is a critical biophysical parameter that has been used for measuring biomass canopy and assessing grassland productivity and carbon balance (He et al., 2007). Leaf Area Index drives the exchange between the atmosphere and the earth surface, hence plays a critical role in the biophysical processes such as photosynthesis, canopy water interception, transpiration, radiation extinction, carbon loads and nutrient sequestration (Leuschner et al., 2006; Chen and Cihlar, 1996; Chason et al., 1991). A number of studies (Shen et al., 2014; Pfeifer et al., 2012; Doraiswamy et al., 2004; Bréda, 2003) have used LAI to model vegetation foliage cover, growth and productivity and effects of disturbances such as climate change, drought and defoliation on vegetation communities.

Currently, direct (e.g. area harvest) and indirect (e.g. use of ceptometer, LAI-2000 canopy analyser, hemispherical canopy photography, spectrometer, aerial and space-borne sensors) approaches are used to determine LAI on grasslands (Shen et al., 2014; Zhang et al., 2012; Jonckheere et al., 2004; Weiss et al., 2004; Bréda, 2003). Direct methods involve LAI using

planimetric or volumetric techniques. In the planimetric technique, individual leaf area is correlated to the number of area units covered by that leaf in a horizontal canopy layer. In the volumetric technique, leaves are dried and correlated to leaf area using predefined ratios between green-leaf-area and dry-weight, referred to as leaf mass per area (Jonckheere et al., 2004). Whereas direct measurements are regarded as more reliable and serve as reference for calibrating indirect measurements, they involve destructive sampling, are labour intensive, costly and time-consuming (Bréda, 2003). Furthermore, they are difficult to implement on large spatial extents and not suited for long-term LAI spatio-temporal monitoring (He et al., 2007; Bréda, 2003; Chason et al., 1991). Indirect methods on the other hand derive LAI by implementing mathematical expressions and/or radiative transfer theory on a related measureable variable (Ryu et al., 2010). For example, to determine the value of LAI from vegetation canopy, a spectrometer measures canopy reflectance, which is used as a proxy for modelling LAI (Jonckheere et al., 2004). In comparison to direct measurements, indirect LAI measurements are relatively quick, non-destructive and can be automatically processed, thus allowing for LAI determination on a larger sampling area. Consequently, indirect methods are increasingly becoming popular (Jonckheere et al., 2004).

Remote sensing techniques are regarded as an indirect approach of determining LAI. Remote sensing techniques are less costly, non-destructive, relatively quick and often the only reliable means for spatio-temporal LAI determination (Shen et al., 2014; Gray and Song, 2012; Pullanagari et al., 2012; Bulcock and Jewitt, 2010; Chen and Cihlar, 1996). However, traditional remotely sensed imagery are characterised by broad band widths which contain consolidated spectral information for biophysical variables, resulting in loss of useful information available in the narrow bands widths (Thenkabail et al., 2000). Shen et al. (2014) for instance noted that adoption of traditional image data for determining LAI could hardly explain 50 % of LAI variablity (Shen et al., 2014). Due to this limitation, use of hyperspectral data has been proposed for LAI modelling.

According to Lee et al. (2004) and Marabel and Alvarez-Taboada (2013), several studies have demonstrated the superiority of hyperspectral data in predicting LAI over traditional remotely sensed multi-spectral data. However, the huge spectral information contained in hyperspectral data makes LAI retrieval challenging (Darvishzadeh et al., 2008). Multi-collinearity presents a high degree of redundancy and affects the performance of hyperspectral dataset (Li et al., 2014). Moreover, hyperspectral dataset are often degraded by a lower signal-to-noise ratio (Marabel and Alvarez-Taboada, 2013). To deal with this limitation, partial least square

regression (PLSR) was introduced (Wold et al., 2001). PLSR decomposes highly collinear explanatory variables (X) into a few non-correlated components using information contained in the dependent variable (Y), then it predicts the Y variable using the new components (Tobias, 1995; Cho et al., 2007). Unlike other algorithms, PLSR can be run on data where the number of predictors is greater than the number of dependent variables. Numerous studies (Darvishzadeh et al., 2008; Cho et al., 2007; Hansen and Schjoerring, 2003) that have modelled LAI or biomass in heterogeneous grasslands using hyperspectral data have demonstrated the superiority of PLSR over traditional regression techniques such as stepwise and univariate linear regressions using vegetation indices. Nevertheless, PLSR uses all available spectral bands during the computational process of model development. Andersen and Bro (2010) and Abdel-Rahman et al. (2014) showed that the removal of uninformative bands from the hyperspectral data improves model accuracies, simplifies interpretation and reduces data dimensionality and collection costs (Atzberger et al., 2004).

Interval partial least squares (iPLSR), proposed by Nørgaard et al. (2000) for chemometric analyses, is one of the variants of PLSR that can reduce hyperspectral data into a portion of bands relevant for prediction. Its principle is to subdivide the electromagnetic spectrum into equidistant intervals and then run PLSR on the spectral intervals. The local PLSR model with the lowest root mean square (RMSE) is finally selected as the best model. iPLSR has the advantage of visually providing a general overview of optimal bands in different spectral regions, thereby sorting out optimal portions of the electromagnetic spectrum from uninformative portions (Navea et al., 2005; Nørgaard et al., 2000). The selected wavebands in the spectral optimal portions may be useful for the development of sensors on satellite and aerial platforms. Whereas studies in chemometrics that have applied this technique have concluded that iPLSR models are more accurate and reliable than full spectrum PLSR models (Zhou et al., 2009; Borin and Poppi, 2005; Navea et al., 2005; Nørgaard et al., 2000), only a few studies have investigated its ability in the field of remote sensing (Mao et al., 2015; Zhang et al., 2012). In this study we hypothesise that iPLSR can be used to improve LAI estimation in heterogeneous grasslands.

Generally, canopy reflectance on heterogeneous grasslands is complicated by multiple surface materials such as varying species composition, phenology, proportions and complex canopy architecture (Darvishzadeh et al., 2008; Röder et al., 2007). In Addition to grass canopy heterogeneity, temporal variability in different growing seasons impedes the performance of remote sensing data in estimating LAI (Shen et al., 2014). For example, at peak season (LAI >

2.5), the saturation problem is observed. This results in non-linearity between canopy reflectance and biophysical variables such as biomass and LAI (Kooistra, 2012; Chen et al., 2009; Wu et al., 2008; Thenkabail et al., 2000; Clevers and Huete et al., 1997). Being a linear regression method, PLSR would not be suited for dataset collected in this period (Wold et al., 2001). Therefore an appropriate multivariate regression method, which accounts for non-linearity, is necessary as a surrogate to PLSR in heterogeneous grasslands.

Support vector regression (SVR) has been proven to quantify linear and nonlinear relationships in dataset (Üstün et al., 2005; Thissen et al., 2004). The SVR, introduced by Petsche et al. (1997) for functions estimation (Smola and Schölkopf, 2004), belongs to the support vector machines (SVMs) family (Vapnik and Vapnik, 1998). SVR works by constructing hyperplane/s in high or infinite-dimensional space, which can separate quantitative estimates for regressions (Malenovsk et al., 2015). While PLSR is widely used in the field of remote sensing for estimating biophysical and biochemical variables from remotely sensed data (Cho et al., 2007; Darvishzadeh et al., 2011; Hansen and Schjoerring, 2003; Herrmann et al., 2011), little is known about the value of SVR (Yang et al., 2011). Whereas some studies that compared PLSR and SVR showed that SVR were more accurate than PLSR models (Üstün et al., 2005; Thissen et al., 2004), others were contrary (Marabel and Alvarez-Taboada, 2013; Shah et al., 2010). A comparison between the two algorithms on hyperspectral data at different temporal scales would therefore indicate the value of the algorithm and the ideal period of application. This is particularly crucial for the development of reliable temporal and multi-temporal models of LAI in heterogeneous grasslands.

1.2 Aims and objectives

The major aims of this study were:

- To investigate the potential of iPLSR on hyperspectral data in estimating LAI on a heterogeneous tropical grassland and
- To compare the performance of PLSR and SVR using optimal bands selected by iPLSR in estimating LAI on a heterogeneous tropical grassland.

The major objectives in the above named aims were:

- To compare PLSR and iPLSR on hyperspectral data.
- To evaluate the robustness of PLSR and iPLSR models at three sampling periods (i.e. onset, mid and end of the planting season) and pooled reflectance data during summer.

- To compare the performance of PLSR and SVR on iPLSR selected optimal bands at three sampling periods within summer (early, mid, late).
- Identify wavebands with Variable Importance in the Projection (VIP) scores above the significant threshold.

1.3 Key research questions

- To what extent can LAI be estimated using ground-based multi-temporal hyperspectral data and regression techniques in tropical grasslands during the growing season?
- What is the best period/s and regression techniques for LAI estimation?
- What are the most optimal bands for LAI estimation?

1.4 Structure of the dissertation

This dissertation comprises four chapters. The first chapter introduces the study. The second chapter focuses on the potential of iPLSR in estimating LAI using hyperspectral data while the third chapter deals with the comparison between PLSR and SVR in modelling LAI using optimal wavebands. The second and third chapters correspond to two research papers (one under review and another in preparation) and therefore include the literature review and methods used in this study. The fourth chapter is a synthesis of different findings.

Chapter 2

The potential of iPLSR in estimating LAI using hyperspectral data

This chapter is based on:

Kiala, Z., Mutanga, O., and Odindi, J., 2015. The potential of interval Partial Least Sqaure Regression (iPLSR) in estimating Leaf Area Index on a tropical grassland using hyperspectral data. *International Journal of Remote Sensing*, In Review.

Abstract

Leaf area index (LAI) is a critical parameter in determining vegetation status and health. In tropical grasslands, reliable determination of LAI, useful in determining above ground biomass, provides a basis for rangeland management, conservation and restoration. In this study, interval partial least square regression (iPLSR) in forward mode was compared to partial least square regression (PLSR) to estimate LAI from in-situ canopy hyperspectral data on a heterogeneous grassland. Canopy reflectance was collected using ASD FieldSpec® 3 spectrometer at different periods (onset, mid and end) during summer. Partial least squares regression (PLSR) and interval partial least squares regression (iPLSR) were then used to select the best spectral intervals. The performance of the two techniques was determined using the least root mean square error (RMSE) and the highest coefficients of determination (R^2) between the predicted and the measured variable. Results show that iPLSR models could explain LAI variation with R^2_p values ranging from 0.809 to 0.933 and low RMSEP values from 0.211 to 0.603 m^2m^{-2}. iPLSR model at the beginning and end of summer could estimate LAI with the highest accuracies (R^2_p = 0.882 and RMSEP = 0.299 m^2m^{-2} ; R^2_p = 0.890 and RMSEP = 0.211 m^2m^{-2} respectively). Pooling data sets from the three assessed periods yielded the highest prediction error (RMSEP=0.603). Results show that iPLSR performed better than the PLSR, which yielded R^2_p and RMSEP values ranging from 0.364 to 0.649 and from 0.542 to 0.694 m^2m^{-2}, respectively. Overall, this study demonstrates the value of iPLSR in predicting LAI and therefore provides a basis for more accurate mapping and monitoring of canopy characteristics of tropical grasslands. The study further provides an indication of the bands useful for development of sensors on aerial and satellite platforms, necessary for large scale tropical grassland monitoring.

Keywords: interval partial least square regression, partial least square regression, Leaf area index, hyperspectral data, tropical grassland.

2.1 Introduction

Measurement of spatio-temporal distribution of quantitative variables like leaf area index (LAI) and biomass are valuable for assessing the health and productivity of tropical grasslands (He et al., 2007). Several studies (Cho et al., 2007; Prins and Beekman, 1989; McNaughton, 1988) have associated vegetation characteristics such as LAI and biomass with animal grazing patterns. Therefore, quantitative assessment of such characteristics offer great potential for determining grassland conditions, useful for generating optimal management guidelines for grazing and rangeland conservation and restoration.

Leaf area index (LAI) has been recognized as a key biophysical parameter for determining vegetation characteristics (Darvishzadeh et al. 2011 ; Broge and Mortensen 2002). Leaf area index determines vegetation biophysical processes such as photosynthesis, canopy water interception, transpiration, radiation extinction, carbon loads and nutrient sequestration (Chen and Cihlar, 1996; Chason et al., 1991). Consequently, LAI is commonly used as a key input for modelling vegetation foliage cover, growth and productivity and effects of disturbances such as drought and climate change on vegetation communities (Bréda, 2003).

Previous studies that estimated LAI on tropical grasslands have emphasized on their spatial variation (Darvishzadeh et al., 2008). However, LAI is a biophysical parameter that is spatially and temporally dynamic across a landscape. According to Shen et al. (2014), the performance of biophysical process models are highly sensitive to the temporal and spatial variation of LAI. Xu and Baldocchi (2004) noted that well timed data collection on changes in LAI could be used to explain more than 84% of the variance in gross primary production, an important input in carbon cycle of an ecosystem. Therefore, analysis of temporal and spatial changes in LAI at the canopy level provides a valuable opportunity for modelling biophysical processes.

Traditionally, direct (e.g. destructive sampling) and indirect (e.g. use of ceptometer LAI-2000 canopy analyser and hemispherical canopy photography) methods are used to determine LAI in grasslands (Shen et al., 2014; Zhang et al., 2012; Jonckheere et al., 2004; Weiss et al., 2004; Bréda, 2003). Typically, the direct methods consist of manually determining LAI using planimetric or volumetric techniques. Whereas, these approaches are simple and reliable (Levy and Jarvis, 1999; Van Gardingen et al., 1999), they involve destructive sampling, are labour intensive, costly and time-consuming (He et al., 2007; Chason et al., 1991). This limits their application for estimating LAI, particularly in large spatial extents that require frequent

monitoring (Bréda, 2003). Indirect methods, like the use of a spectrometer on the other hand quantify LAI by measuring spectral reflectance which is then used as a proxy for modelling LAI. Generally, such indirect methods are quick and can be automatically processed, thus allowing their application in a larger sampling area (Jonckheere et al., 2004).

Remotely sensed spectral data presents an opportunity to indirectly retrieve LAI in heterogeneous grasslands (He et al., 2007). Techniques that rely on remotely sensed spectral data are non-destructive, relatively quick and cost-effective, and therefore valuable for large spatial and multi-temporal monitoring (Shen et al., 2014; Pullanagari et al., 2012; Bulcock and Jewitt, 2010). Literature shows that canopy hyperspectral data, acquired using hand-held spectrometers has been widely adopted to derive LAI in heterogeneous grasslands (Shen et al., 2014; Si et al., 2012; Banskota, 2006; Atzberger et al., 2004; Thenkabail et al., 2004; Hansen and Schjoerring, 2003). According to Hansen and Schjoerring (2003), such data provide hundreds or even thousands of spectral bands with information sensitive to specific vegetation variables valuable for modelling. However, whereas Lee et al. (2004) demonstrated that models generated from hyperspectral data predicted LAI better than broadband spectral data, the large spectral information that characterise hyperspectral data makes derivation of LAI from heterogeneous grasslands data challenging (Darvishzadeh et al., 2008). Additionally, hyperspectral datasets suffer from multi-collinearity that often occurs when many adjacent spectral bands present a high degree of redundancy and correlation (Li et al., 2014). Tropical grasslands LAI retrieval using canopy reflectance is further complicated by varying species composition, phenology and proportions and complex canopy architecture.

A number of studies (Nguyen and Lee, 2006; Atzberger et al., 2004; Yeniay and Goktas, 2002) that have adopted canopy reflectance hyperspectral data to derive LAI demonstrated the superiority of partial least square regression (PLSR) over traditional regression techniques. The technique was introduced to solve multi-collinearity and over-fitting problems by reducing variables to fewer components (Li et al., 2014). The PLSR technique is a full spectrum method that simultaneously use all available wavebands to create models. Compared to other algorithms, PLSR is less restrictive because it can be run on data where sample size is smaller than predictor variables (Dorigo et al., 2007). The technique is particularly useful for removing uninformative bands and retains those useful for predicting response variables. Consequently, it has become valuable for improving inter alia model predictions by reducing data collection costs, interpretation complexity and data dimensionality (Abdel-Rahman et al., 2014; Andersen and Bro, 2010).

Whereas use of PLSR, a full spectrum technique, has gained popularity in hyperspectral data modelling (Li et al., 2014; Nguyen and Lee, 2006; Atzberger et al., 2004; Yeniay and Goktas, 2002), studies in fields like chemometrics have suggested that interval partial least squares (iPLSR), a variant of PLSR, can reduce hyperspectral data into band portions valuable for more accurate prediction (Zhou et al., 2009; Borin and Poppi, 2005; Navea et al., 2005; Nørgaard et al., 2000). Developed by Nørgaard et al. (2000), iPLSR is a graphically oriented technique for local regression modelling of spectral data. Unlike PLSR, it visually provides a general overview of relevant information in different spectral regions, thereby screening out important portions of the electromagnetic spectrum and discarding interference from irrelevant portions. Nørgaard et al. (2000) for instance used spectra for beer samples to retrieve original extract concentration by comparing iPLSR, PLSR and other algorithms. They found that iPLSR improved determination coefficient and root mean square error of prediction of full spectrum PLSR model from 0.993 and 0.40 % to 0.998 and 0.17 %, respectively. Whereas this approach offers great promise in improving landscape modelling accuracy, no documented studies have used iPLSR on ground-based hyperspectral data collected from heterogeneous landscapes such as tropical grasslands. Consequently, this study sought to pursue two objectives, firstly, to compare heterogeneous tropical grasslands LAI estimates using iPLSR and PLSR models based on hyperspectral data and secondly, to evaluate the robustness of the two models in estimating multi-temporal tropical grassland LAI (i.e. onset, mid and end) and pooled reflectance data during summer.

2.2 Materials and Methods

2.2.1 The study area

The study area is located in the Ukulinga Research Farm at the University of KwaZulu-Natal in Pietermaritzburg (Figure 2-1). The area is characterized by warm to hot summers and mild winters, often accompanied by occasional frost. Mean monthly temperature range from 13.2ºC to 21.4ºC, with a 17ºC annual mean (Everson et al., 2013; Mills and Fey, 2004). The farm receives over 106 days of rain with an annual precipitation of about 680 mm. Soils originate from shallow marine shales of Lower Permian Ecca Group classified as Westleigh forms. The area is under the Southern Tall Grassveld and is predominately herbaceous due to frequent mowing and long term burnings (Mills and Fey, 2004). *Themeda triandra* Forssk, *Heteropogon contortus* (L.) P. Beauv. ex Roem. Schult. and *Tristachya leucothrix* Trin. ex Nees dominate the area (Ghebrehiwot et al., 2013).

2.2.2 Field sampling

Data for the study was collected during the southern hemisphere summer (October of 2014 to March of 2015). Stratified random sampling with clustering was adopted to select sampling sites. The grassland area was first digitized from an aerial photograph (Figure 2-1) and stratified into North, South, East and West aspects. To select the plots, 10 x-y coordinates were randomly generated from the stratum using the Hawth tool. In total, 40 plots (30 m x 30 m) were selected and located in the field using a GPS (Trimble GEO XT, with an estimated 10 cm accuracy). Two to three subplots of 1 m x 1 m were randomly chosen within each plot to generate a final sample size of 100 plots. Spectral and LAI data were then collected within the subplots at the on-set, mid and end of summer.

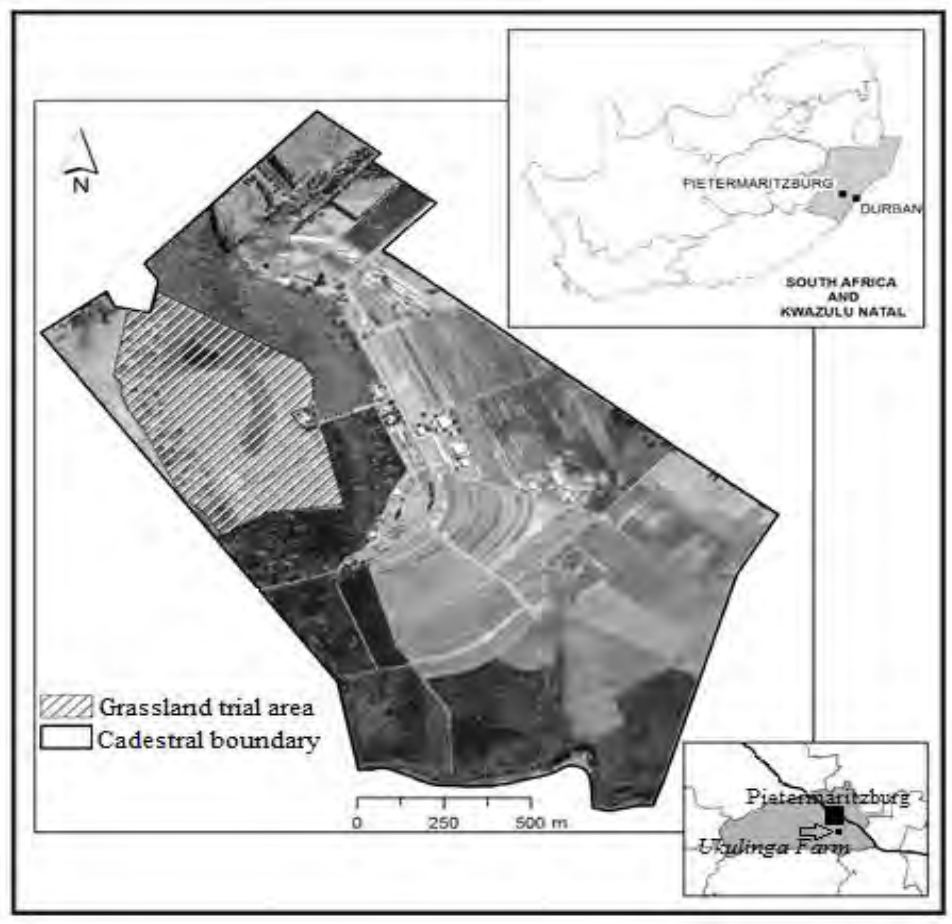

Figure 2-1 The study area

2.2.3 Data collection

Leaf Area Index at sampling points was acquired using LAI-2200C Plant Canopy Analyzer using the procedure described by Darvishzadeh et al. (2008) while canopy reflectance was acquired using an Analytical Spectral Device (ASD), ASD FieldSpec® 3 spectrometer (Inc., Boulder, CO, USA). The spectral resolution of the ASD FieldSpec® 3 spectrometer ranges from 350nm to 2500nm with 1.4 nm and 2 nm sampling intervals for the ultraviolet to visible and near infrared region (350-1000 nm) and the short-wave infrared region (1000-2500 nm) respectively. To normalize the spectra collected, the radiance of a white standard panel coated with Barium Sulphate (BaSO4) and of known reflectivity was first recorded. Canopy reflectance measurements were made under clear sky between 10:00 and 14:00 hrs local time to minimize atmospheric effects. To account for any changes in the atmospheric condition and the sun irradiance, reflectance measurements were recorded with frequent normalization using the standard panel (Adjorlolo et al., 2013). Fifteen replicates of canopy reflectance within each subplot were collected and averaged, allowing for elimination of measurement noise arising from soil background (Darvishzadeh et al., 2008).

2.2.4 Data analysis

2.2.4.1 Pre-processing of hyperspectral data

To separate overlapping bands, thereby amplifying fine differences in the electromagnetic spectrum, the first-order derivative at three nanometers was applied on the resulting mean spectral data (Archontaki et al., 1999; Holden and LeDrew, 1998). First-order derivative is also known to be useful in minimising atmospheric and background noise (Pullanagari et al., 2012; Dorigo et al., 2007). A number of researchers (Darvishzadeh et al., 2008; Wang et al., 2008; Thenkabail et al., 2004) have applied first order derivative on hyperspectral data for LAI estimation. The transformed spectra data were then exported to Microsoft Excel wherein noise bands were removed. The spectral regions between 350-399 nm, 1355-1420 nm, 1810-1940 nm, 2470-2500 nm (Figure 2-2) are known to be noisy and were discarded from the spectra (Rajah et al., 2015; Abdel-Rahman et al., 2014; Adjorlolo et al., 2013).

2.2.4.2 Analysis of variance (ANOVA) and Brown-Forsythe test

The combined test of skewness and kurtosis was first employed to evaluate the distribution of the collected LAI data. The test of normality is a prerequisite to assessing data variability. A

perfect normal distribution has skewness and kurtosis values equal to zero (Peat and Barton, 2014). To assess LAI variations between periods within summer, one-way ANOVA and Brown-Forsythe test ($\alpha = 0.05$) were implemented. The use of Brown-Forsythe test, in addition to ANOVA, was justified by the smaller sample size at the end of summer (n = 73) due to the spectrometer failure. According to Maxwell and Delaney (2004) and Sheskin (2003), Brown-Forsythe test is preferred to ANOVA when sample sizes are heterogeneous and is less affected by abnormally distributed data.

2.2.4.3 Partial least squares regression (PLSR)

Partial least squares regression (PLSR) is originally an econometric technique created by Herman Wold in the 1960s that construct predictive models from highly collinear explanatory variables (Yeniay and Goktas, 2002). The principle of PLSR is to firstly decompose explanatory variables (X) into a few non-correlated latent variables or components using information contained in the response variable (Y); then to regress the new components against the response variable (Cho et al., 2007; Tobias, 1995). According to Wang et al. (2011a), Tan and Li (2008) and Yeniay and Goktas (2002), the model that underlies PLSR consists of three phases. In the first phase, explanatory variables (X) and response variable (Y) are decomposed based on the expression:

$$X = TPT + E \quad (1)$$

$$Y = UQT + F \quad (2)$$

Where T and U are respective matrices of scores of X and Y; P and Q stand for the matrices of loadings; E and F, errors of X and Y matrices. In the second phase, the Y-scores (U) are predicted using the X-scores (T) based on the expression:

$$U = bT + e \quad (3)$$

Where b represents the regression coefficient and e, the error matric of the relationship between Y-scores and X-scores. In the final phase, the predicted Y -scores are used to build predictive models of response variable using the expression:

$$Y = bTQ + G \quad (4)$$

Where G is the error matrix related to estimating Y.

In the present study, PLS-toolbox (Eigenvector Research Inc.) used with MATLAB (version R2013b) was used to build PLSR models. Before running PLSR, pre-processed hyperspectral data along with LAI data were autoscaled (Zhang et al., 2012). This procedure scales mean-centres of each waveband to unit standard deviation (Wise et al., 2006). The PLSR was then run on data using a leave-one-out cross-validation method. The least root mean square error (RMSE) and the highest coefficients of determination (R^2) between the predicted and the measured Y variable were the two criteria used to select the best model with optimal number of components. The best model was suggested by the software.

2.2.4.4 Interval partial least squares regression (iPLSR)

Interval partial least squares regression (iPLSR) is a variant of PLS that locally develops PLS models on equidistant portions of the full spectrum (Navea et al., 2005; Nørgaard et al., 2000). To predict a Y variable from spectra using iPLSR, the spectrum is split into a number of intervals of equal distance. A PLSR model is then built on each spectral interval. Thereafter, all the models built on the wavebands of different intervals are compared to the full-spectrum model based on calibration parameters such as root mean squared error of cross-validation (RMSECV). Finally, the local model with the lowest RMSECV is selected (Xiaobo et al., 2007; Andersen and Bro, 2010; de Lira et al., 2010). The iPLSR can operate in two modes or variable selection directions: backward and forward mode. In forward mode, the algorithm starts without any variable selection and then develops the best PLSR model from the interval with the lowest RMSECV. This process can be repeated by including more intervals to enhance the model. In backward mode, the algorithm starts by selecting all variables and then discards the interval with the largest RMSECV (Balabin and Smirnov, 2011; Mehmood et al., 2012).

In this study, iPLSR in forward mode was used to select best spectral intervals. As predictive bands of LAI are known to spread across the entire electromagnetic spectrum (Cho et al. 2007; Darvishzadeh et al., 2008), the interval size was set to a single variable. This approach is recommended when there is uniqueness of information in variables (Wise et al., 2006). After several adjustments, the process was repeated 40 times. Therefore, the output local model had 40 intervals or bands. The iPLSR in forward mode was implemented using PLS-toolbox.

2.2.4.5 Validation

Models were validated using leave-one-out cross validation on the training data set and then validated again on independent test dataset. LAI and spectral data for each period during

summer were split into training data set (70%) and independent test data set (30%) (Kohavi, 1995). To avoid arbitrary data splitting that may cause biased results (Darvishzadeh et al., 2008), onion algorithm was applied. The principle of onion algorithm is to keep outside covariant data plus those that are randomly inner-spaced (Sousa et al., 2015). After splitting data, PLSR and iPLSR were run on training dataset to develop models. The developed models were validated using leave-one-out cross-validation. The leave-one-out cross-validation successively removes a sample from the entire calibration dataset and uses it for validation (Arlot and Celisse, 2010). Then, for all the iterations, RMSECV and R^2_{cv} are calculated and averaged to assess the performance of a predictive model (Knox et al., 2012). Finally, the models obtained through leave-one-out cross-validation were tested on independent test dataset.

2.3 Results

2.3.1 Variation in LAI and spectra data

The values of skewness (between 0.397 and -0.449) and kurtosis (between 0.856 and -0.111) indicate that the LAI of grass species canopy in the sampling plots had a normal distribution. That made the LAI data in the present study suitable for ANOVA and Brown-Forsythe Test. In the three multi-temporal periods, samples in mid-summer had the highest mean (3.626 m^2m^{-2}) and variability (standard deviation= 1.099 m^2m^{-2}) ($p < 0.01$) while samples at the end of summer had the second highest mean (2.015 m^2m^{-2}) and lowest variability (standard deviation =0.705 m^2m^{-2}) during the study.

To assess the change in reflectance at the different sampling periods, the mean spectra of all the sampling plots were averaged and upper and lower 95% confidence limits derived. Results show that there was a change in averaged reflectance during the sampling periods (Figure 2-2). Visually, averaged reflectance was noticeably different across the electromagnetic spectrum. Canopy reflectance at the end, beginning and mid-summer presented the highest mean reflectance in the Visible, NIR and SWIR region respectively. Figure 2-2 shows that first derivative spectra differed in some spectral portions at the different sampling periods.

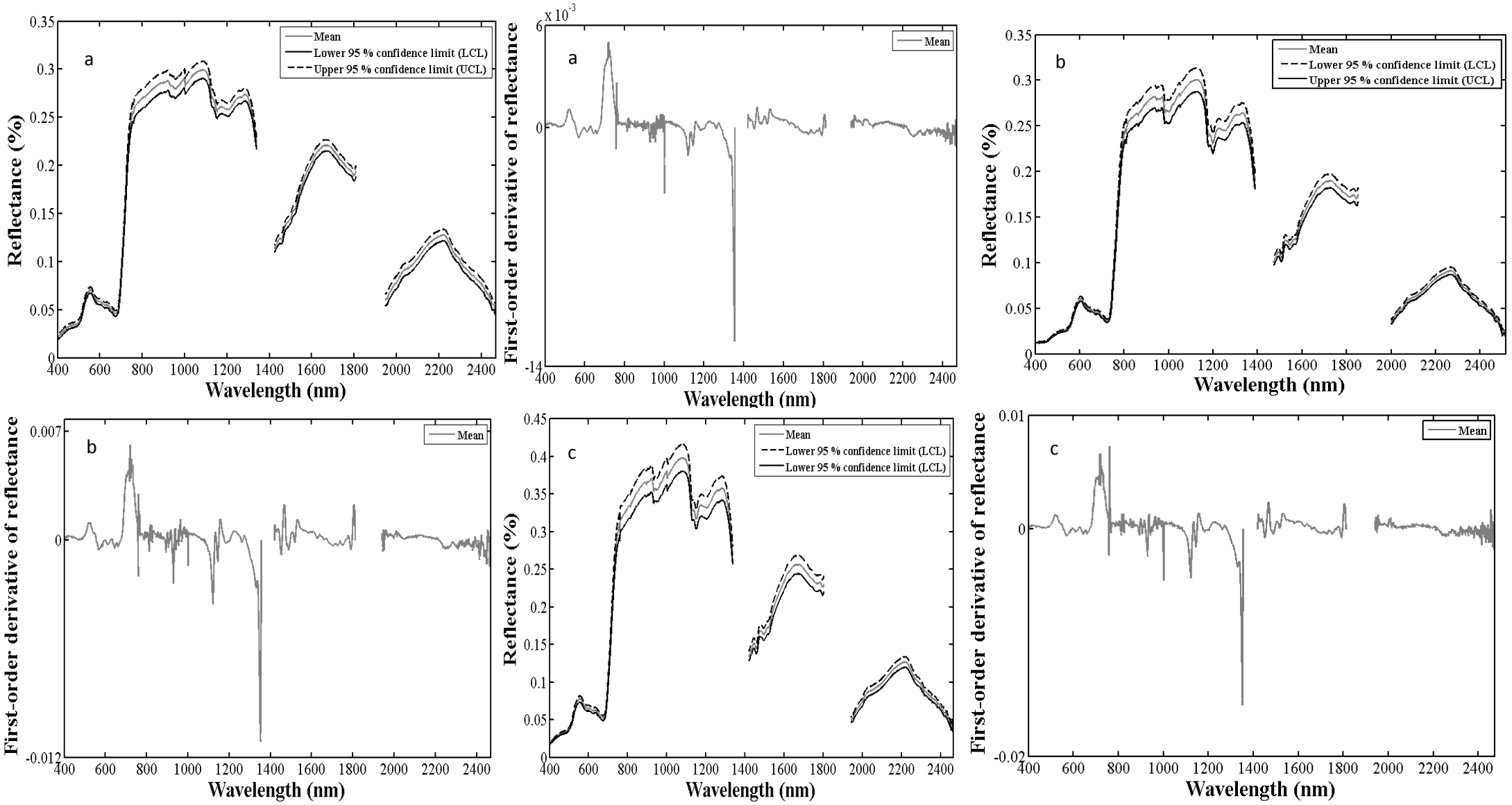

Figure 2-2 Mean and respective first-order derivative of canopy spectra of all grass subplots at the beginning (a), mid (b) and end (c) of summer.

The highest values of first-order derivative of reflectance are located in the NIR and SWIR region of the electromagnetic spectrum.

2.3.2 PLSR and iPLSR models

Table 2-1 presents results of the PLSR and iPLSR models performance for training dataset at each of the sampling periods within summer. Based on RMSECV and R^2, results show that the iPLSR models perform better that the PLSR models. At each period, iPLSR models were able to explain more than 85% of LAI variability (88.8% at the beginning, 90.3% in mid and 89.6% at the end of summer) with RMSECV values that vary from 0.237 to 0.321 (m^2m^{-2}). Although iPLSR had a slightly higher RMSECV value (0.529 m^2m^{-2}) it had a better estimation of LAI variability across the entire summer (R^2_{cv} = 0.809). PLSR models on the other hand yielded high RMSECV values (0.551 - 0.768 m^2m^{-2}) and poorly explained the LAI variation (31.3 – 67.1 %).

Table 2-1 R^2_{cv}, RMSECV and number of factors of training PLSR and iPLSR models

Regression algorithm	R^2_{cv}	RMSECV	Number of factors
Beginning of summer			
PLSR (full-spectrum)	0.313	0.745	6
iPLSR (40 intervals)	0.888	0.288	6
Middle of summer			
PLSR (full-spectrum)	0.537	0.768	4
iPLSR (40 intervals)	0.903	0.321	5
End of summer			
PLSR (full-spectrum)	0.391	0.551	5
iPLSR (40 intervals)	0.896	0.237	6
Combined-period (pooled dataset)			
PLSR (full-spectrum)	0.671	0.750	5
iPLSR (40 intervals)	0.809	0.529	6

The contribution of each waveband in the selected PLSR factors is displayed in Figure 2-3. The most valuable bands for estimating LAI were distributed across the electromagnetic spectrum. However, the highest peaks for all the periods within summer, including all the periods combined, were mostly located in the NIR and SWIR region.

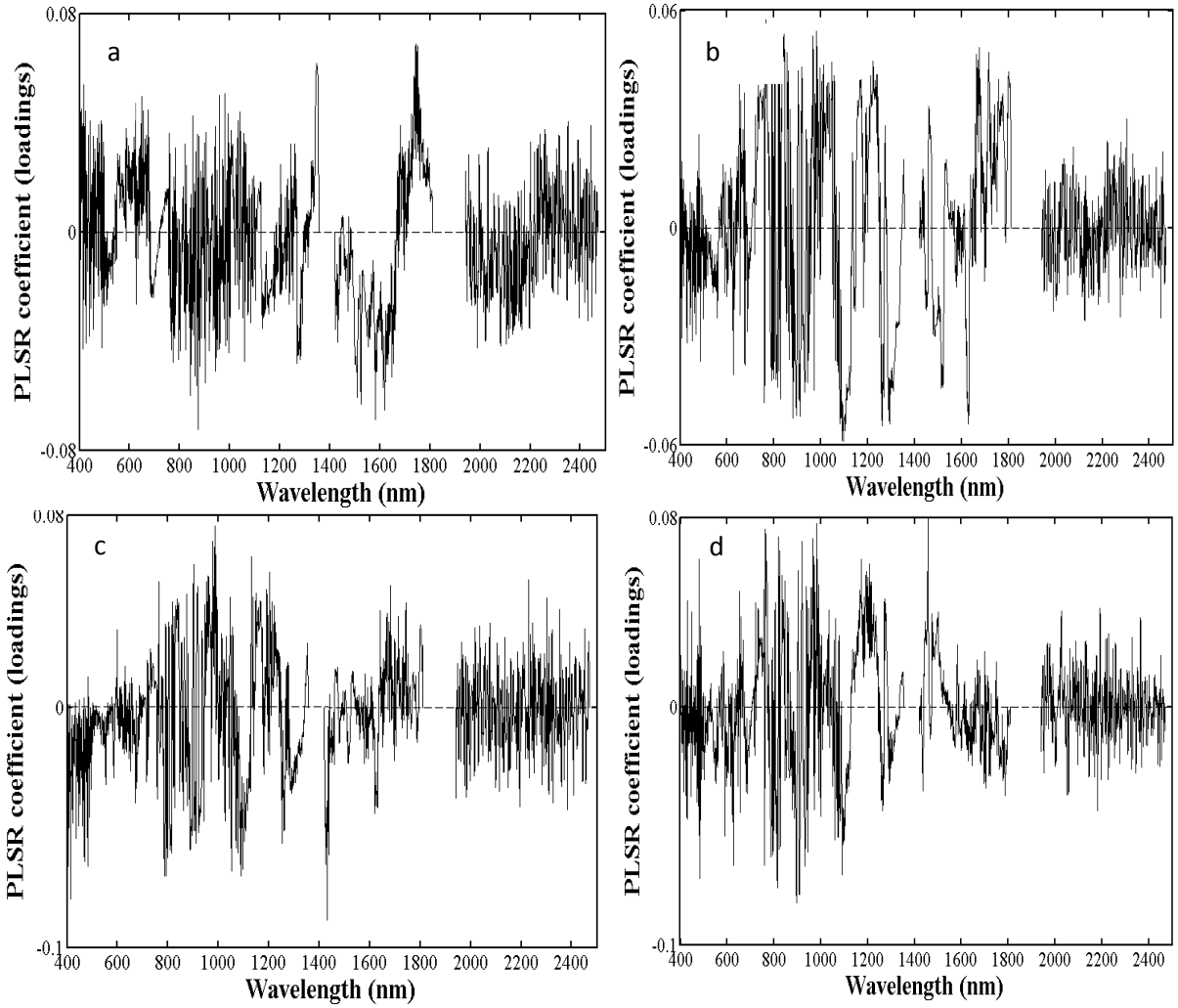

Figure 2-3 PLSR loadings for beginning (a), mid (b) and end (c) of summer and pooled data (d)

Using iPLSR models with 40 intervals, Table 2-2 and Figure 2-4 present the selected bands and their location within the four regions of the electro-magnetic spectrum, respectively while Figure 2-5 provides a percentage number of predictive bands in relation to the regions within the electro-magnetic spectrum.

Table 2-2 Selected bands (nm) and spectral regions at the beginning, mid and end of summer and pooled data

	Visible	Red Edge (RE)	Near InfraRed (NIR)	Short Wave InfraRed (SWIR)
Beginning of summer	461, 764	-	793, 1020, 1061, 1201, 1267	1633, 1640, 1656, 1681, 1708, 1741, 1956, 1997, 2003, 2021, 2071, 2086, 2097, 2117, 2127, 2140, 2165, 2167, 2201, 2219, 2220, 2221, 2286, 2291, 2321, 2344, 2347, 2369, 2388, 2398, 2429, 2436, 2439
Middle of summer	413, 442, 443	-	995, 1132, 1134, 1174, 1240 , 1275	1693, 1944, 1947, 1951, 1959,1969, 1978, 2011, 2042, 2048, 2065, 2181, 2206, 2207, 2216, 2218, 2219, 2258, 2281, 2290, 2319, 2333, 2353, 2388, 2390, 2394, 2424, 2427 ,2434, 2437, 2450
End of summer	-	-	874, 943, 1003, 1010, 1058, 1059	1427, 1430, 1782, 1783, 1960, 1961, 1981, 1985, 1986, 2012, 2018, 2052, 2067, 2102, 2114, 2119, 2141, 2152, 2190, 2208, 2250, 2262, 2301, 2321, 2344, 2364, 2383, 2394, 2396, 2417, 2448, 2455, 2462, 2469
Combined-period	433, 489, 490, 535, 551	732, 752	957, 961, 968, 1062, 1183, 1244	1471, 1478, 1585, 1626, 1656, 1672, 1693, 1708, 1733, 1742, 1780, 2047, 2060, 2075, 2097, 2133, 2136, 2148, 2241, 2259, 2280, 2323, 2325, 2367, 2372, 2403, 2417

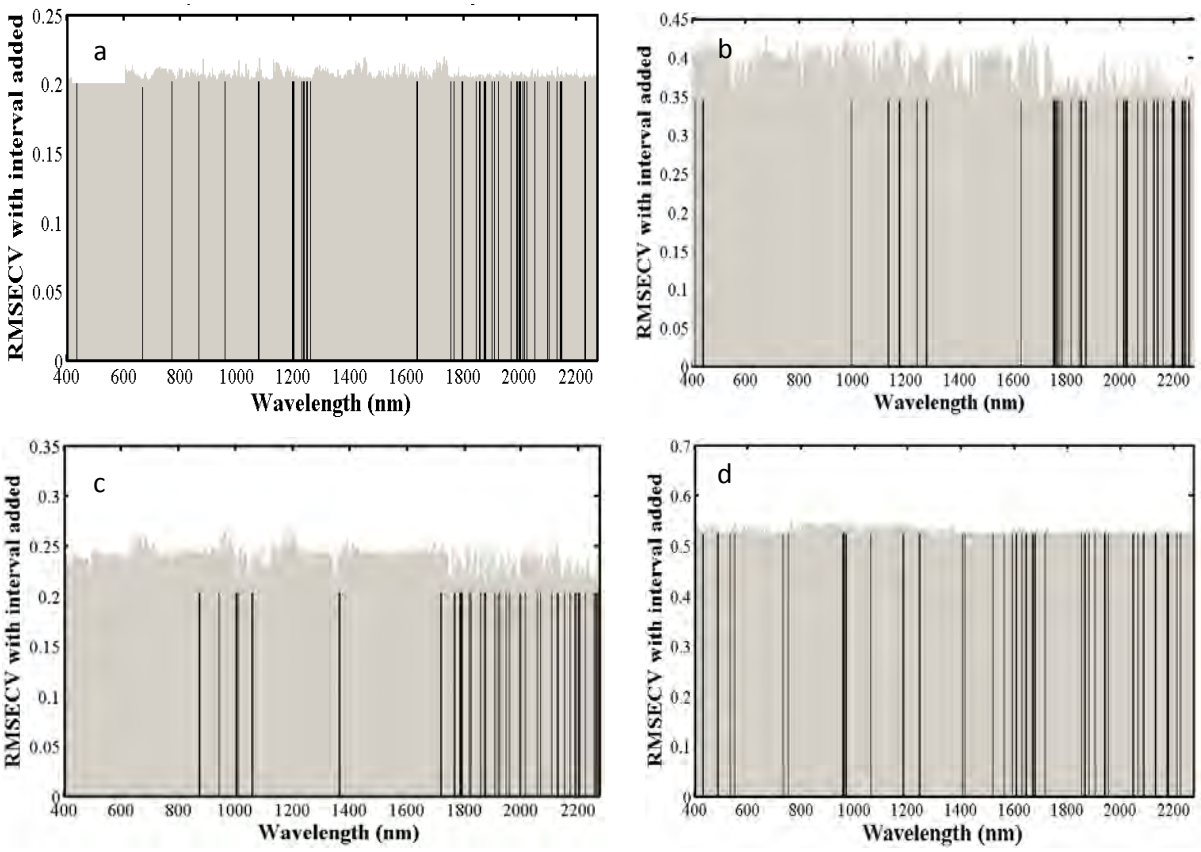

Figure 2-4 Chosen bands (in dark) for modelling LAI at the beginning (a), mid (b) and end (c) of summer and pooled data (d)

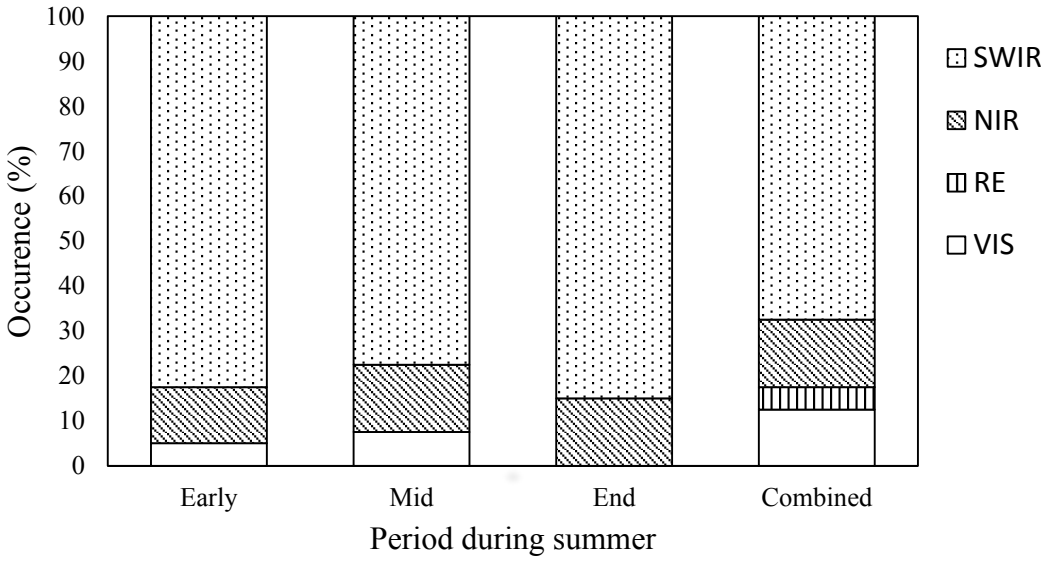

Figure 2-5 Summary of predictive bands of LAI in different spectral regions

2.3.3 Model validation

Figure 2-6 shows the performance of PLSR and iPLSR (40 intervals) models on independent test dataset. PLSR models of all the periods within summer (including all the periods combined) increased the coefficient of determination for prediction (R^2_p) and slightly decreased the root mean square error for prediction (RMSEP). The values of R^2_p and RMSEP respectively, varied from 0.364 to 0.649 and from 0.542 to 0.694 (m^2m^{-2}). However, iPLSR models performed better than the full-spectrum PLSR models for all the sampling periods in summer. The predictive power of iPLSR models did not change much on validation dataset. More than 80 % of new data of LAI could be explained by the iPLSR models at all periods within summer (including all the periods combined).

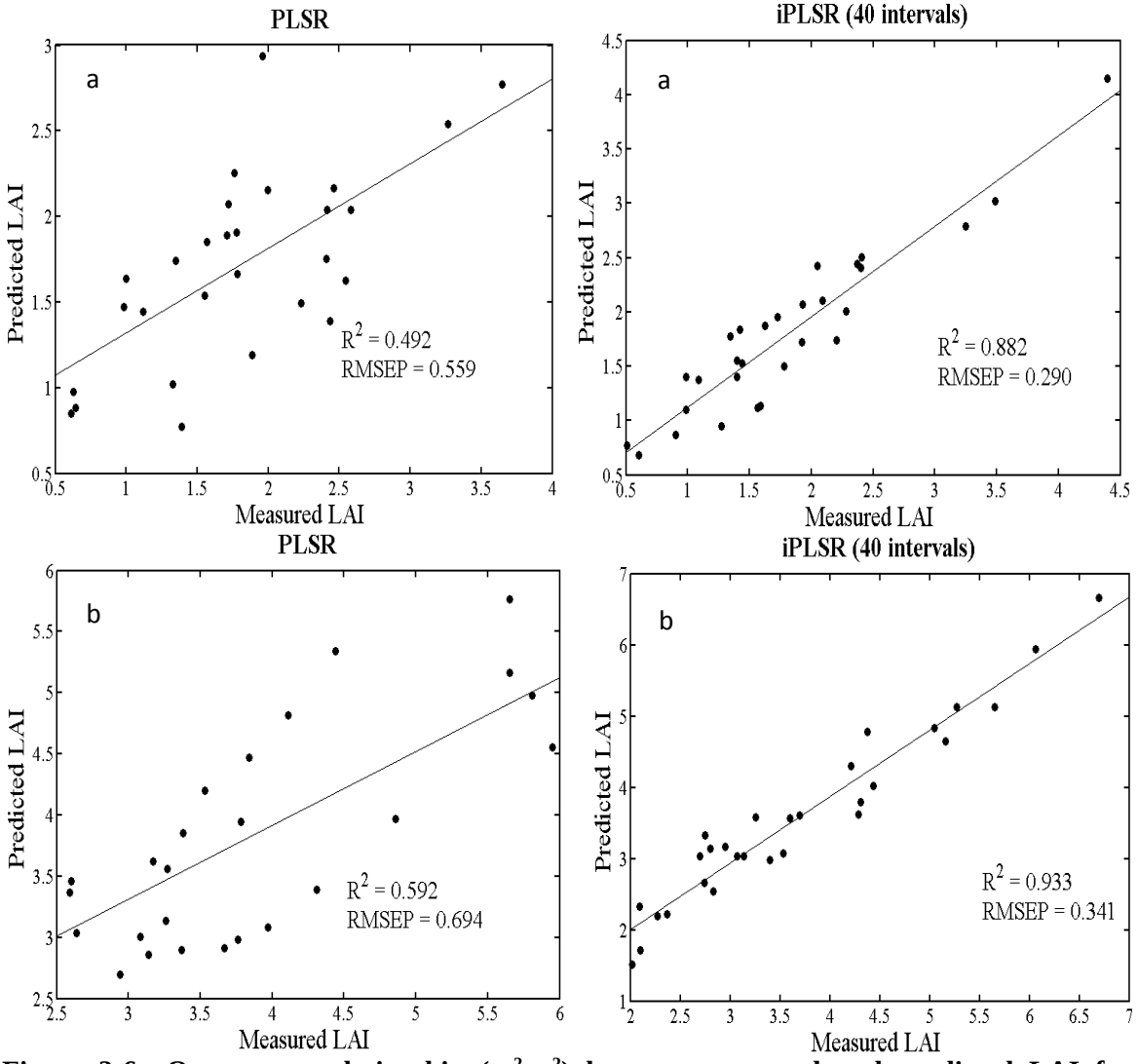

Figure 2-6a One-to-one relationship (m^2m^{-2}) between measured and predicted LAI for validating PLSR and iPLSR models on independent test dataset, early summer (a), mid-summer (b).

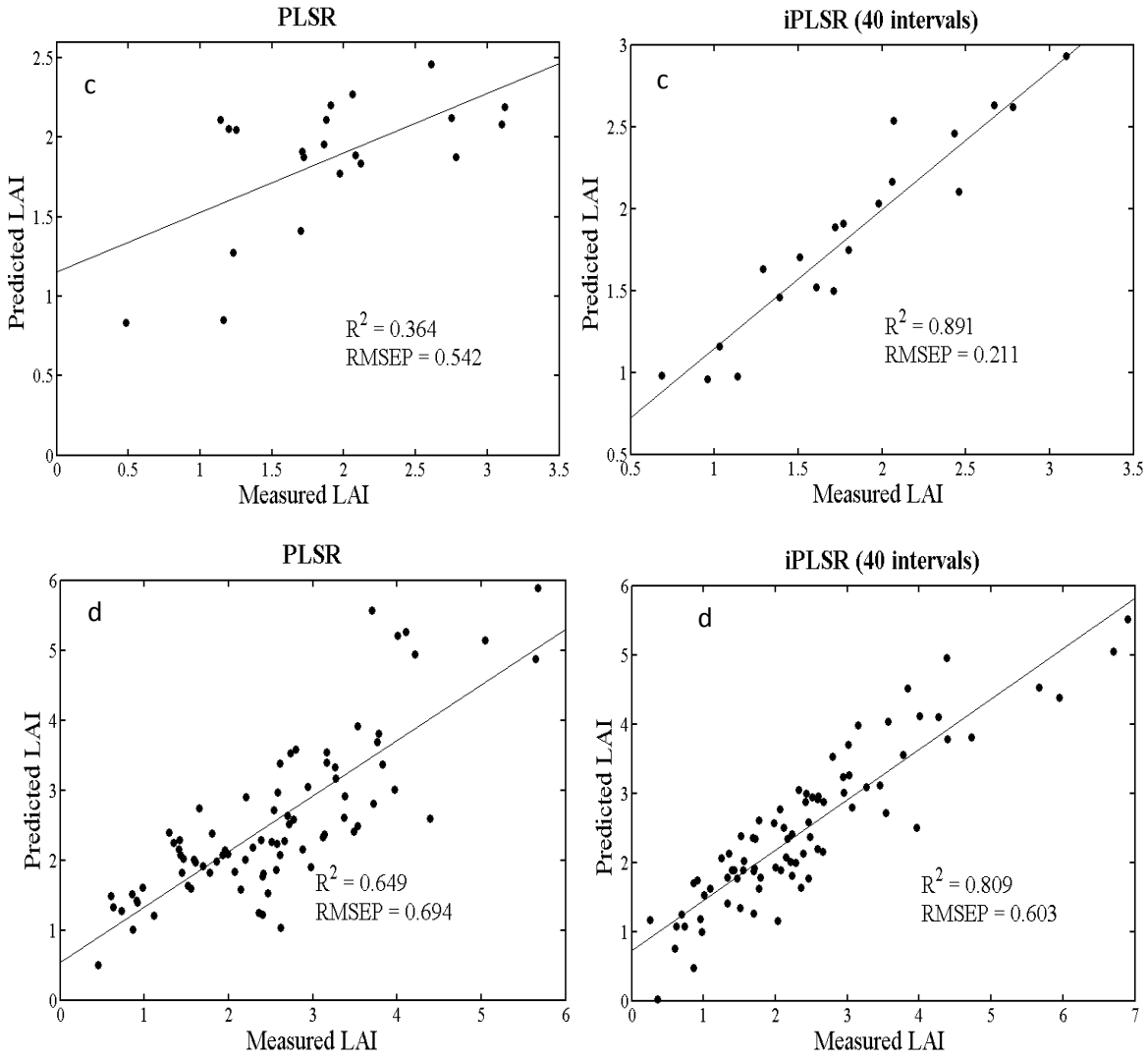

Figure 2-6b One-to-one relationship (m^2m^{-2}) between measured and predicted LAI for validating PLSR and iPLSR models on independent test dataset, end of summer (c) and pooled data (d).

2.4 Discussion

This study sought to determine the performance of two multivariate regression models (PLSR and iPLSR) in estimating canopy level LAI on tropical grassland during summer. Comparisons were determined using the coefficient of determination (R^2) and the root mean square (RMSE). Specifically, this study examined the possibility of developing a model that can estimate LAI at different periods within summer (beginning, mid and end) and across the entire summer period. Use of iPLSR to select the optimal bands for predicting LAI was also investigated.

Results showed that PLSR algorithm run on first derivative spectra to assess LAI variation at different periods did not perform well. The values of R^2_p and RMSE, respectively, ranged from 0.364 to 0.649 and 0.542 to 0.694 (m^2m^{-2}). Albeit PLSR is known to reduce hyperspectral data

to a few useful bands, inclusion of all the wavebands was not useful in the predictive performance of PLSR models, results consistent with Liu (2014), Chung and Keles (2010), Filzmoser (2012) and Karaman et al. (2013). However, when data dimensionality was reduced to useful bands using iPLSR, the performance of models (R^2 and RMSE) significantly improved. Overall, there were very close relationships between measured and predicted LAI values, with low values of RMSE and higher values of determination coefficients (R^2) (Figure 2-6). In consistency with Zhou et al. (2009), Navea et al. (2005), Borin and Poppi (2005) and Nørgaard et al. (2000), these findings confirmed the superiority of iPLSR over full spectrum PLSR.

The best predictive models were derived from early summer canopy reflectance (R^2_p = 0.882 and RMSEP = 0.290 m^2m^{-2}) and end of summer (R^2_p = 0.891 and RMSEP = 0.211 m^2m^{-2}. This was expected as most of reflectance (visible and near infrared) do not saturate at low LAI (Dorigo et al., 2007; Darvishzadeh et al., 2008). This finding concurs with Zhang et al. (2012) and Li et al. (2006) who showed that reflectance models at early wheat growth (tillering, jointing and booting stage) performed better than reflectance models at maximal period of growth (filling stage). In this study, the lower early summer prediction in comparison to end-summer can be attributed to higher soil background noise. According to Darvishzadeh et al. (2008), soil background has a negative effect on the predictive power of hyperspectral data for LAI estimation. In comparison to early and end of summer, model accuracies in mid-summer and all the periods combined marginally dropped. Higher prediction error in mid-summer and combined-period model can be explained by taller and denser grass volume and therefore shadows (Adjorlolo et al., 2013).

Adoption of iPLSR was useful in identifying relevant wavebands for predicting LAI. In total, 40 intervals were identified for all the sampling periods. The success of iPLSR for band selection in this study may be attributed to successful separation of overlapping bands performed by the first-derivative technique on the spectra. The spectral regions (NIR and SWIR) of bands selected by iPLSR are consistent with the findings by Darvishzadeh et al. (2008), Thenkabail et al. (2004), Brown (2000), Schlerf et al. (2005) and Gong et al. (2003). Within ±12 nm, the bands chosen (Table 2-1) in this study showed a consistency with the known bands for estimating LAI. For example, bands near 793 nm, 1061 nm, 1062 nm, 1633 nm, 442 nm, 443 nm, 535 nm, 551 nm, 732 nm, 2190 were also identified by Wang et al. (2008) for estimating rice LAI at different growth phases. Furthermore, Gong et al. (2003) found that bands centred near 1201 nm, 1240 nm, 1062 nm, 1640 nm, 2097 nm, 2259 nm were

useful for estimating forest LAI.

It is worth noting that the contribution of different spectral regions along with their wavebands to LAI estimation depends on a particular period within summer (Table 2-2). This might be explained by the fact that the positions of selected wavebands are sensitive to changes in LAI as indicated by ANOVA and Brown-Forsythe test. Thus, the positions vary when factors like biochemical (e.g. chlorophyll) and biophysical (e.g. canopy closure) parameters and background effects change with canopy growth phases (Wang et al., 2008). For example, at the end of summer, as canopy senesce and the amount of chlorophyll decline, NIR and SWIR become more important in predicting LAI (Zhao et al., 2007). Furthermore, in the combined-period model, the selected bands can be explained by the fact that they were insensitive to changes in LAI. Delegido et al. (2013) found that vegetation indices combining bands at 674 nm and 712 nm could overcome the aforementioned saturation problem while Kim et al. (1994) found similar results with the ratio of 550 and 700 nm, which were insensitive to changes in chlorophyll concentration.

2.5 Conclusion

From the study findings, the following conclusions can be drawn:

- iPLSR can be used to simplify the relationship between LAI and canopy reflectance transformed using first derivative technique better than PLSR.
- The best iPLSR relationship is at the beginning and end of summer. By including all the variables, full-spectrum PLSR models yield higher prediction error.
- iPLSR used as a single variable selection algorithm for LAI estimation can generate stable and reliable models with 40 bands.
- The period within summer, which is associated with vegetation growth, determines the selection and accuracy of LAI predictive bands.

This study has analysed the multi-temporal variation of LAI at the canopy level in a tropical grassland. Results show that appropriate band selection on in-situ hyperspectral data using iPLSR can overcome the challenge faced by remotely sensed data to accurately estimate LAI in a heterogeneous grassland. The findings in this study have paved the way to more accurate mapping and monitoring of canopy characteristics in a tropical grassland from airborne and space borne hyperspectral data. However, the development of iPLSR model for all the

combined periods within summer needs further investigations, as its prediction error was higher than all the models created at different periods.

Chapter 3

Comparison between PLSR and SVR in estimating LAI using optimal bands

This chapter is based on:

Kiala, Z., Mutanga, O., and Odindi, J., and Kabir, P. 2015. A comparison of Partial Least Square and Support Vector regressions in predicting Leaf Area Index on a tropical grassland using hyperspectral data. *International Journal of Remote Sensing*, in preparation.

Abstract

Leaf area index (LAI) is a key biophysical parameter commonly used to determine vegetation status, productivity and health in tropical grasslands. Therefore, accurate estimates of LAI are useful in supporting sustainable rangeland management. Due to the vast amount of information they provide, hyperspectral remotely sensed data in concert with multivariate regression techniques offer new opportunities to accurately estimate LAI in tropical grasslands. Modelling techniques like partial least square regression (PLSR) have become popular in remote sensing, however, recent literature has shown that irrelevant variables affects its performance. Whereas other robust modelling techniques like support vector regression (SVR) have been successful in fields like chemometrics, their potential in remote sensing remain unexplored. In this study, the performance of support vector regression (SVR) was compared to partial least square regression (PLSR) on optimal hyperspectral bands on a heterogeneous grassland at different periods (early, mid and late) within summer. Furthermore, Variable of Importance on the Projection (VIP) of the bands was investigated. The comparison of the two multivariate modelling regressions were based on the root mean square error (RMSE) and the coefficients of determination (R^2) between the predicted and the measured variable. Results show that PLSR performed better than SVR at the beginning and end of summer. For the two sampling periods, PLSR models on the new dataset (30 % of the entire dataset) could respectively explain 86.5 % and 85.1 % of LAI variance with RMSEP values of 0.263 m^2m^{-2} and 0.204 m^2m^{-2}. The LAI variance in the SVR models was 85.8 % and 83.2 % with RMSEP values of 0.287 m^2m^{-2} and 0.218 m^2m^{-2}, respectively. However, at the peak of the growing season (mid-summer), at

reflectance saturation, SVR models yielded higher accuracies (R^2 = 0.902 and RMSE= 0.371 m^2m^{-2}) than PLSR models (R^2 = 0.886 and RMSE= 0.379 m^2m^{-2}). Similarly, for pooled dataset, SVR models were slightly more accurate (R^2= 0.74 and RMSE= 0.578 m^2m^{-2}) than PLSR models (R^2 = 0.732 and RMSE= 0.58 m^2m^{-2},). This finding confirms the ability of SVR to deal with nonlinearity in hyperspectral datasets. With respect to Variable of Importance on the Projection (VIP) scores of bands, result show that most of the bands were located in the near infrared (NIR) and shortwave (SWIR) regions of the electromagnetic spectrum. This study introduces the application of SVR in predicting LAI from sensors on aerial and satellite platforms, necessary for large scale tropical grassland monitoring.

Keywords: partial least square regression, support vector regression, leaf area index, hyperspectral data, tropical grassland.

3.1 Introduction

Grasslands are valuable economic and ecological resources. They provide grazing lands and goods and services (e.g. fuel wood, edible herbs and fruit, insect) (Shackleton et al., 2002; Chen et al., 2009). Grassland health and productivity are determined by, inter alia, biophysical variables like leaf area index (LAI) and biomass, which are spatially and temporally dynamic (He et al., 2007). In the context of southern Africa, overgrazing associated with poor planning and management of grazing lands have been observed in the communal grasslands ecosystems (Ramoelo et al., 2013). According to Snyman (1999), 66 % of rangelands have undergone a moderate to serious land degradation in South Africa. Therefore, accurate estimates of biophysical variables such as LAI and biomass and their spatial and temporal changes may enhance decision and policy making process in grassland management, restoration and conservation in communal rangelands.

Leaf area index is a key biophysical parameter of vegetation characteristics that has been used as a surrogate of canopy biomass (Darvishzadeh et al., 2011; He et al., 2007). Leaf area index has a direct implication on plant productivity as it determines the amount of water and energy exchange between plants and the atmosphere (Leuschner et al., 2006; Chen and Cihlar, 1996). Consequently, LAI has been used as an input to model among others vegetation foliage cover, growth and productivity and effects of disturbances such as climate change, drought, and defoliation (Bréda, 2003).

Hyperspectral sensing techniques, compared to other approaches of measuring LAI, offer new opportunities to accurately estimate LAI at the regional scale because of the large amount of information they provide (Fava et al., 2009). The use of hyperspectral data has significantly improved leaf area index (LAI) estimation (Darvishzadeh et al., 2008; Broge and Mortensen, 2002). However, previous studies on modelling LAI using hyperspectral data have concentrated on its spatial variation (Darvishzadeh et al., 2011; Darvishzadeh et al., 2008; He et al., 2006). Leaf area index is a biophysical parameter that also varies temporally across an ecosystem (Shen et al., 2014). According to Shen et al. (2014), temporal LAI variation determines the performance of biophysical processes models. Therefore, failure to periodically determine LAI estimates may lead to errors in those models. For example, Li (2010) found that temporal changes in LAI could explain more than 84 % of the variance in gross primary production. Hence, there is a need to analyse temporal changes in LAI to reliably model a grassland's biophysical processes (Si et al., 2012).

Although hyperspectral data have proven their superiority in predicting LAI over traditional remote sensing datasets (Marabel and Alvarez-Taboada, 2013; Lee et al., 2004), the large spectral information makes deriving LAI from hyperspectral data challenging (Darvishzadeh et al., 2008). Hyperspectral datasets also suffer from multicollinearity that often occurs when many adjacent spectral bands present a high degree of redundancy and correlation (Li et al., 2014). Moreover, the performance of hyperspectral data are often deteriorated by a lower signal-to-noise ratio (Marabel and Alvarez-Taboada, 2013). Furthermore, spatial heterogeneity and temporal variability of canopy characteristics in heterogeneous ecosystems like tropical grasslands are other major factors that have limited the performance of remotely sensed data (Shen et al., 2014).

Many approaches have been proposed to overcome the above aforementioned limitations. Among others, studies have proposed the selection of informative bands that best correlate with investigated biophysical variables using variable selection algorithms such as stepwise multiple linear regression (SMLR) (Darvishzadeh et al., 2008; Wang et al., 2008; Jensen et al., 2009), interval partial least square regression (iPLSR) (Zhang et al., 2012) and Partial least square regression (PLSR) (Kawamura et al., 2010; Banskota, 2006). Another approach has been the implementation of appropriate statistical modelling methods. Partial least square Regression (PLSR) and Support Vector Regression (SVR) have been among the most dominantly adopted in statistical analysis. Both techniques are full spectrum methods and have been commonly applied in chemometrics (Marabel and Alvarez-Taboada, 2013; Mountrakis et al., 2011; Thissen et al., 2004). Unlike PLSR, which is a linear regression method (Wold et al., 2001), SVR is a non-linear method (Vapnik and Vapnik, 1998), therefore tailored to dealing with non-linearity in dataset, often observed in saturation reflectance from high canopy density (Chen et al., 2009; Thenkabail et al., 2000). Whereas PLSR has been commonly used in determining LAI from hyperspectral data (Li et al., 2014; Darvishzadeh et al., 2011; Atzberger et al., 2010; Hansen and Schjoerring, 2003), the performance of SVR is yet to be established (Yang et al., 2011).

A number of studies, mostly in chemometrics, have compared SVR and PLSR on spectral data (Yue et al., 2015; Marabel and Alvarez-Taboada, 2013; Shah et al., 2010; Li et al., 2009; Üstün et al., 2005; Thissen et al., 2004). Some (Yue et al., 2015; Üstün et al., 2005; Thissen et al., 2004) demonstrated the superiority of SVR models over PLSR models whereas others (Marabel and Alvarez-Taboada, 2013; Shah et al., 2010) have demonstrated the superiority of PLSR over SVR. For example, Üstün et al. (2005) showed that SVR could create models on

NIR data better than PLSR. This performance was attributed to its ability to be insensitive to spectral noise and to quantify nonlinear relationships in dataset. This study sought to: a) compare the performance of PLSR and SVR models in estimating LAI at different periods (early, mid, late) within summer using iPLSR selected wavebands and b) retrieve the bands with the most significant Variable Importance in the Projection (VIP).

3.2 Materials and Methods

3.2.1 Study area

This study was conducted at the University of KwaZulu-Natal research farm in Pietermaritzburg (Figure 3-1). The area is characterised by warm to hot summers and mild winters, accompanied with occasional frost. Mean monthly and annual temperature range from 13.2 ^{o}C to 21.4 ^{o}C and 17 ^{o}C respectively (Everson et al., 2013; Mills and Fey, 2004). Ukulinga farm receives over 106 days of rain with an annual precipitation of about 680 mm. Soils originate from shallow marine shales of Lower Permian Ecca Group classified as Westleigh forms. The area is under the Southern Tall Grassveld and is predominately herbaceous due to frequent mowing and long term burnings. (Mills and Fey, 2004). The following grass species are the most dominant in the area: *Themeda triandra* Forssk, *Heteropogon contortus* (L.) P. Beauv. ex Roem. Schult. And *Tristachya leucothrix* Trin. ex Nees (Ghebrehiwot et al., 2013).

3.2.2 Field plots sampling

Stratified random sampling with clustering was adopted. The grassland trial area was first identified and digitized from an aerial photograph and stratified into North, South, East and West aspects. Coordinates were then randomly generated from the stratum to select the plots using the Hawth tool in ArcGIS 9. In total, 40 plots (30 m x 30 m) were selected and located in the field using a GPS (Trimble GEO XT, with an estimated 10 cm accuracy). Finally, 2 or 3 subplots of 1 m x 1 m were randomly chosen within each plot for a final sample size of 100 plots. Spectral and LAI data were then collected within the subplots at the on-set, middle and end of summer (October of 2014 to March of 2015).

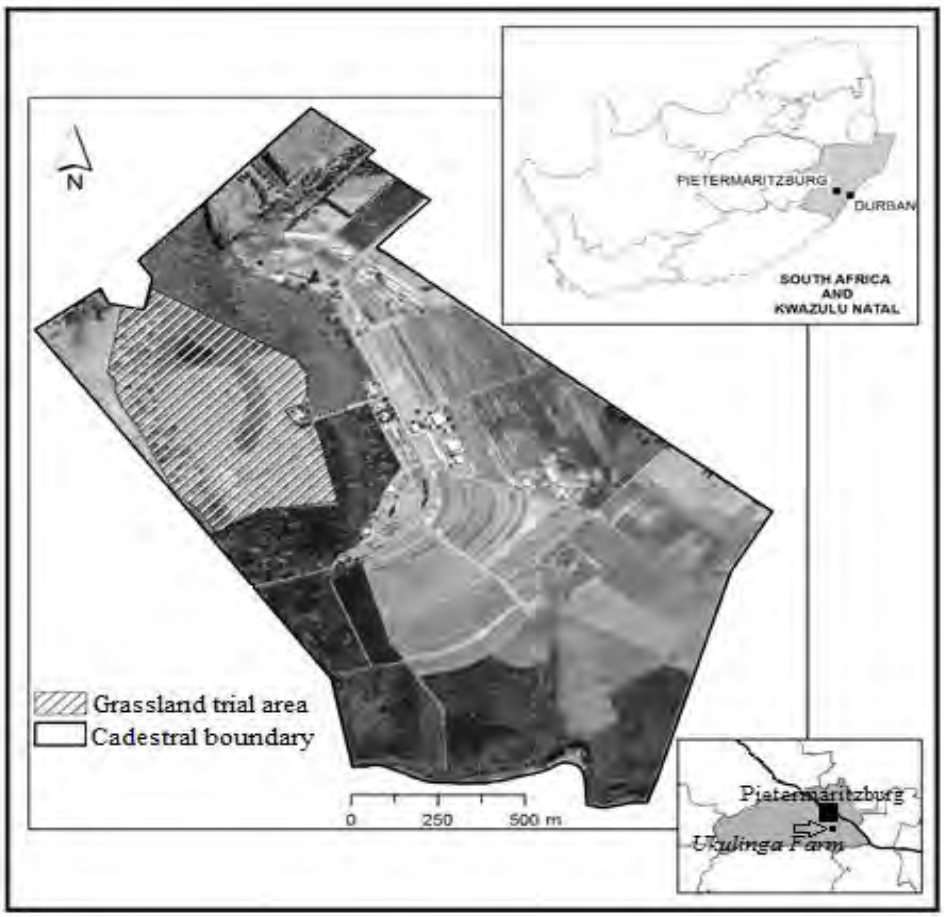

Figure 3-1 The study area within the Ukulinga research farm.

3.2.3 Leaf area index and canopy reflectance measurement

Leaf area index and canopy reflectance were measured from the canopy of most dominant grass species within each sampling subplots. To determine LAI, a 2200C Plant Canopy Analyzer was used according to the protocol described in its instruction manual (LI-COR, 2010). Leaf area index measured in this study corresponds to effective plant area index which partially account for clumping effects (Liu et al., 2012). The canopy spectra were measured using an Analytical Spectral Device (ASD), ASD FieldSpec® 3 spectrometer (Inc., Boulder, CO, USA). The ASD has a spectral resolution that ranges from 350nm to 2500nm with 1.4 nm and 2 nm sampling intervals for the ultraviolet to visible and near infrared region (350-1000 nm) and the short-wave infrared region (1000-2500 nm) respectively. Before reflectance measurement, target measurements were normalised by recording the radiance of a white standard panel coated with a barium sulphate (BaSO4) of known reflectivity (Adjorlolo et al., 2013). This was done to account for any changes in the atmospheric condition and the sun irradiance. This process was repeated every fifteen minutes. Fifteen replicates of canopy reflectance were

collected within each subplot under clear skies between 10:00 and 14:00 hrs local time as recommended by Darvishzadeh et al. (2008).

3.2.4 Data analysis

3.2.4.1 Pre-processing of hyperspectral data and selection of optimal bands

The fifteen spectra of each subplot were averaged to reduce noise in the measured canopy reflectance (Darvishzadeh et al., 2008). The resulting mean spectral data were then transformed using a first-order derivative at three nanometers (Archontaki et al., 1999; Holden and LeDrew, 1998). ViewSpecPro^{R} software was used for the computation. The mean spectra were then exported to Microsoft Excel wherein bands with noise were removed. The spectral regions between 350-399 nm, 1355-1420 nm, 1810-1940 nm, 2470-2500 nm have been reported to be noisy and were thus removed from the analysis (Rajah et al., 2015; Abdel-Rahman et al., 2014; Adjorlolo et al., 2013). On the resulting 1873 bands, interval partial least square regression (iPLSR) in forward mode was applied to select best spectral intervals for estimating LAI at different periods (onset, middle and end) within Summer. By setting the interval size to a single variable, models with 40 intervals or bands yielded better R^2 and RMSE accuracies. The results of selected bands for each period within summer are shown in Table 3-1.

3.2.4.2 Descriptive statistic, Analysis of variance (ANOVA) and Brown-Forsythe test

Skewness and kurtosis were used to determine the distribution of the collected LAI data. The test of normality was done to evaluate the suitability of LAI data using an ANOVA (Peat and Barton, 2014). One factor-ANOVA and Brown-Forsythe test ($\alpha = 0.05$) were implemented to assess the significance of LAI changes between the investigated periods within summer. Brown-Forsythe test was used to complement ANOVA because the malfunctioning ASD towards the end of summer. Brown-Forsythe test is less affected by heterogeneous sample sizes and non-distributed data (Maxwell and Delaney, 2004; Sheskin, 2003).

Table 3-1 Optimal wavebands (nm) selected by iPLSR at the beginning, mid and end of summer and for pooled data

	Visible	Red Edge (RE)	Near InfraRed (NIR)	Short Wave InfraRed (SWIR)
Beginning of summer	461, 764	-	793, 1020, 1061, 1201, 1267	1633, 1640, 1656, 1681, 1708, 1741, 1956, 1997, 2003, 2021, 2071, 2086, 2097, 2117, 2127, 2140, 2165, 2167, 2201, 2219, 2220, 2221, 2286, 2291, 2321, 2344, 2347, 2369, 2388, 2398, 2429, 2436, 2439
Middle of summer	413, 442, 443	-	995, 1132, 1134, 1174, 1240 , 1275	1693, 1944, 1947, 1951, 1959,1969, 1978, 2011, 2042, 2048, 2065, 2181, 2206, 2207, 2216, 2218, 2219, 2258, 2281, 2290, 2319, 2333, 2353, 2388, 2390, 2394, 2424, 2427 ,2434, 2437, 2450
End of summer	-	-	874, 943, 1003, 1010, 1058, 1059	1427, 1430, 1782, 1783, 1960, 1961, 1981, 1985, 1986, 2012, 2018, 2052, 2067, 2102, 2114, 2119, 2141, 2152, 2190, 2208, 2250, 2262, 2301, 2321, 2344, 2364, 2383, 2394, 2396, 2417, 2448, 2455, 2462, 2469
Combined-period	433, 489, 490, 535, 551	732, 752	957, 961, 968, 1062, 1183, 1244	1471, 1478, 1585, 1626, 1656, 1672, 1693, 1708, 1733, 1742, 1780, 2047, 2060, 2075, 2097, 2133, 2136, 2148, 2241, 2259, 2280, 2323, 2325, 2367, 2372, 2403, 2417

3.2.4.3 Statistical modelling

a) Partial least squares regression (PLSR)

Partial least squares regression (PLSR) was first introduced by Herman Wold in the 1960s to construct predictive models from multicollinear variables (Yeniay and Goktas, 2002). Firstly, it decomposes independent variables (X) into a few non-correlated latent variables or factors using information contained in the dependent variable (Y); then it regresses the new latent variables against the response variable (Cho et al., 2007; Tobias, 1995). Generally, the model that underlies PLSR consists of three steps. In the first step, independent variables (X) and response variable (Y) are decomposed as follows:

$$X = TPT + E \quad (5)$$

$$Y = UQT + F \quad (6)$$

Where T and U are respectively the matrices of scores of X and Y; P and Q stand for the matrices of loadings; E and F, errors of X and Y matrices; in the second step, the Y-scores (U) are predicted using the X-scores (T) as follows:

$$U = bT + e \quad (7)$$

Where b represents the regression coefficient and e, the error matric of the relationship between Y-scores and X-scores; in the third phase, the predicted Y -scores are used to build predictive models of response variable (Wang et al., 2011a; Tan and Li, 2008; Yeniay and Goktas, 2002).

$$Y = bTQ + G \quad (8)$$

Where G is the error matrix related to estimating Y.

b) Support vector regression (SVR)

SVR is a machine learning technique that was first introduced by Petsche et al. (1997). It was initially designed for function estimation (Smola and Schölkopf, 2004). Being a member of the support vector machines (SVMs) family, SVR can model linear or nonlinear relationship in dataset (Vapnik and Vapnik, 1998). The principle of SVR is to construct a hyperplane (s) in high- or infinite-dimensional space, which can separate quantitative estimates for regressions (Malenovsk et al., 2015).

In this study, epsilon-SVR algorithm based on the radial basis function was used to estimate LAI at investigated periods within summer. In summary, the theory of SVR algorithm is discussed as follows: given a set of calibration sample set: X= $(x_i y_i)|x \in R^n, i = 1,2, \ldots \ldots l\}$, where x_i represents the n-dimensional input vector and $y_i = \mathrm{f}(x_{i)}$, the predicted output variable. The SVR transforms the input vector into a high-dimensional feature space using a non-linear function, called kernel function$(k(x_i, x))$, in order to approximate a linear function.

$$f(x) = w.x + b \tag{9}$$

Where w, b are the weight vector and offset of the equation, respectively. With the introduction of a kernel function, the SVR equation is expressed as follows:

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) k(x_i, x) + b \tag{10}$$

Where α_i, α_i^* represent the Lagrange multipliers. The introduction of a kernel function enables SVR to model nonlinear relationship in dataset. Another feature of SVR is the ability to reduce the complexity of models (structural risk) in addition to minimizing calibration error (empirical risk) as in traditional regression methods. This makes SVR models less prone to overfitting and highly general in performance regardless the dimensionality of dataset (Axelsson et al., 2013; Yang et al., 2011). More detailed description on the theory behind SVR can be found in Smola and Schölkopf (2004) and Smola and Vapnik (1997).

3.2.4.4 Evaluation of model performance and chemometrics software

Comparison between SVR and PLSR models for different periods within summer was made on training dataset and independent test dataset using calibration measures. Data was first split into training (70 %) and test (30%) using Kennard-Stone algorithm (Kohavi, 1995). Using geometric distance, Kennard stone method first selects two samples of data that are farthest apart. Then, it adds another sample from the remaining dataset which is farthest away from the previously selected samples subset, thereby ensuring maximum coverage of the dataset (Comments and Source, 2011; May et al., 2010). After splitting the data, SVR and PLSR models were developed on same training dataset and validated using Venetian blinds cross-validation with 10 data splits. Finally, to ensure their robustness, PLSR and SVR models were tested on the remaining validation dataset (Arlot and Celisse, 2010). Root mean squared error for cross validation (RMSE) and correlation coefficients between the predicted and observed LAI (R^2) were used as calibration measures to evaluate the performance of models. RMSE has

the advantage of directly estimating the error of models. It is expressed in the same unit as original LAI units (Wang et al., 2011a). Models with better performance were indicated with smaller RMSE and greater correlation coefficient. Once the PLSR and SVR models were calibrated and validated, Variable Importance in the Projection (VIP) was generated to analyse the contribution of each variable in a model. VIP scores were considered to be applicable for both regression methods. This was because PLS-toolbox does not compute VIP scores for SVR. Also, some studies reported relatively similar variables of importance between the two algorithms (Axelsson et al., 2013; Thissen et al., 2004).

Partial least square and Support vector regression models were developed in a MATLAB version 2013 environment using PLS-toolbox (Eigenvector Research Inc.) (Wise et al., 2006). The reflectance of useful wavebands and LAI values were auto scaled before running PLSR and SVR to set all the variables on an equal basis (Wise et al., 2006; Zhang et al., 2012). PLS-toolbox suggested the best PLSR models with the optimal number of latent variables. It also automatically tuned the kernel parameter and the regularization factor of the SVR models.

3.3 Result

3.3.1 Variation in LAI and spectral data

The test for skewness and kurtosis indicated that the LAI data in the present study had a positive or less symmetric distribution (Figure 3-2). Skewness ranged between 0.856 and -0.111 and kurtosis, between 0.397 and -0.449. That made the LAI data suitable for ANOVA and Brown-Forsythe Test.

Leaf area index variation in grass species canopy was significant among the three sampling periods ($p < 0.01$). The highest mean (3.626 m^2m^{-2}) and standard deviation (1.099 m^2m^{-2}) values of LAI were observed at mid-summer. Grass species canopies at the end of summer had the second highest mean (2.015 m^2m^{-2}) and the least standard deviation (0.705 m^2m^{-2}) values of LAI. Beginning of summer had the least mean value of LAI (1.667 m^2m^{-2}) in grass species canopies, with the second least variability (0.821 m^2m^{-2}) in LAI (Figure 3-2).

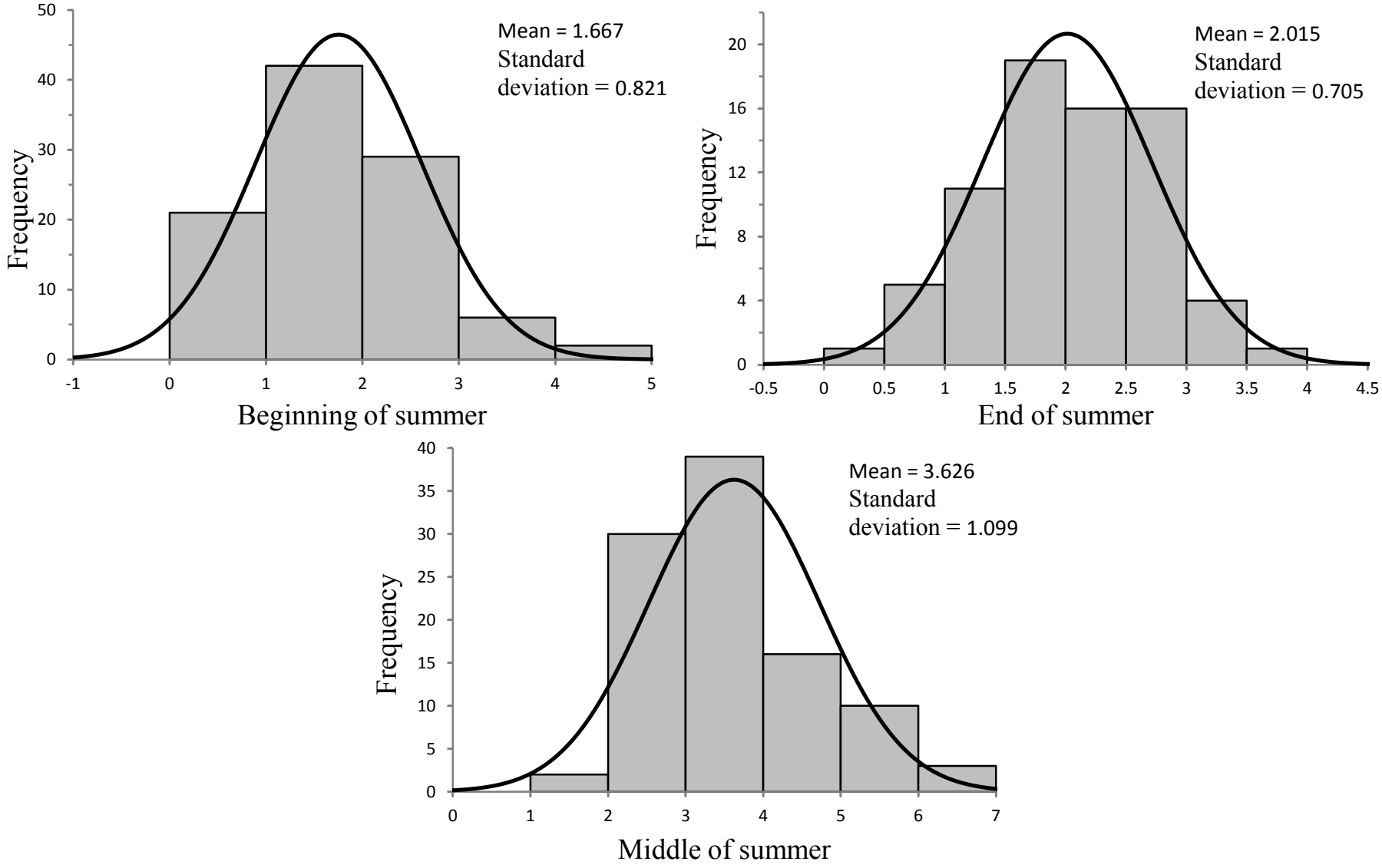

Figure 3-2 Descriptive statistics of LAI data (m^2m^{-2}) at the three sampling periods.

3.3.2 PLSR and SVR models

3.3.2.1 Evaluation of PLSR and SVR models on calibration dataset

Table 3-2 shows the performance of PLSR and SVR models on calibration dataset at each sampling period within summer. PLSR models performed better than SVR models at the beginning of summer and for pooled dataset. PLSR models yielded R^2_{cv} and RMSECV values of 0.886 and 0.311 m^2m^{-2}, respectively, at the beginning of summer and 0.831 and 0.537 m^2m^{-2}, respectively, for pooled dataset. However, at the middle and end of summer, SVR models were more accurate than PLSR models. The RMSECV and R^2_{cv} of SVR models were 0.903 and 0.351 m^2m^{-2}, respectively, at mid-summer and 0.876 and 0.272 m^2m^{-2}, respectively, at the end of summer.

The optimal number of factors to avoid overfitting in PLSR models ranged between 5 and 7. PLSR models at the end of summer displayed the highest optimal number of factors (n=7). Models at the beginning, mid-summer and all periods combined (pooled dataset) had the lowest optimal number of factors (n=5).

Table 3-2 R^2_{cv} and RMSE of PLSR (including number of factors) and iPLSR models on training dataset

Regression algorithm	R^2_{cv}	RMSECV	Number of factors
	Beginning of summer		
PLSR	0.886	0.311	5
SVR	0.871	0.335	-
	Middle of summer		
PLSR	0.894	0.368	5
SVR	0.903	0.351	-
	End of summer		
PLSR	0.862	0.286	7
SVR	0.876	0.272	-
	Combined-period (pooled dataset)		
PLSR	0.831	0.537	5
SVR	0.823	0.552	-

3.3.2.2 Variables of importance in the projection (VIP) in PLSR and SVR models

Figure 3-3 shows the importance of each waveband in the PLSR and SVR models at the investigated periods within summer. Thirteen bands had VIP scores above the significant threshold at the beginning of summer, eleven at mid-summer, thirteen at the end of summer and ten for pooled dataset.

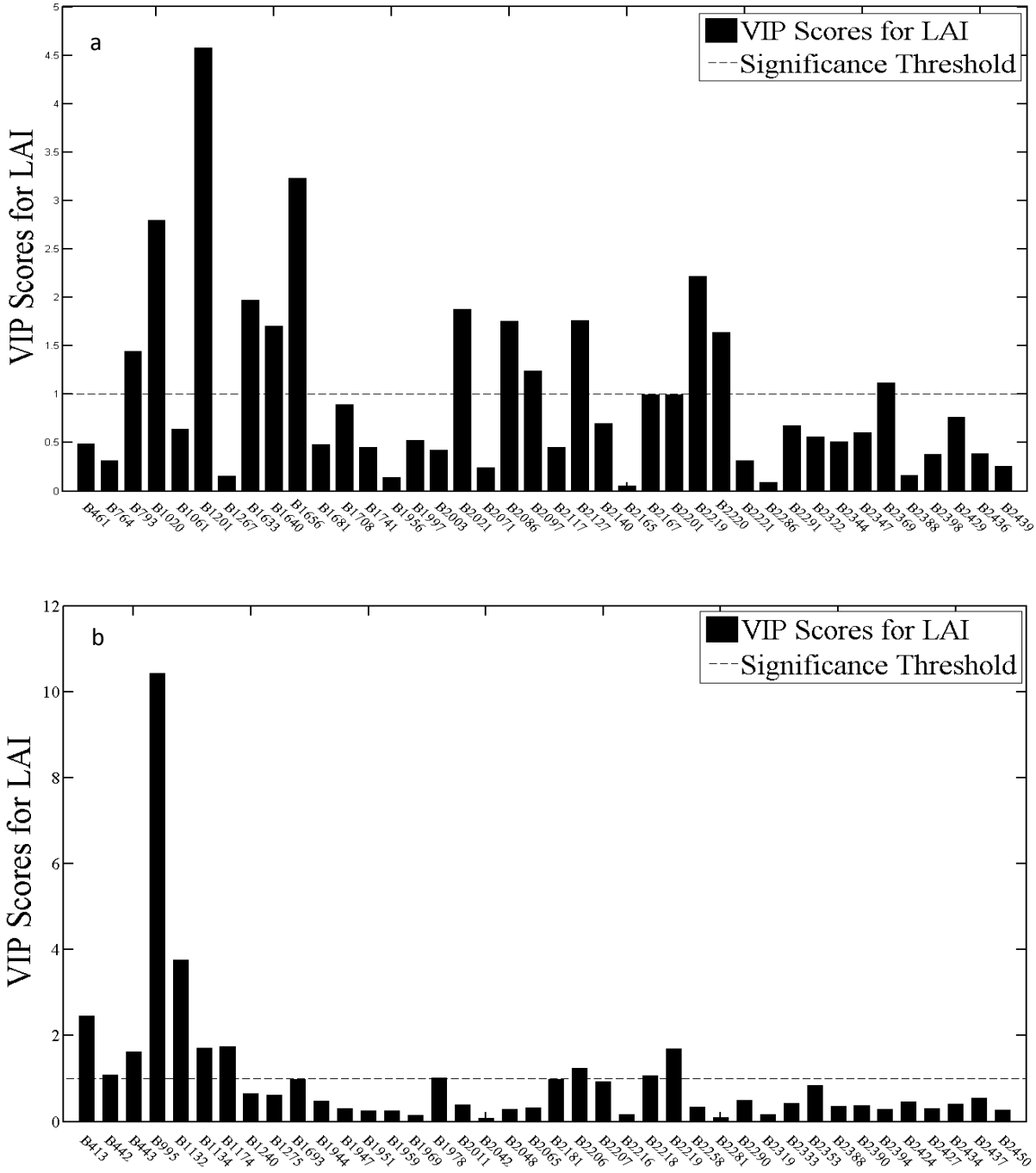

Figure 3-3a VIP scores of PLSR models at the beginning (a), mid (b) [B= Band].

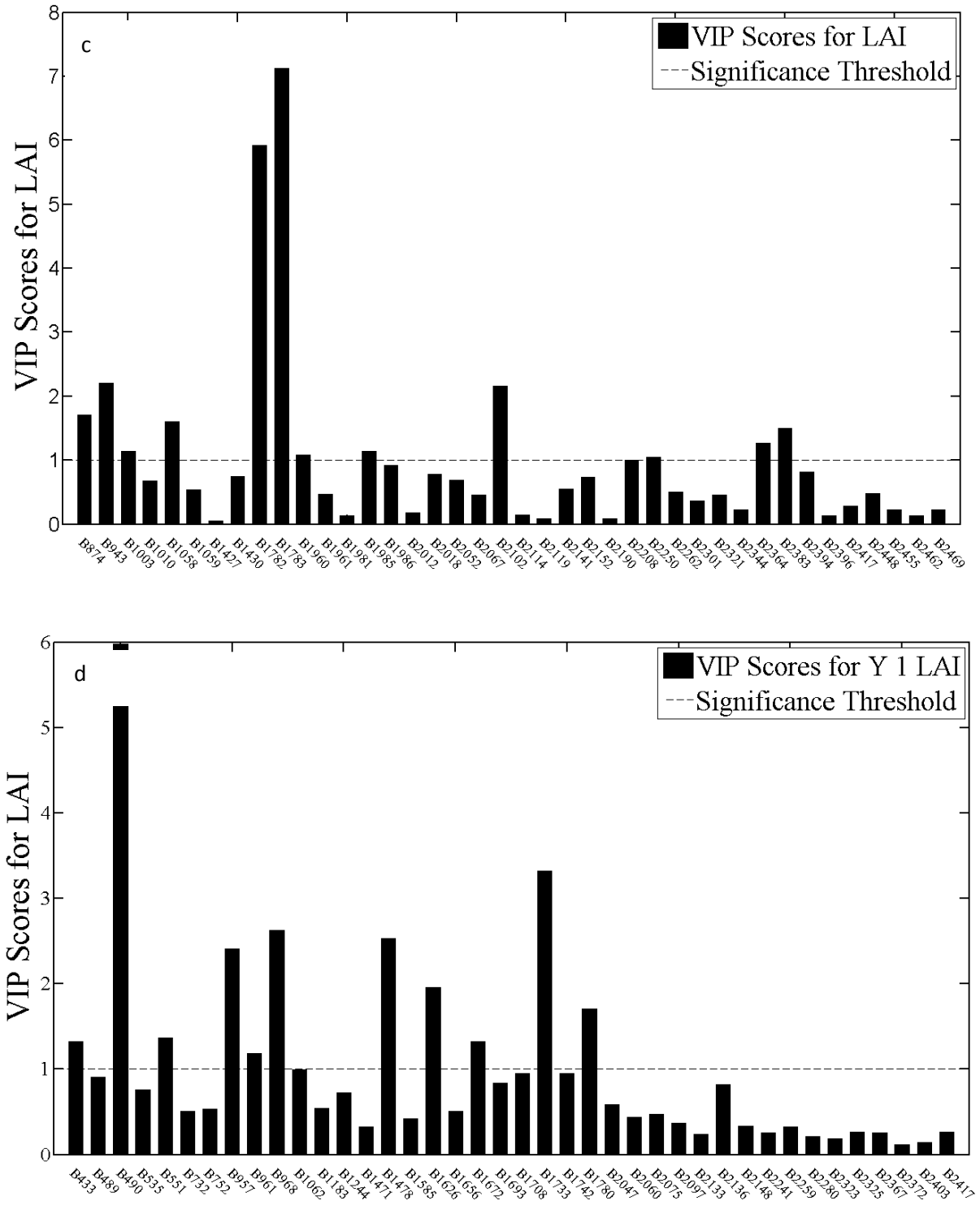

Figure 3-3b VIP scores of PLSR models at the end of summer (c) and pooled data (d) [B= Band].

3.3.3 Model validation

Figure 3-4 shows the performance of PLSR and SVR models on validation dataset. PLSR models for beginning and end of summer yielded higher accuracies than SVR models. The values of RMSEP and R^2_p of PLSR models were 0.263 m^2m^{-2} and 0.865, respectively at the beginning of summer and 0.204 m^2m^{-2} and 0.851, respectively at the end of summer. The SVR models demonstrated their superiority over PLSR models at the middle of summer and all the periods combined within summer (pooled dataset). The SVR models could respectively predict more than 90.2 % and 74 % of LAI variation in the two sampling periods. They were also characterized by lower values of RMSEP than PLSR models at those periods (0.371 m^2m^{-2} and 0.578 m^2m^{-2} versus 0.379 m^2m^{-2} and 0.580 m^2m^{-2}).

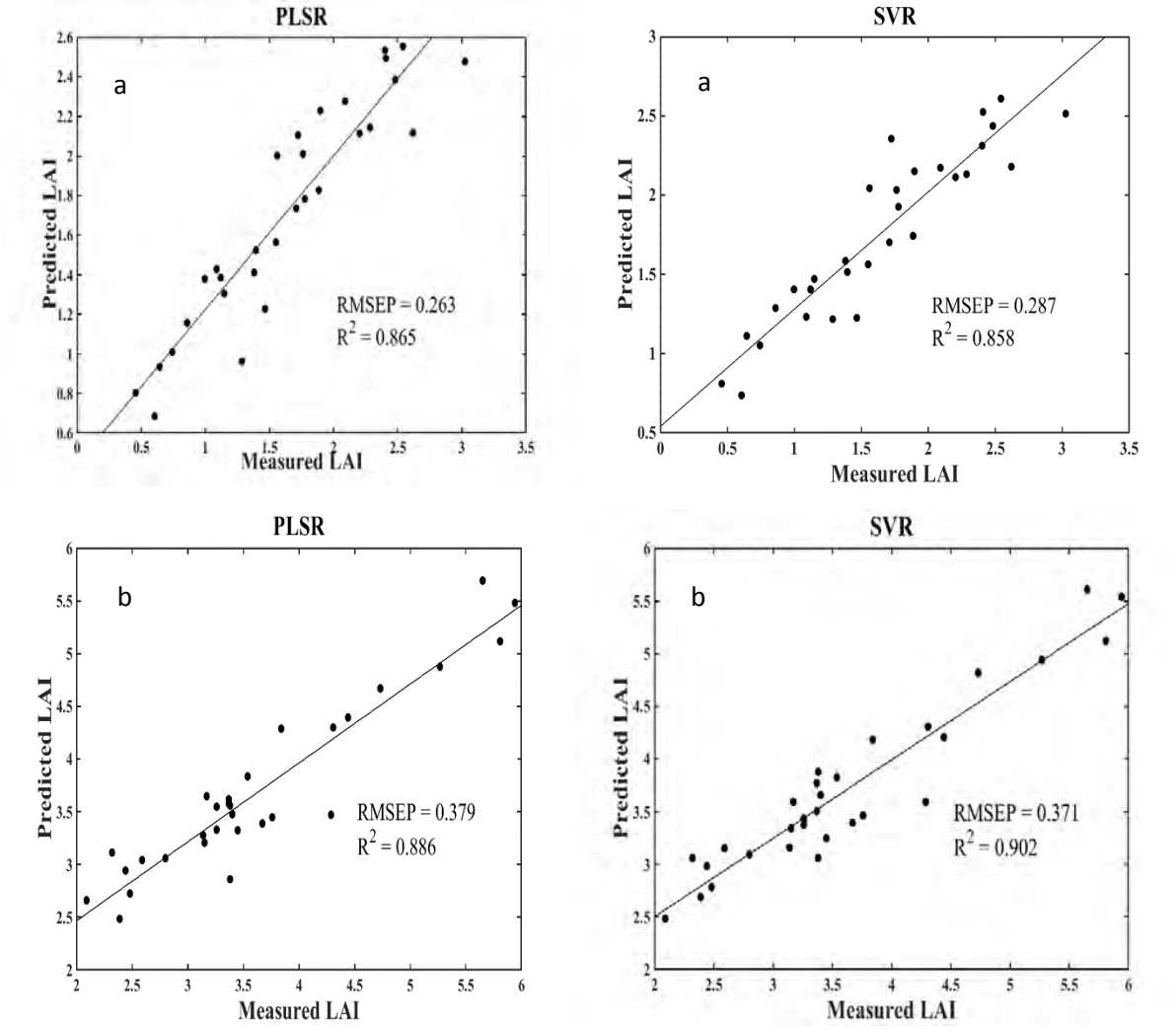

Figure 3-4a One-to-one relationship (m^2m^{-2}) between measured and predicted LAI for validating PLSR and SVR models on validation dataset, early summer (a), mid-summer (b).

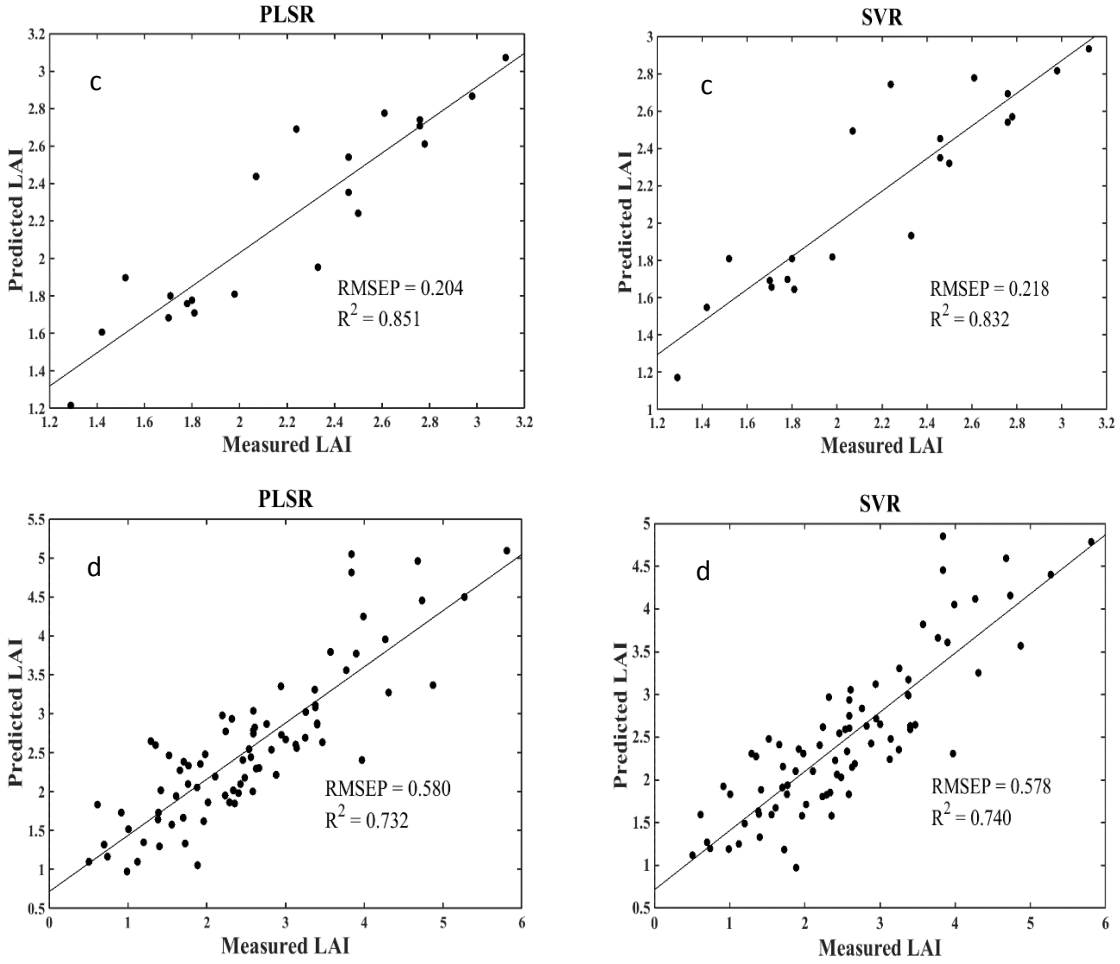

Figure 3-4b One-to-one relationship (m^2m^{-2}) between measured and predicted LAI for validating PLSR and SVR models on validation dataset, end of summer (c) and pooled data (d).

3.4 Discussion

This study sought to determine the performance of SVR and PLSR models on valuable wavebands selected by iPLSR in estimating LAI at different periods (mid, late and combined-period). The comparison of the two multivariate algorithms was based on two measures of accuracy: coefficient of determination (R^2) and the root mean square error (RMSE). Significance of the contribution of bands in models was evaluated on the basis of their VIP scores.

a) Performance of PLSR and SVR models on selected optimal bands

Overall, results of the current study are comparable to existing literature. Darvishzadeh et al. (2011) modeled LAI in heterogenious grasslands with R^2 value of 0.87 using PLSR on Hymap data subsets, but with higher prediction error (0.64 m^2m^{-2} for spectral subset) and four optimal

number of factors. Wang et al. (2011a) applied least square support vector machine (LS-SVM), an optimized variant of support vector machine, on 15 optimal hyperspectral bands to estimate LAI from rice canopy reflectance. They found that LS-SVM model could explain 90 % of LAI variance. Most of models in this study yielded higher accuracies than existing studies. For instance, Darvishzadeh et al. (2008) showed that PLSR built on useful wavebands, selected by SMLR, could predict 64 % of LAI variabilty (R_p^2) with a prediction error of 0.34 (nRMSEp). Yang et al. (2011) developed SVR models on the full spectrum of rice canopy reflectance using different pre-processing techniques (e.g. first-order derivative) and obtained model with R^2 and RMSE values of 0.8024 and 1.0496 LAI units respectively.

b) Comparison of PLSR and SVR at different sampling periods

In the three periods within summer, the results showed that PLSR models outperformed SVR models in estimating LAI at the beginning and end of summer on validation dataset. PLSR models could respectively explain 86.5 % and 85.1.2% in LAI variability against 85.8 % and 83.2 % for SVR models at the end of summer (Figure 3-4). At those two sampling periods, the mean values of LAI were low (1.667 $m^2m^{-2}$and 2.015 m^2m^{-2}), thus the saturation problem was not critical (Fava et al., 2009; Hansen and Schjoerring, 2003). This resulted to a more linear relationship between LAI and useful bands. Given that PLSR algorithm is a multivariable linear regression method and is less affected by background effects (Chen et al., 2009; Wold et al., 2001), the dataset at the beginning and end of summer might be more suitable to PLSR (Zhang et al., 2012; Li et al., 2006). These could explain the superiority of PLSR over SVR. This finding is consistent with Marabel and Alvarez-Taboada (2013) who found that PLSR on appropriate absorption features outperforms SVR in estimating less dense biomass (total aboveground biomass = 45.05 g/m^2). Shah et al. (2010) found similar results by comparing PLSR and SVR in predicting peptide drift times. They attributed the outperformance of PLSR to a strong linear relationship between the drift times and a set of properties depicting peptide structure.

However SVR models outperformed PLSR models at the middle of summer and all the periods combined (pooled dataset). Respectively, the R^2_p values were 0.886 and 0.732 for PLSR and 0.902 and 0.74 for SVR (Figure 3-4). Mid-summer coincides with the peak productivity period within summer, which is characterised by dense vegetation. According to Hansen and Schjoerring (2003), saturation problem is observed in canopies with LAI value above 2.5, which was the case of this period (mean value of LAI = 3.626 m^2m^{-2}). Various studies (Clevers

and Kooistra, 2012; Chen et al., 2009; Wu et al., 2008; Thenkabail et al., 2000; Huete et al., 1997) have associated non-linearity in dataset and reflectance saturation. Non-linearity and spectral noise are efficiently dealt by SVR (Üstün et al., 2005). Therefore, at mid-summer, SVR may be the most robust algorithm in comparison to PLSR, a finding consistent with Axelsson et al. (2013) who compared different variants of SVR and PLSR in mapping mangrove foliar biochemicals. Thissen et al. (2004) also found that SVR outperformed PLSR in estimating ethanol, water, and iso-propanol concentrations using near infrared spectra, which were affected by nonlinear temperature-induced variation.

c) The model's VIP bands

Most of the selected VIP bands fell in the near infrared (NIR) and shortwave infrared (SWIR) sections of the electromagnetism spectrum (Table 3-1). In previous studies, the two spectral regions have been reported to contain bands that correlate to LAI on heterogeneous grasslands (Darvishzadeh et al., 2011; Fava et al., 2009; Darvishzadeh et al., 2008). At the end of summer, only bands in the NIR and SWIR region were selected. This could be due to canopy senescence, causing a decline of the amount of chlorophyll. According Zhao et al. (2007), NIR and SWIR reflectance becomes more important at this growth stage. In the pooled dataset (all combined periods within summer), results showed that bands in the visible spectral region noticeably contributed to models. However, the contribution of bands in the near infrared region was significant. These findings are consistent with Fava et al. (2009) who found a strong correlation between visible reflectance and green biomass, LAI and nitrogen. In the same study near infrared reflectance exhibited a weak correlation within autumn and spring. Regardless of the sampling period, some of the selected bands (e.g. 793 nm 1201 nm, 1058 nm, 633 nm, 1640 nm, 2097 nm, 442 nm, 443 nm, 433 nm,1672 nm) that had a significant contribution in models are similar (within ±12 nm) to known bands for predicting LAI in heterogeneous or homogeneous canopies (Darvishzadeh et al., 2011; Darvishzadeh et al., 2008; Wang et al., 2008; Gong et al., 2003).

This study has demonstrated that a proper variable selection algorithm coupled with PLSR and SVR on hyperspectral data can improve estimation of biochemical and biophysical variables in heterogeneous grass canopies. Temporal variability in heterogeneous grassland have been an impediment to remotely sensed data in estimating vegetation characteristics due to an assortment of grass species and soil background (Darvishzadeh et al., 2008; Röder et al., 2007). Results in this study have potential for broader spatial scaling using airborne or satellite sensors

by adopting selected optimal bands for image acquisition. This would immensely contribute to knowledge on grassland condition, thus enhance rangeland management. Moreover, wavebands of high VIP scores may be useful in developing new generation of imaging sensors. For example, sensors that specialise in quantifying LAI would focus on NIR and SWIR spectral regions. However, it may still be necessary to investigate the efficacy of some other useful variable algorithms (e.g. genetic algorithms) in combination with PLSR or SVR, and their respective variants, for effective spectral transformations and therefore more accurate models for heterogeneous grassland condition assessment (Li et al., 2011; Wang et al., 2011b; Kawamura et al., 2010; Üstün et al., 2005; Yao and Tian, 2003).

3.5 Conclusion

Major conclusions in this study can be summarized as:

- PLSR and SVR can successfully be used to simplify the relationship between LAI and useful spectral bands on a heterogeneous grassland
- PLSR models are suited for modelling LAI at relatively sparse vegetation, beginning and end of summer were the most ideal periods.
- On denser grassland, SVR outperformed PLSR. This can be attributed to its ability to quantify nonlinear relationships in samples.
- Most of bands that significantly contributed in modelling LAI are located in the NIR and SWIR regions.

Overall, this study has endeavoured to compare the performance of PLSR and SVR on multi-temporal variation of LAI at the canopy level in a tropical grassland. It was found that the performance of either multivariate regression method depends on the phenological stage. The findings of this study shed more light on the use of SVR in estimating biophysical variables in a heterogeneous grassland at different temporal scales. As only one season was used in this study, we recommend these findings be validated using dataset from similar sites over several years.

Chapter 4

Synthesis

This study focused on exploring the capability of in-situ multi-temporal hyperspectral data and regression techniques in estimating leaf area index (LAI) on heterogeneous tropical grassland. In this chapter, aims and respective objectives, which were set out in the first chapter are reviewed against the findings. Major conclusions and recommendations for future research are also highlighted.

4.1 First aim and its objectives

Aim: To investigate the potential of iPLSR on hyperspectral data in estimating LAI on a heterogeneous tropical grassland.

Objectives: -To compare PLSR and iPLSR on hyperspectral data
-To evaluate the robustness of PLSR and iPLSR models at three sampling periods (i.e. onset, mid and end) and pooled reflectance data during summer)

Hyperspectral dataset are known to suffer from multicollinearity and high degree of redundancy. Partial least square regression was introduced to overcome these shortcomings. However, due to heterogeneous grass canopies and temporal variability of grass canopies in different growing seasons, the performance of PLSR is hampered by mixture of canopy reflectance. Darvishzadeh et al. (2008) applied PLSR on hyperspectral dataset to model LAI in heterogeneous grassland. They concluded that PLSR models yielded moderate accuracies. Therefore, this section aimed at investigating the potential of iPLSR in estimating LAI using hyperspectral data at different periods within summer. iPLSR is a variant of PLSR that reduces data dimensionality to optimal bands. Results showed that iPLSR outperformed PLSR at all the sampling periods. The iPLSR models were more accurate at the beginning and end of summer. Forty wavebands, located in the near and shortwave infrared regions, were selected by iPLSR. The superior performance of iPLSR over PLSR may be attributed to its ability to remove uninformative wavebands from hyperspectral data and retain those which are useful for LAI estimation. Reduction of data dimensionality has been proven to be valuable in improving model estimations. The reason early and end-summer models performed better than

mid-summer model might be explained by the absence of saturation problem, which affects model performance, observed at those sampling periods.

4.2 The second aim and its objectives

Aim: To compare the performance of PLSR and SVR using optimal bands selected by iPLSR in estimating LAI on a heterogeneous tropical grassland.

Objectives: -To compare the performance of PLSR and SVR on iPLSR selected optimal bands at three sampling periods within summer (early, mid, late).

-To identify wavebands with VIP scores above the significant threshold.

Canopy heterogeneity and temporal variability of grasslands at different periods within growing season are two major reasons that have hindered the performance of remote sensing data in estimating LAI (Shen et al., 2014). Saturation problem which is effective at peak productivity accentuate this hindrance by creating nonlinear relationship between canopy reflectance and biophysical variables such as LAI and biomass. In this study, partial least square regression (PLSR), a linear regression method, and support vector regression (SVR), a nonlinear regression, were tested and compared on optimal wavebands at the beginning, middle and end of summer. Overall, results showed that PLSR and SVR models developed in this study were more accurate than models developed in previous studies undertaken in homogeneous (rice canopies) or heterogeneous (heterogeneous grass canopies) environment. Results also showed that PLSR produced best models at low canopy density (LAI < 2.5 m^2m^{-2}) (at the beginning and end of summer). This could be explained by the fact that at low LAI, the saturation problem was absent. That resulted to a linear relationship between LAI and useful bands. As PLSR algorithm is a multivariable linear regression method and is less affected by background reflectance, the dataset turned out to be more suitable to PLSR. However, when vegetation became dense (LAI > 2.5 m^2m^{-2}) or dataset pooled, SVR outperformed PLSR. In this case, saturation in reflectance was present. So, relationship between LAI and useful bands was nonlinear. SVR is known to efficiently deal with non-linear dataset. That could be the reason of the superior performance of SVR over PLSR. VIP analysis revealed that most of wavebands that significantly contributed in SVR and PLSR models were located in the NIR and SWIR regions. The occurrence of wavebands in the two spectral regions depended on the phenological stage.

4.3 Conclusion and recommendations

The current study aimed at investigating the potential of iPLSR on hyperspectral data and comparing PLSR and SVR using optimal bands selected by iPLSR in estimating LAI on a heterogeneous tropical grassland. The conclusions are consolidated on the basis of research questions asked raised.

To which extent can LAI be estimated using ground-based multi-temporal hyperspectral data and regression techniques in tropical grasslands during the growing season?

Partial least square regression models poorly estimated LAI. When data dimensionality was reduced to optimal bands using iPLSR, models, accuracies improved considerably. iPLSR models could explain more than 80 % of new LAI data for all the sampling periods, including all the periods combined. Furthermore, model accuracies moderately improved using SVR on denser canopies, which are characterized by the presence of nonlinearity in dataset, and at all the periods combined (pooled dataset).

What is the best period (s) and regression techniques for LAI estimation?

The best periods for estimating LAI among the investigated periods within summer are at the beginning and the end. The two periods were characterized by low LAI means. Previous studies obtained similar results at low vegetation density. Literature reported that PLSR could deal with noise encountered in lower density canopies.

What are the most optimal bands for LAI estimation?

Interval partial least square regression was applied on hyperspectral data to select useful bands for LAI estimation. Forty bands were useful at each sampling period within summer. Majority of selected bands were within the NIR and SWIR spectral region. This finding was confirmed by preceding research undertaken on either heterogeneous or homogeneous landscape. VIP analysis on the contribution of optimal bands selected using iPLSR also showed that the most influential bands were located in the NIR and SWIR. Further light shed by VIP analysis was that phenological stage drove the influence of bands in different portions of the electromagnetic spectrum.

In summary, an appropriate variable coupled with multivariate regression techniques such as PLSR or SVR can reliably model LAI on heterogeneous tropical grassland using in situ multi-temporal hyperspectral data. Density in vegetation canopies should determine the adoption of either regression method. Findings in this study provide a practical insight in mapping LAI on heterogeneous grasslands at regional scale using airborne or space borne hyperspectral data such as MERIS, HYPERION, MODIS and CHRIS. This would enhance decision and policy making on grassland management, thereby mitigating land degradation observed in South Africa and indeed the world. For future research, the methods used in this work should be extended to other biophysical and biochemical variables of canopy characteristics in heterogeneous grasslands to validate the findings. The performance of other variable selection algorithms in concert with relevant multivariate regression methods and reflectance transformations should also be pursued in future endeavours.

References

Abdel-Rahman, E. M., Mutanga, O., Odindi, J., Adam, E., Odindo, A., and 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.

Adler, P., Raff, D., and Lauenroth, W. (2001). The effect of grazing on the spatial heterogeneity of vegetation. *Oecologia*, 128(4), 465-479.

Adjorlolo, C., Mutanga, O., Cho, M. A., and Ismail, R. (2013). Spectral resampling based on user-defined inter-band correlation filter: C3 and C4 grass species classification. *International Journal of Applied Earth Observation and Geoinformation*, 21, 535-544.

Andersen, C. M., and Bro, R. (2010). Variable selection in regression—a tutorial. *Journal of Chemometrics*, 24(11-12), 728-737.

Archontaki, H. A., Atamian, K., Panderi, I. E., and Gikas, E. E. (1999). Kinetic study on the acidic hydrolysis of lorazepam by a zero-crossing first-order derivative UV-spectrophotometric technique. *Talanta*, 48(3), 685-693.

Arlot, S., and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, 4, 40-79.

Atzberger, C., Jarmer, T., Schlerf, M., Kötz, B., and Werner, W. (2004). Spectroradiometric determination of wheat bio-physical variables: comparison of different empirical-statistical approaches. *In Remote Sensing in Transitions, Proc. 23rd EARSeL symposium, Belgium* (pp. 2-5).

Atzberger, C., Guérif, M., Baret, F., and Werner, W. (2010). Comparative analysis of three chemometric techniques for the spectroradiometric assessment of canopy chlorophyll content in winter wheat. *Computers and Electronics in Agriculture*, 73(2), 165-173.

Axelsson, C., Skidmore, A. K., Schlerf, M., Fauzi, A., and Verhoef, W. (2013). Hyperspectral analysis of mangrove foliar chemistry using PLSR and support vector regression. *International Journal of Remote Sensing*, 34(5), 1724-1743.

Balabin, R. M., and Smirnov, S. V. (2011). Variable selection in near-infrared spectroscopy: benchmarking of feature selection methods on biodiesel data. *Analytica chimica acta*, 692(1), 63-72.

Banskota, A. (2006). Estimating leaf area index of salt marsh vegetation using airborne hyperspectral data.

Borin, A., and Poppi, R. J. (2005). Application of mid infrared spectroscopy and iPLS for the quantification of contaminants in lubricating oil. *Vibrational Spectroscopy*, 37(1), 27-32.

Bréda, N. J. (2003). Ground-based measurements of leaf area index: a review of methods, instruments and current controversies. *Journal of experimental botany*, 54(392), 2403-2417.

Broge, N. H., and Mortensen, J. V. (2002). Deriving green crop area index and canopy chlorophyll density of winter wheat from spectral reflectance data. *Remote sensing of environment*, 81(1), 45-57.

Brown, L., Chen, J. M., Leblanc, S. G., and Cihlar, J. (2000). A shortwave infrared modification to the simple ratio for LAI retrieval in boreal forests: An image and model analysis. *Remote sensing of environment*, 71(1), 16-25.

Bulcock, H. H., and Jewitt, G. P. W. (2010). Spatial mapping of leaf area index using hyperspectral remote sensing for hydrological applications with a particular focus on canopy interception. *Hydrology and Earth System Sciences*, 14 (2), 383-392.

Chason, J. W., Baldocchi, D. D., and Huston, M. A. (1991). A comparison of direct and indirect methods for estimating forest canopy leaf area. *Agricultural and Forest Meteorology*, 57(1), 107-128.

Chen, J., Gu, S., Shen, M., Tang, Y., and Matsushita, B. (2009). Estimating aboveground biomass of grassland having a high canopy cover: an exploratory analysis of in situ hyperspectral data. *International Journal of Remote Sensing*, 30 (24), 6497-6517.

Chen, J. M., and Cihlar, J. (1996). Retrieving leaf area index of boreal conifer forests using Landsat TM images. *Remote sensing of Environment*, 55(2), 153-162.

Cho, M. A., Skidmore, A., Corsi, F., Van Wieren, S. E., and 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(4), 414-424.

Chung, D., and Keles, S. (2010). Sparse partial least squares classification for high dimensional data. *Statistical applications in genetics and molecular biology*, 9(1).

Clevers, J. G., and Kooistra, L. (2012). Using hyperspectral remote sensing data for retrieving canopy chlorophyll and nitrogen content. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 5(2), 574-583.

Comments, S., and Source, M. (2011). American Society for Quality. *Quality*, 15(1), 661-675.

Darvishzadeh, R., Atzberger, C., Skidmore, A., and Schlerf, M. (2011). Mapping grassland leaf area index with airborne hyperspectral imagery: A comparison study of statistical approaches and inversion of radiative transfer models. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(6), 894-906.

Darvishzadeh, R., Skidmore, A., Schlerf, M., Atzberger, C., Corsi, F., and Cho, M. (2008). LAI and chlorophyll estimation for a heterogeneous grassland using hyperspectral measurements. *ISPRS Journal of Photogrammetry and Remote Sensing*, 63(4), 409-426.

de Lira, L. F. B., de Albuquerque, M. S., Pacheco, J. G. A., Fonseca, T. M., de Siqueira Cavalcanti, E. H., Stragevitch, L., and Pimentel, M. F. (2010). Infrared spectroscopy and multivariate calibration to monitor stability quality parameters of biodiesel. *Microchemical journal*, 96(1), 126-131.

Delegido, J., Verrelst, J., Meza, C. M., Rivera, J. P., Alonso, L., and Moreno, J. (2013). A red-edge spectral index for remote sensing estimation of green LAI over agroecosystems. *European Journal of Agronomy*, 46, 42-52.

Doraiswamy, P. C., Hatfield, J. L., Jackson, T. J., Akhmedov, B., Prueger, J., and Stern, A. (2004). Crop condition and yield simulations using Landsat and MODIS. *Remote sensing of environment*, 92(4), 548-559.

Dorigo, W. A., Zurita-Milla, R., de Wit, A. J., Brazile, J., Singh, R., and Schaepman, M. E. (2007). A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling. *International journal of applied earth observation and geoinformation*, 9(2), 165-193.

Everson, C. S., Mengistu, M. G., and Gush, M. B. (2013). A field assessment of the agronomic performance and water use of Jatropha curcas in South Africa. *Biomass and Bioenergy*, 59, 59-69.

Fava, F., Colombo, R., Bocchi, S., Meroni, M., Sitzia, M., Fois, N., and Zucca, C. (2009). Identification of hyperspectral vegetation indices for Mediterranean pasture characterization. *International Journal of Applied Earth Observation and Geoinformation*, 11(4), 233-243.

Filzmoser, P., Gschwandtner, M., and Todorov, V. (2012). Review of sparse methods in regression and classification with application to chemometrics.*Journal of Chemometrics*, 26(3-4), 42-51.

Ghebrehiwot, H. M., Kulkarni, M. G., Szalai, G., Soós, V., Balázs, E., and Van Staden, J. (2013). Karrikinolide residues in grassland soils following fire: Implications on germination activity. *South African Journal of Botany*, 88, 419-424.

Gong, P., Pu, R., Biging, G. S., and Larrieu, M. R. (2003). Estimation of forest leaf area index using vegetation indices derived from Hyperion hyperspectral data. *Geoscience and Remote Sensing, IEEE Transactions on*, 41(6), 1355-1362.

Gray, J., and Song, C. (2012). Mapping leaf area index using spatial, spectral, and temporal information from multiple sensors. *Remote Sensing of Environment*, 119, 173-183.

Hansen, P. M., and Schjoerring, J. K. (2003). Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indices and partial least squares regression. *Remote sensing of environment*, 86(4), 542-553.

He, Y., Guo, X., and Wilmshurst, J. (2006). Studying mixed grassland ecosystems I: suitable hyperspectral vegetation indices. *Canadian Journal of Remote Sensing*, 32(2), 98-107.

He, Y., Guo, X., and Wilmshurst, J. F. (2007). Comparison of different methods for measuring leaf area index in a mixed grassland. *Canadian Journal of Plant Science*, 87(4), 803-813.

Herrmann, I., Pimstein, A., Karnieli, A., Cohen, Y., Alchanatis, V., and Bonfil, D. J. (2011). LAI assessment of wheat and potato crops by VENμS and Sentinel-2 bands. *Remote Sensing of Environment*, *115*(8), 2141-2151.

Holden, H., and LeDrew, E. (1998). Spectral discrimination of healthy and non-healthy corals based on cluster analysis, principal components analysis, and derivative spectroscopy. *Remote sensing of environment*, *65*(2), 217-224.

Huete, A. R., Liu, H., and Van Leeuwen, W. J. (1997). The use of vegetation indices in forested regions: issues of linearity and saturation. InGeoscience and Remote Sensing, 1997. IGARSS'97. *Remote Sensing-A Scientific Vision for Sustainable Development, 1997 IEEE International* (Vol. 4, pp. 1966-1968). IEEE.

Jensen, R. R., Hardin, P. J., Bekker, M., Farnes, D. S., Lulla, V., and Hardin, A. (2009). Modeling urban leaf area index with AISA+ hyperspectral data. *Applied Geography*, 29(3), 320-332.

Jonckheere, I., Fleck, S., Nackaerts, K., Muys, B., Coppin, P., Weiss, M., and Baret, F. (2004). Review of methods for in situ leaf area index determination: Part I. Theories, sensors and hemispherical photography. *Agricultural and forest meteorology*, 121(1), 19-35.

Karaman, İ., Qannari, E. M., Martens, H., Hedemann, M. S., Knudsen, K. E. B., and Kohler, A. (2013). Comparison of Sparse and Jack-knife partial least squares regression methods for variable selection. *Chemometrics and Intelligent Laboratory Systems*, 122, 65-77.

Kawamura, K., Watanabe, N., Sakanoue, S., Lee, H. J., Inoue, Y., and Odagawa, S. (2010). Testing genetic algorithm as a tool to select relevant wavebands from field hyperspectral

data for estimating pasture mass and quality in a mixed sown pasture using partial least squares regression. *Grassland science*, 56 (4), 205-216.

Kim, M. S., Daughtry, C. S. T., Chappelle, E. W., McMurtrey, J. E., and Walthall, C. L. (1994). The use of high spectral resolution bands for estimating absorbed photosynthetically active radiation (A par).

Knox, N. M., Skidmore, A. K., Prins, H. H., Heitkönig, I. M., Slotow, R., van der Waal, C., and de Boer, W. F. (2012). Remote sensing of forage nutrients: Combining ecological and spectral absorption feature data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 72, 27-35.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, No. 2, pp. 1137-1145).

Lee, K. S., Cohen, W. B., Kennedy, R. E., Maiersperger, T. K., and Gower, S. T. (2004). Hyperspectral versus multispectral data for estimating leaf area index in four different biomes. *Remote Sensing of Environment*, 91(3), 508-520.

Leuschner, C., Voß, S., Foetzki, A., and Clases, Y. (2006). Variation in leaf area index and stand leaf mass of European beech across gradients of soil acidity and precipitation. *Plant Ecology*, 186(2), 247-258.

Levy, P. E., and Jarvis, P. G. (1999). Direct and indirect measurements of LAI in millet and fallow vegetation in HAPEX-Sahel. *Agricultural and Forest Meteorology*, 97(3), 199-212.

Li, H., Liang, Y., and Xu, Q. (2009). Support vector machines and its applications in chemistry. *Chemometrics and Intelligent Laboratory Systems*, 95(2), 188-198.

Li, S., Wu, H., Wan, D., and Zhu, J. (2011). An effective feature selection method for hyperspectral image classification based on genetic algorithm and support vector machine. *Knowledge-Based Systems*, 24(1), 40-48.

Li, X., Zhang, Y., Bao, Y., Luo, J., Jin, X., Xu, X., and Yang, G. (2014). Exploring the best hyperspectral features for LAI estimation using partial least squares regression. *Remote Sensing*, 6(7), 6221-6241.

Li, Y., Zhu, Y., Tian, Y., and Cao, W. (2006). [Quantitative relationships between leaf area index and canopy reflectance spectra of wheat]. *Ying yong sheng tai xue bao= The journal of applied ecology/Zhongguo sheng tai xue xue hui, Zhongguo ke xue yuan Shenyang ying yong sheng tai yan jiu suo zhu ban*, 17(8), 1443-1447.

Li, Z. (2010). *Improved leaf area index estimation by considering both temporal and spatial variations* (Doctoral dissertation, University of Saskatchewan Saskatoon).

LI-COR Inc. *LAI-2200 plant canopy analyzer, introduction manual*. LI-COR, Inc., Lincoln (2010), p. 209.

Liu, J. (2014). Developing a soft sensor based on sparse partial least squares with variable selection. *Journal of Process Control*, 24(7), 1046-1056.

Liu, J., Pattey, E., and Jégo, G. (2012). Assessment of vegetation indices for regional crop green LAI estimation from Landsat images over multiple growing seasons. *Remote Sensing of Environment*, 123, 347-358.

Malenovský, Z., Turnbull, J. D., Lucieer, A., and Robinson, S. A. (2015). Antarctic moss stress assessment based on chlorophyll content and leaf density retrieved from imaging spectroscopy data. *New Phytologist*, 208(2), 608-624.

Mao, H., Gao, H., Zhang, X., and Kumi, F. (2015). Nondestructive measurement of total nitrogen in lettuce by integrating spectroscopy and computer vision. *Scientia Horticulturae*, 184, 1-7.

Marabel, M., and Alvarez-Taboada, F. (2013). Spectroscopic determination of aboveground biomass in grasslands using spectral transformations, support vector machine and partial least squares regression. *Sensors*, 13(8), 10027-10051.

Maxwell, S. E., and Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (Vol. 1). Psychology Press.

May, R. J., Maier, H. R., and Dandy, G. C. (2010). Data splitting for artificial neural networks using SOM-based stratified sampling. *Neural Networks*, 23(2), 283-294.

McNaughton, S. J. (1988). Mineral nutrition and spatial concentrations of African ungulates.

Mehmood, T., Liland, K. H., Snipen, L., and Sæbø, S. (2012). A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 118, 62-69.

Mills, A. J., and Fey, M. V. (2004). Frequent fires intensify soil crusting: physicochemical feedback in the pedoderm of long-term burn experiments in South Africa. *Geoderm*a, 121(1), 45-64.

Mountrakis, G., Im, J., and Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66 (3), 247-259.

Navea, S., Tauler, R., and de Juan, A. (2005). Application of the local regression method interval partial least-squares to the elucidation of protein secondary structure. *Analytical biochemistry*, 336(2), 231-242.

Nguyen, H. T., and Lee, B. W. (2006). Assessment of rice leaf growth and nitrogen status by hyperspectral canopy reflectance and partial least square regression. *European Journal of Agronomy*, 24(4), 349-356.

Norgaard, L., Saudland, A., Wagner, J., Nielsen, J. P., Munck, L., and Engelsen, S. B. (2000). Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. *Applied Spectroscopy*, 54(3), 413-419.

Peat, J., and Barton, B. (2014). Medical Statistics: A Guide to SPSS, Data Analysis and Critical Appraisal. Vol. 4. *Wiley & Sons*.

Petsche, T., Mozer, M. C., and Jordan, M. I. (Eds.). (1997). *Advances in Neural Information Processing Systems 9: Proceedings of the 1996 Conference*. MIT Press.

Pfeifer, M., Gonsamo, A., Disney, M., Pellikka, P., and Marchant, R. (2012). Leaf area index for biomes of the Eastern Arc Mountains: Landsat and SPOT observations along precipitation and altitude gradients. *Remote Sensing of Environment*, 118, 103-115.

Prins, H. H. T., and Beekman, J. H. (1989). A balanced diet as a goal for grazing: the food of the Manyara buffalo. *African Journal of Ecology*, 27(3), 241-259.

Pullanagari, R. R., Yule, I. J., Tuohy, M. P., Hedley, M. J., Dynes, R. A., and King, W. M. (2012). In-field hyperspectral proximal sensing for estimating quality parameters of mixed pasture. *Precision Agriculture*, 13(3), 351-369.

Rajah, P., Odindi, J., Abdel-Rahman, E. M., Mutanga, O., and Modi, A. (2015). Varietal discrimination of common dry bean (*Phaseolus vulgaris* L.) grown under different watering regimes using multitemporal hyperspectral data. *Journal of Applied Remote Sensing*, *9*(1), 096050-096050.

Ramoelo, A., Skidmore, A. K., Cho, M. A., Mathieu, R., Heitkönig, I. M. A., Dudeni-Tlhone, N., and Prins, H. H. T. (2013). Non-linear partial least square regression increases the estimation accuracy of grass nitrogen and phosphorus using in situ hyperspectral and environmental data. *ISPRS journal of photogrammetry and remote sensing*, *82*, 27-40.

Röder, A., Kuemmerle, T., Hill, J., Papanastasis, V. P., and Tsiourlis, G. M. (2007). Adaptation of a grazing gradient concept to heterogeneous Mediterranean rangelands using cost surface modelling. *Ecological Modelling*, 204(3), 387-398.

Ryu, Y., Sonnentag, O., Nilson, T., Vargas, R., Kobayashi, H., Wenk, R., and Baldocchi, D. D. (2010). How to quantify tree leaf area index in an open savanna ecosystem: a multi-instrument and multi-model approach. *Agricultural and Forest Meteorology*, 150(1), 63-76.

Schlerf, M., Atzberger, C., and Hill, J. (2005). Remote sensing of forest biophysical variables using HyMap imaging spectrometer data. *Remote Sensing of Environment*, 95(2), 177-194.

Shackleton, S. E., Shackleton, C. M., Netshiluvhi, T. R., Geach, B. S., Ballance, A., and Fairbanks, D. H. K. (2002). Use patterns and value of savanna resources in three rural villages in South Africa. *Economic Botany*, 56(2), 130-146.

Shah, A. R., Agarwal, K., Baker, E. S., Singhal, M., Mayampurath, A. M., Ibrahim, Y. M., and Smith, R. D. (2010). Machine learning based prediction for peptide drift times in ion mobility spectrometry. *Bioinformatics*, 26(13), 1601-1607.

Shen, L., Li, Z., and Guo, X. (2014). Remote Sensing of Leaf Area Index (LAI) and a Spatiotemporally Parameterized Model for Mixed Grasslands.International *Journal of Applied*, 4(1).

Sheskin, D. J. (2003). *Handbook of parametric and nonparametric statistical procedures*. crc Press.

Si, Y., Schlerf, M., Zurita-Milla, R., Skidmore, A., and Wang, T. (2012). Mapping spatio-temporal variation of grassland quantity and quality using MERIS data and the PROSAIL model. *Remote Sensing of Environment*, 121, 415-425.

Smola, A. J., and Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, 14(3), 199-222.

Smola, A., and Vapnik, V. (1997). Support vector regression machines. *Advances in neural information processing systems*, 9, 155-161.

Snyman, H. A. (1999). Short-term effects of soil water, defoliation and rangeland condition on productivity of a semi-arid rangeland in South Africa. *Journal of Arid Environments*, 43(1), 47-62.

Sousa, A. G., Ahl, L. I., Pedersen, H. L., Fangel, J. U., Sørensen, S. O., and Willats, W. G. (2015). A multivariate approach for high throughput pectin profiling by combining glycan microarrays with monoclonal antibodies. *Carbohydrate research*, 409, 41-47.

Tan, C., and Li, M. (2008). Mutual information-induced interval selection combined with kernel partial least squares for near-infrared spectral calibration. *Spectrochimica Acta Part A*: *Molecular and Biomolecular Spectroscopy*, 71(4), 1266-1273.

Thenkabail, P. S., Enclona, E. A., Ashton, M. S., and Van Der Meer, B. (2004). Accuracy assessments of hyperspectral waveband performance for vegetation analysis applications. *Remote sensing of environment*, 91(3), 354-376.

Thenkabail, P. S., Smith, R. B., and De Pauw, E. (2000). Hyperspectral vegetation indices and their relationships with agricultural crop characteristics. *Remote sensing of Environment*, 71(2), 158-182.

Thissen, U., Pepers, M., Üstün, B., Melssen, W. J., and Buydens, L. M. C. (2004). Comparing support vector machines to PLS for spectral regression applications. *Chemometrics and Intelligent Laboratory Systems*, 73(2), 169-179.

Tobias, R. D. (1995, April). An introduction to partial least squares regression. In Proc. *Ann. SAS Users Group Int. Conf., 20th, Orlando, FL* (pp. 2-5).

Üstün, B., Melssen, W. J., Oudenhuijzen, M., and Buydens, L. M. C. (2005). Determination of optimal support vector regression parameters by genetic algorithms and simplex optimization. *Analytica Chimica Acta*, 544(1), 292-305.

Van Gardingen, P. R., Jackson, G. E., Hernandez-Daumas, S., Russell, G., and Sharp, L. (1999). Leaf area index estimates obtained for clumped canopies using hemispherical photography. *Agricultural and Forest Meteorology*, 94(3), 243-257.

Vapnik, V. N., and Vapnik, V. (1998). *Statistical learning theory* (Vol. 1). New York: Wiley.

Wang, F. M., Huang, J. F., Zhou, Q. F., and Wang, X. Z. (2008). Optimal waveband identification for estimation of leaf area index of paddy rice. *Journal of Zhejiang University Science B*, 9(12), 953-963.

Wang, F. M., Huang, J. F., and Lou, Z. H. (2011a). A comparison of three methods for estimating leaf area index of paddy rice from optimal hyperspectral bands. *Precision Agriculture*, 12(3), 439-447.

Wang, X., Fu, L., and He, C. (2011b). Applying support vector regression to water quality modelling by remote sensing data. *International journal of remote sensing*, 32(23), 8615-8627.

Weiss, M., Baret, F., Smith, G. J., Jonckheere, I., and Coppin, P. (2004). Review of methods for in situ leaf area index (LAI) determination: Part II. Estimation of LAI, errors and sampling. *Agricultural and Forest Meteorology*, 121(1), 37-53.

Wise, B. M., Gallagher, N. B., Bro, R., Shaver, J. M., Windig, W., and Koch, R. S. (2006). Chemometrics tutorial for PLS_Toolbox and Solo. *Eigenvector Research, Inc*, 3905.

Wold, S., Sjöström, M., and Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2), 109-130.

Wu, C., Niu, Z., Tang, Q., and Huang, W. (2008). Estimating chlorophyll content from hyperspectral vegetation indices: *Modeling and validation. Agricultural and forest meteorology*, 148(8), 1230-1241.

Xiaobo, Z., Jiewen, Z., Xingyi, H., and Yanxiao, L. (2007). Use of FT-NIR spectrometry in non-invasive measurements of soluble solid contents (SSC) of 'Fuji'apple based on different PLS models. *Chemometrics and Intelligent Laboratory Systems*, 87(1), 43-51.

Xu, L., and Baldocchi, D. D. (2004). Seasonal variation in carbon dioxide exchange over a Mediterranean annual grassland in California. *Agricultural and Forest Meteorology*, 123(1), 79-96.

Yang, X., Huang, J., Wu, Y., Wang, J., Wang, P., Wang, X., and Huete, A. R. (2011). Estimating biophysical parameters of rice with remote sensing data using support vector machines. *Science China Life Sciences*, 54(3), 272-281.

Yao, H., and Tian, L. (2003). A genetic-algorithm-based selective principal component analysis (GA-SPCA) method for high-dimensional data feature extraction. *Geoscience and Remote Sensing, IEEE Transactions on*, 41(6), 1469-1478.

Yeniay, O., and Goktas, A. (2002). A comparison of partial least squares regression with other prediction methods. *Hacettepe Journal of Mathematics and Statistics*, 31(99), 99-101.

Yue, X., Quan, D., Hong, T., Wang, J., Qu, X., and Gan, H. (2015). Non-destructive hyperspectral measurement model of chlorophyll content for citrus leaves. *Transactions of the Chinese Society of Agricultural Engineering*, 31(1), 294-302.

Zhang, R., Ba, J., Ma, Y., Wang, S., Zhang, J., and Li, W. (2012). A comparative study on wheat leaf area index by different measurement methods. In *Agro-Geoinformatics (Agro-Geoinformatics), 2012 First International Conference on* (pp. 1-5). IEEE.

Zhao, D., Huang, L., Li, J., and Qi, J. (2007). A comparative analysis of broadband and narrowband derived vegetation indices in predicting LAI and CCD of a cotton canopy. *ISPRS Journal of Photogrammetry and Remote Sensing*, 62(1), 25-33.

Zhou, Y., Xiang, B., Wang, Z., and Chen, C. (2009). Determination of chlorpyrifos residue by near-infrared spectroscopy in white radish based on interval partial least square (iPLS) model. *Analytical Letters*, 42(10), 1518-1526.