Estimation of Maize grain yield using multispectral satellite data sets (SPOT 5) and the random forest algorithm

A. Ngie[1] and F. Ahmed[2]

[1] Department of Geography, Environmental Management and Energy Studies, University of Johannesburg, P.O. Box 524 Auckland Park 2006, South Africa, adelinengie@gmail.com, Tel.: +27 (0)11 559 4641

[2] School of Geography, Archaeology and Environmental Studies, University of the Witwatersrand, Johannesburg, South Africa

DOI: http://dx.doi.org/10.4314/sajg.v7i1.2

ABSTRACT

Crop yield estimation is a very important aspect in food production as it provides information to policy and decision makers that can guide food supply not only to a nation but also influence its import and export dynamics. Remote sensing has the ability to provide the given tool for crop yield predictions before harvesting. This study utilised canopy reflectance from a multispectral sensor to develop vegetation indices that serve as input variables into an empirical pre-harvest maize (Zea mays) yield prediction model in the north eastern section in Free State province of South Africa. Some fields in this region that were grown of maize under rain-fed conditions were monitored and the grain harvested after 7-8 months with actual yields measured. The acquisition of suitable medium resolution SPOT 5 images over this area was in March and June before the grains were harvested in July of 2014. A number of well known spectral indices were developed using the visible and near infrared bands. Through the random forest algorithm predictive models, maize grain yields were estimated successfully from the March images. The accuracies of these models were of an R^2 of 0.92 (RMSEP = 0.11, MBE = -0.08) for the Agnes field and for Cairo the R^2 was 0.9 (RMSEP = 0.03, MBE = 0.004). These results were produced by the SAVI and NDVI respectively for both fields. It was therefore evident that the predictive model applied in this study was site specific and would be interesting to be tested for an optimal period during the plant life cycle to predict grain yields of maize in South Africa.

*Keywords**:** maize, non-linear regressions, prediction, random forest, spectral indices, SPOT 5, variable importance, yield*

1. Introduction

1.1. Background to study

Crop yield prediction is production estimates that are made a couple of months or weeks depending on the crop in question before the actual harvest. This is frequently done through computer programmes that utilize agro-meteorological data, soil data, remotely sensed and agricultural statistics to describe quantitatively the plant-environment interactions (Zere *et al.*, 2004). In some instances,

meteorological data is included to run some of the yield models. The meteorological data is usually generated from weather stations and cover a given area. Hence, crop yield can be described as involving the effect of biotic and abiotic factors cumulatively which could however vary not just across fields but among fields and seasons alike (Bullock, 2004).

The traditional methods turn to be time-consuming and cannot consider yield variations over a field or space; therefore they are prone to large errors due to incomplete ground observations, leading to poor crop yield assessment and crop area estimations or predictions (Reynolds & Yittayew, 2000; Sau *et al.,* 2004). In the light of these limitations, remote sensing methods were introduced (Mo *et al.,* 2005). While remote sensing methods seemed to have responded to the above challenges, they were not without problems among which availability of suitable satellite data is enlisted. These challenges have led to the continuous seeking of improvements in yield estimation through either repeated application of already existing methods in different fields with different satellite data and crop types (Ngie *et al.,* 2014).

Over the years remotely sensed data has proven worthy through its extracted spectral information to give information that relates statistically to crop yields and mapping of the spatial variability across regions as well as fields (Sun, 2000; Li *et al.,* 2007). The frequently researched field crops have included wheat (Singh *et al.,* 2002; Thenkabail, 2003; Bullock, 2004; Kastens *et al.,* 2005; Ren *et al.,* 2008), potatoes (Al-Gaadi *et al.,* 2016) rice (Casanova *et al.,* 1998; Noureldin *et al.,* 2013), soybeans (Kastens *et al.,* 2005; Li *et al.,* 2007, You *et al.,* 2017) and maize (Lewis *et al.,* 1998; Shanahan *et al.,* 2001; Baez-Gonzalez *et al.,* 2002; Ferencz *et al.,* 2004; Baez-Gonzalez *et al.,* 2005; Kastens *et al.,* 2005; Kogan *et al.,* 2005; Mkhabela *et al.,* 2005; Li *et al.,* 2007; Inman *et al.,* 2007; Salazar *et al.,* 2008; Panda *et al.,* 2010; Bognár *et al.,* 2011). Most of these studies made used of the normalised difference vegetation index (NDVI) generated from the coarse resolution sensors such as the Advanced Very High Resolution Radiometer (AVHRR) and the Moderate Resolution Imaging Spectroradiometer (MODIS) to model yields. The use of these sensors was instrumental since the research focus was mostly at regional or county levels. Among the above cited studies, Singh *et al.* (2002) and Thenkabail (2003) worked on wheat fields at local levels and for the maize, only Inman *et al.* (2007) used a handheld sensor to collect remotely sensed data at field level for yield estimation. However, the listed studies above can only show the trend and importance to maize yield predictions which cannot be overemphasized as the crop is relevant not just for food security but other economic sectors like energy.

The use of remote sensing to estimate biological crop yields is being explored in many countries such as the United States, China and India, and likely will become the keystone of agricultural statistics in the future (Zhao *et al.,* 2007). The fact that crop productivity vary greatly across climatic regions since it depends on agroclimatic conditions, the application of remote sensing in this field would be necessary to show the differences. The variability of these conditions warrants models to be developed based on the conditions of different areas where the crops are planted. Moreover, there

is room to improve on methodologies and principles such as applying non-linear statistical algorithms.

South Africa is among the top ten maize producers in the world and a major player on the African continent, which makes it necessary to monitor productivity through quick and reliable methods such as remote sensing. The medium resolution Satellite Pour l'Observation de la Terre (SPOT 5) images were accessed through the South African Space Agency (SANSA). This study also sets out to test a non-linear statistical method in analysing canopy reflectance values for precise or fairly accurate prediction of maize grains models for crops grown under field conditions.

2. Materials and methods

2.1. Description of study area

The fields cultivated with maize in the 2011/2012 farming season by the farming group with which collaboration was reached were around Sasolburg and Parys. The latter is the major town in the Metsimaholo local municipality (Metsimaholo LM) and the former is of the Ngwathe local municipality (Ngwathe LM), all in the northeastern section of the Free State province of South Africa (Figure 1). However, the two fields for this study where access was granted by the farmer were located within the Ngwathe local municipality (Figure 1). This area is located within the "Maize Triangle" of South Africa that is seated within two other provinces being the North West and Gauteng.

This region where the fields were located is fairly flat with an altitude of less than 1500 m above sea level and mostly covered by the grassland ecosystem. The area is well watered by some of the main river systems of South Africa such as the Vaal and Orange Rivers. Rainfall (500 mm per annum) over this region is during the summer months (October to April) followed by the winter season which can get frosty.[1] It is made up of rich soils and greatly covered by commercial farms in grains (maize, soybeans, sunflower and sorghum). The Cairo field was made up of 97.83 hectares and the Agnes was 109.89 hectares.

2.2. Satellite image acquisition and pre-processing

The SPOT 5 L3 data set of path/row 133/404 acquired in March and June 2012 was obtained from the *fundisa* disc provided to universities by the South African National Space Agency (SANSA). The multispectral digital imagery has a 10 m pixel value with the green band range of 500 - 590 nm, red band range of 610 - 680 nm, near infrared (NIR) band range of 780 - 890 nm and the shortwave infrared (SWIR) band range of 1.58 -1.75 nm. The green and red bands make up the visible region of the electromagnetic spectrum for this sensor. For this study, the interest was with the visible and NIR bands which are vital in measuring the vigour or photosynthetic capacity of the plants.

[1] http://www.fallingrain.com/world/SF/03/Parys.html Accessed 18/10/2017

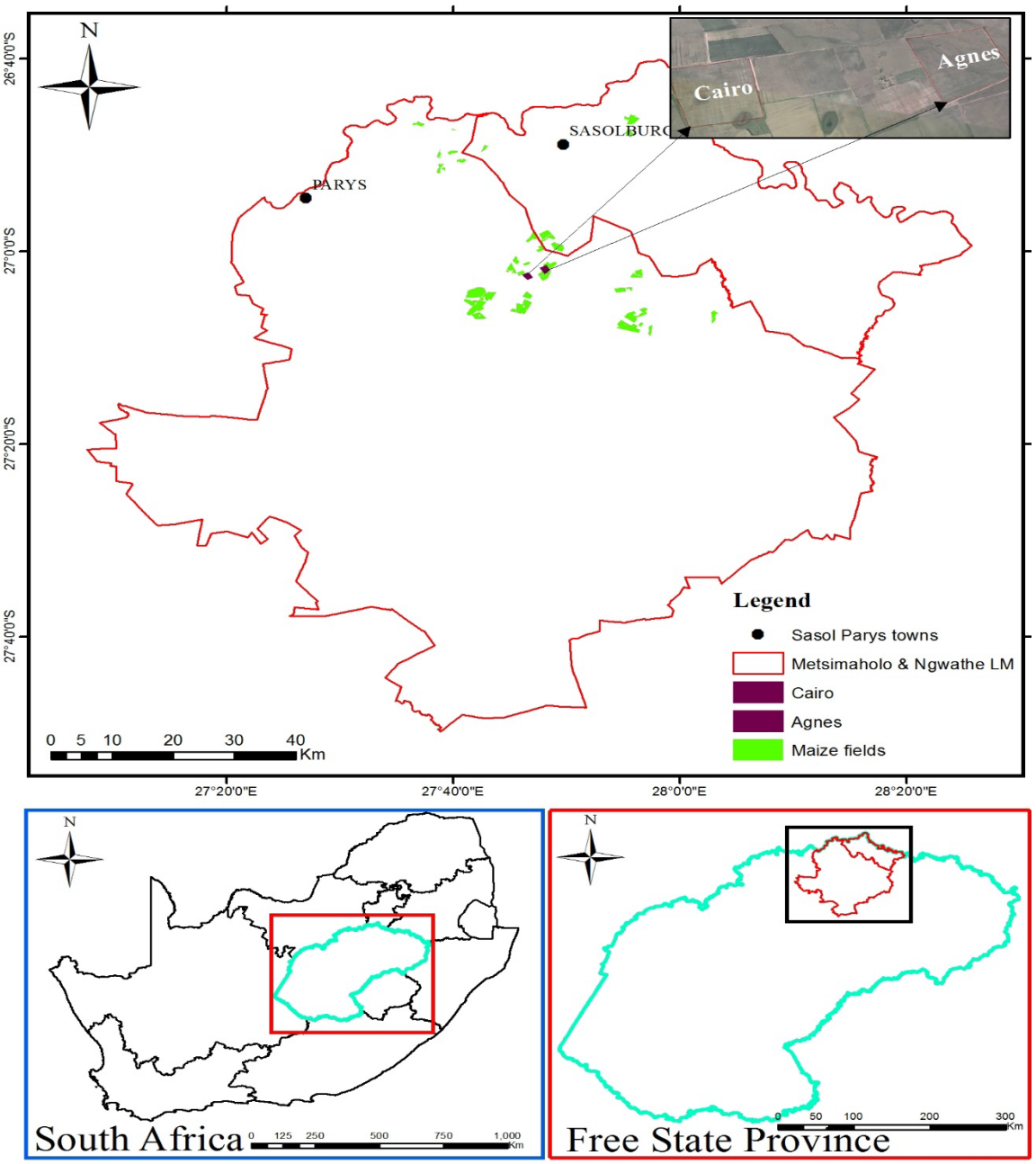


Figure 1: Map of study fields Cairo and Agnes within the local municipalities (LM) in the Free State province of South Africa. The maize fields indicated here were only those cultivated by the farming group that the researchers collaborated with for this study (Insert image from GoogleEarth 8/5/2014)

2.3. Field data acquisition

Field visits were conducted firstly to ascertain accessibility to farms with proper harvesting systems that record the yields across the fields. Secondly, the visits to the identified fields were to ascertain the conditions of the plants as well as the plots. The fields were planted with maize and treated under the normal field conditions for maximal production as marketed by the seed company. After maturity, the plants were left on the field to lose over 90% of its moisture to enable grain harvest and storage. The fields were harvested using a combine harvester that recorded the grain weight per

hectare within 20 m X 20 m ranges at (kg/ha). The harvester has an onboard system with a GPS and records the coordinates of the plots against the dry weight of the harvested grain.

In order to avoid the effects from the field boundaries, only plots that were completely within the fields were selected for the extraction of reflectance values from the developed indices (Thenkabail, 2003). The demarcated polygons were comprised of 4 pixels each to correspond with the harvest area of 20 m X 20 m plots. The reflectance value from the polygons was an average of the 4 pixels and the value per polygon served as sample observation (n) for the yield prediction models. The sampling of plots across the field was crucial since spatial variability was evident and replicates at other locations were considered based on the actual yield provided by the farmer after harvest. The plots were randomly chosen but considering areas of high and low actual grain weight.

3. Data analysis

3.1. Spectral vegetation indices

The principal reason for using spectral vegetation indices in crop studies has been to compensate the effects of factors of disturbance between the spectral reflectance measured from the vegetation and its characteristics such as canopy biomass or vegetation type (Bouman, 1995). The indices obtained from optical sensors have been valuable in crop production estimates through leaf interception media being mainly the leaf area index (LAI) (Tucker, 1979). The main index from which this interception medium is derived is the NDVI thereby making it a building block towards crop yield estimation using remote sensing. However, this relationship between the NDVI and the LAI is not without vices as it saturates with fully covered plant canopy (Pontailler *et al.,* 2003).

There are some disturbing factors such as soil background to measured-reflectance where other distance-based vegetation indices were included to this analysis such as the Soil adjusted vegetation index (SAVI) (Huete, 1988) to overcome. The SAVI has the ability to completely cancel or reduce the effect of soil brightness wherever pixels have a combination of soil and vegetation reflectance (Huete & Jackson, 1988). It is a calibration factor in the NDVI equation that accounts for the first order soil-vegetation optical interactions and its potential has been successfully proven (Huete, 1988). All the spectral vegetation index images were recreated from the identified indices (Table 1) in the Environment for Visualizing Images (ENVI) software (v. 5.0, ITT Visual Information Systems, 2012) and had the various areas of interest (AOI) as the 4-pixels identified. The average reflectance value for the 4-pixel plots from each index served as samples for the statistical modelling.

Table 1: Spectral vegetation indices used in this study

Index acronym	Name and Description	Formula	Reference
NDVI	Normalised difference index: related to changes in amount of green biomass: pigment content and water stress	$(R_{NIR}-R_{RED})/(R_{NIR}+R_{RED})$	Rouse *et al.*, 1974
TNDVI	Transformed normalised difference index relates to the green leaf material or photosynthetically active biomass in the plant canopy	(NDVI+ 0.5)1/2	Tucker (1979)
RDVI	Re-normalised difference vegetation index is used to linearise relationships between the index and surface parameters that tend to be nonlinear	$(R_{NIR}-R_{RED})/(R_{NIR}+R_{RED})1/2$	Roujean & Breon (1995)
SR	Simple ratio which relates to changes in the amount of green biomass, pigment content as well as leaf water stress	R_{NIR}/R_{RED}	Rouse *et al.* (1974)
Sqrt SR	Square root of simple ratio that relates primarily to the green leaf area or biomass	$(R_{NIR}/R_{RED})1/2$	Tucker (1979)
MSR	Modified simple ratio that de-linearises relationships between the index and biophysical parameters	$(R_{NIR}/R_{RED}-1)/(R_{NIR}/R_{RED})1/2+1$	Chen (1996)
GVI or NDVIgreen or GNDVI	Green vegetation index which determines nitrogen influences from the green colour of the leaf through green reflectance	$(R_{NIR}-R_{GREEN})/(R_{NIR}+R_{GREEN})$	Gitelson *et al.* (1996)
GRDI	Green red difference index which is the visible light normalised index and can relate to plant pigment content	$(R_{GREEN}-R_{RED})/(R_{GREEN}+R_{RED})$	Gianelle & Vescovo (2007)
VI	Vegetation index is sensitive to the green leaf material or the photosynthetically active biomass in plant canopy	$R_{NIR}-R_{RED}$	Tucker (1979)
GDI	Green difference index which is a non-normalised index. Could therefore be used when the impact of factors such as slope and aspect is not pronounced	$R_{NIR}+R_{RED}+R_{GREEN}$	Gianelle & Vescovo (2007)
SAVI	Soil adjusted vegetation index is used to reduce soil brightness in vegetation reflectance	$(R_{NIR}+R_{RED})(1+L^2)$ $(R_{NIR}+R_{RED}+L)$	Huete (1988)

3.2. Random forest (RF) algorithm

The RF operates on the principle of constructing through recursive partitioning to split data into homogenous regression trees independently to maximum size without pruning and averages the results of all trees. There are two important parameters in the construction of the RF algorithm namely: the number of trees (*ntree*), and number of variables randomly chosen at each split (*mtry*) (Breiman, 2001). The robustness of the RF causes it to fit against the challenge of over-fitting that is experienced in linear models (Prasad *et al.*, 2006; Palmer *et al.*, 2007). The RF regression algorithm operates through bootstrapping samples from randomly divided original data set into the two third

[2] L is a canopy background adjustment factor (a correlation factor for soil line between red and near infrared reflectance) set at 0.5 in this study.

(2/3) training sample and one third (1/3) testing sample. There was a variation of the values for the *ntree* and *mtry* parameters accordingly. Liaw and Weiner (2002) in their study recommended the optimum number of *mtry* to be defined by one third of the total number of the input variables (32 for Agnes and 36 for Cairo). Meanwhile the *ntree* was regularised through the model for selection from 500 up to 2500 at 500 intervals (Prasad *et al.,* 2006). The algorithm was performed for each growth stage (March and June) for the two fields being calibrated on training sample, n = 64 for Agnes field; n = 72 for Cairo and testing sample with n = 32 for the Agnes; n = 36 for Cairo fields. The calibration was evaluated through the root mean square error of calibration (RMSEC). The same sample plots were used for both the March and June sample dates or growth stages of the maize plants.

3.3. Selection of variables (vegetation indices)

The RF algorithm also has the ability to calculate variable importance (*varimp*) (Breiman, 2001). This function has been critiqued for its bias nature of selecting variables (Strobl *et al.,* 2007) which would be able to extract relevant indices to maize yield prediction. The conditional forest (*cforest*) has been proposed for such variable selection analysis. The *cforest* function reduces the level of biased selection of variables in individual classification trees unlike the variable selection embedded in the original RF by Breiman (2001) (Strobl *et al.,* 2008; Strobl *et al.,* 2009). The *cforest* is included within the development of the regression script for the RF algorithm to run simultaneously.

The RF ensemble uses the out-of-bag (OOB) error estimates to rank variables and is derived by predicting the data that are in each tree being considered (error of prediction). Prasad *et al.* (2006) then describes the variable importance as an evaluation of how worse the prediction becomes when the data for a variable were randomly permuted.

For the evaluation of different variables (selected spectral vegetation indices from *varimp*) in the performance of the RF regression algorithms, the predicted and actual or measured maize grain yields were related in one-to-one sets. During the relationship match, the coefficient of determination (R^2), root mean squared error of prediction (RMSEP) and the mean bias error (MBE) were calculated for every model run with the selected variables (spectral vegetation index).

4. Results and discussion

4.1. Measured maize yields

The actual or measured yields of the maize grains from the Agnes and Cairo field for 2012 harvesting season from the combined harvester were recorded. The descriptive analysis indicated the Cairo field more productive with a higher mean yield of 4077.26 kg/ha (4.08 t/ha) as opposed to the 3207.1 kg/ha (3.21 t/ha) for the Agnes field (Table 2).

Table 2: Descriptive statistics of the actual yields (kg/ha) for both fields

Description	Agnes field	Cairo field
Total Area (ha)	110.45	97.83
Count (number of quadrants)	2936	2597
Minimum yield	983.69	2745.1
Maximum yield	4569.82	5457.59
Sum	9416046.2	10588637.84
Mean	3207.1	4077.26
Standard Deviation	542.45	522.05

4.2. Spectral index of importance selection

The RF prediction models for maize grain yield including all the developed spectral vegetation indices in this study proved successful through the varied values of the parameters. The optimum performing parameter values were identified (Table 3). However, the contribution of the various spectral vegetation indices to the success would have varied according to their relationships with the grain productive parameters of the maize plants such as chlorophyll content, water content and others.

Table 3: Summary results of the RF prediction models for maize with all indices showing validation parameters

Field	Data period	*ntree*	*mtry*	R^2	RMSEC
Agnes	March	500	11	0.79	4.18
		1000	11	0.35	31.82
		1500	11	0.42	13.38
		2000	11	0.38	29.34
		2500	11	0.27	34.24
	June	500	11	0.49	8.54
		1000	11	0.23	37.09
		1500	11	0.34	32.09
		2000	11	0.21	38.19
		2500	11	0.18	42.63
Cairo	March	500	12	0.76	3.06
		1000	12	0.82	2.48
		1500	12	0.42	12.35
		2000	12	0.48	8.04
		2500	12	0.39	26.32
	June	500	12	0.51	10.42
		1000	12	0.47	8.91
		1500	12	0.19	42.83
		2000	12	0.24	36.02
		2500	12	0.19	41.58

The *ntree* value of 500 generally outperformed the others (up to 2500) by the higher R^2 values and lower RMSEC for both images (March and June). The evaluations of accuracy for both images showed the March image with higher R^2 and lower RMSEC values. The spectral vegetation indices were input into the model as independent variables and in running the variable selection function reported on their contribution through the OOB error estimates. The results once again showed the mid cropping growth stage (March for maize in this region) to be the relevant period for assessing crop vigour towards yields. Hence, the indices that relate most to chlorophyll concentrations were selected as contributing the most to the error of predictions or show of importance. The selected indices ranked at top three included NDVI, SAVI and GVI (Figure 2; Figure 3).

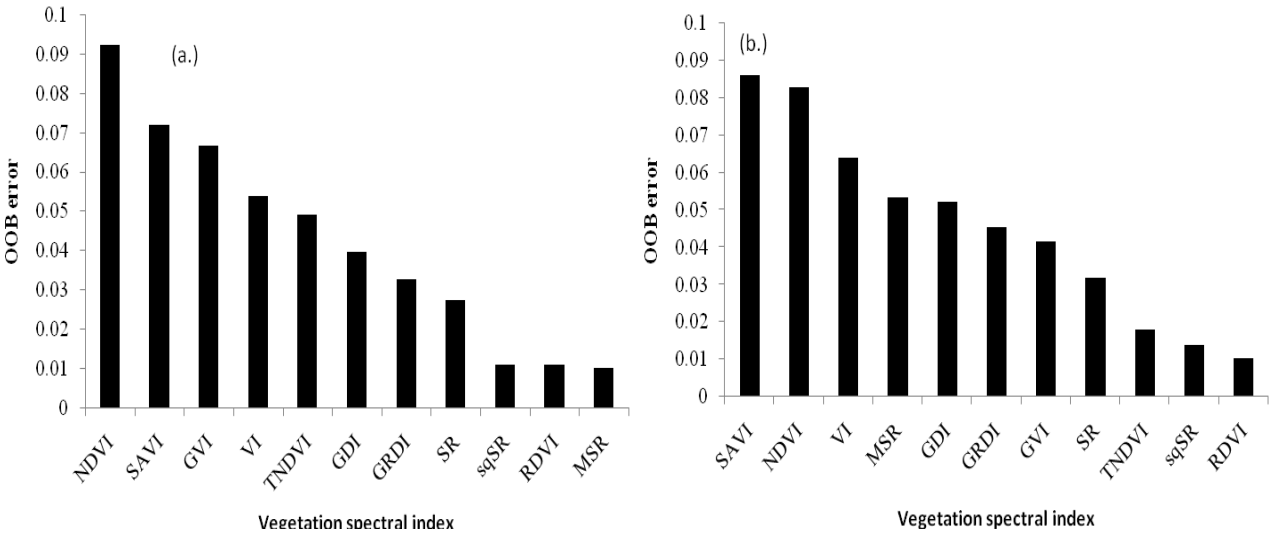


Figure 2: The importance of spectral vegetation indices for predicting maize grain yield (ton/ha) in the Agnes field (a.) March and (b.) June

The ranking of individual spectral vegetation indices according to their importance in predicting maize yield from the March image were recorded (Figure 2). The importance of a spectral vegetation index in predicting maize yields could have depended on the growth stage or period that the satellite images were acquired. This is proven by the higher OOB errors registered by the selected indices in March than for the June images (Figure 2; Figure 3). The larger error relates to the importance of the variable when left out of the permutation or better still it shows how bad the algorithm perform with that specific variable.

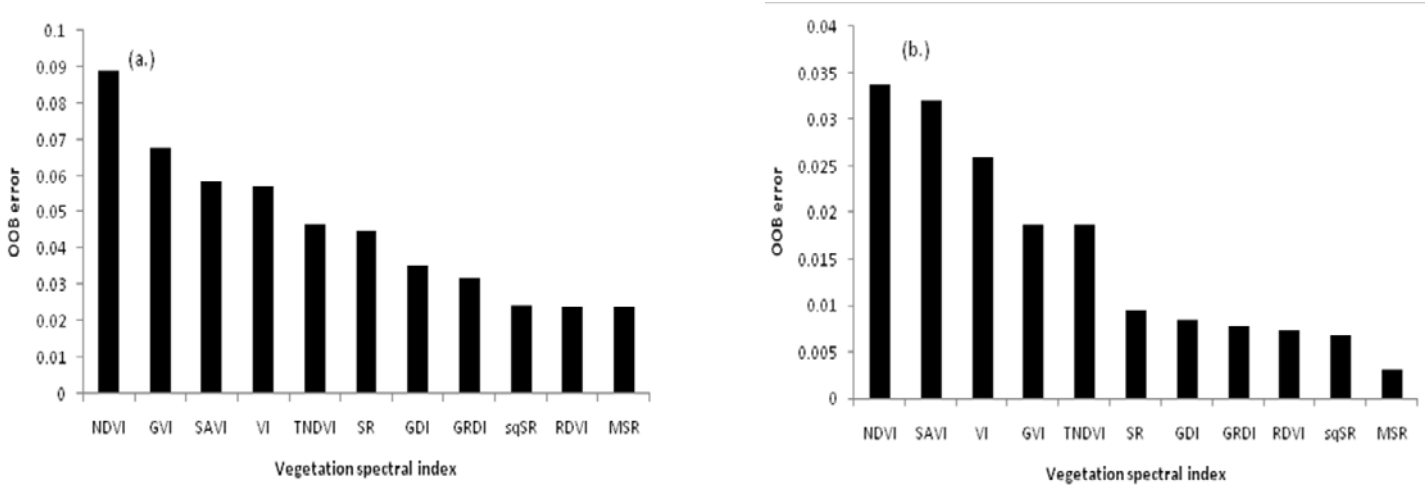


Figure 3: The importance of spectral vegetation indices for predicting maize grain yield (ton/ha) in the Cairo field (a.) March and (b.) June

The observed results with the NDVI as variable of importance in both fields, was in confirmation of its importance in relating the chlorophyll concentration. This is through the absorption levels of the red light and reflection levels in the NIR region of the electromagnetic spectrum of healthy plants (Rouse *et al.,* 1974). The GVI was among the top three selected indices by the *cforest* which confirms its importance in maize yield estimation as reported in previous studies (Shanahan *et al.,* 2001) where it highly correlated with maize grain yields. The relevance of the SAVI in maize yield predictions was also confirmed to Panda *et al.* (2010) where it was considered successful at 95.04% (R^2 of 0.53) equally with the NDVI. It should be clear that as the maize plants matured towards harvesting of the grains, soil background had a significant effect on the maize spectral features. Then it accounts for identification of the SAVI as an important index to maize yield prediction at such growth stage.

4.3. Random forest regression algorithm for maize yields using selected indices

The results of the selected indices showed marked increase of the R^2 values and relatively lower RMSEP values as well as the MBE for the corresponding growth stages (March or June) (Figure 4; Figure 5; Figure 6; Figure 7). Once again, results from March images showed significantly better results than those from the June images for both fields. The results could be attributed to the fact that the visible and NIR regions of the electromagnetic spectrum mostly relates to the chlorophyll light absorption feature in plants. According to Rouse *et al.* (1974) the concentration of this pigment dictates the amount of reflectance measured. With the maize plants being greener in March than in June not just because of the distinctive summer and winter seasons but also growth stages where March was the vegetative stage and June the reproductive stage. The NDVI which is developed from the difference of the region of maximum chlorophyll absorption (red region) and the corresponding region of maximum reflectance of incident light (NIR) has proven successful in all maize yield estimation studies applying remote sensing (Ngie *et al.,* 2014) and this study was not left out. It proved to be the best out of the indices for both growth stages though with varying accuracies as well as the two fields.

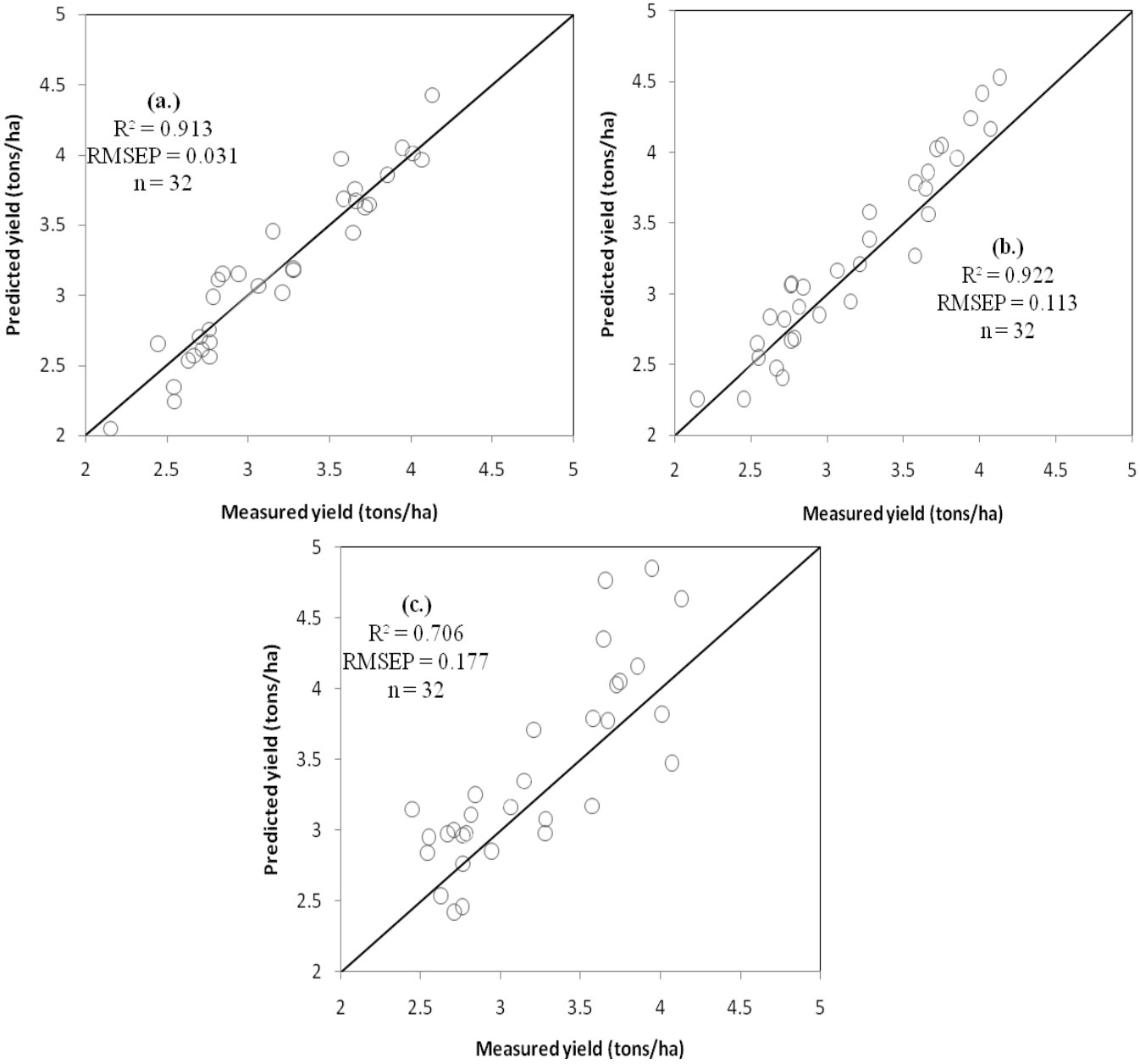


Figure 4: One-to-one relationship between predicted and measured maize yields from March images over Agnes using individual selected indices for Agnes with (a) = NDVI, (b) = SAVI and (c) = GVI

The accuracies of predictions obtained for this study however, do vary from previously recorded studies (Ngie *et al.*, 2014) which could be accounted for by the different sensors used which ranged from coarse to medium and to fine resolution satellite images. The choice of the sensor also depends on the spatial extent of the study area. In some situation also where the same sensor was used, other factors such as soil characteristics, climatic conditions, cultivar types and the growth stage at which the yields were predicted varied, thereby causing discrepancies in the accuracy. For instance the results of this study which comprised of the same cultivar planted under same field conditions and monitored with the same sensor as well as dates did not provide the same accuracies in yield predictions. The difference in accuracies with R^2 of 0.91 and 0.89 for Agnes and Cairo fields obtained from NDVI in March (Figure 4; Figure 6) was however slightly lower but looks insignificant.

The NDVI and SAVI variables performed better in the model than the GVI which has problems of underestimation and overestimation. The GVI even though resulting with a high R^2 (over 70%) and a low RMSEP (0.18) overestimated the grain yield at higher grain output (Figure 4 (c)) which could be a challenge to stakeholders requiring adequate information for financial planning.

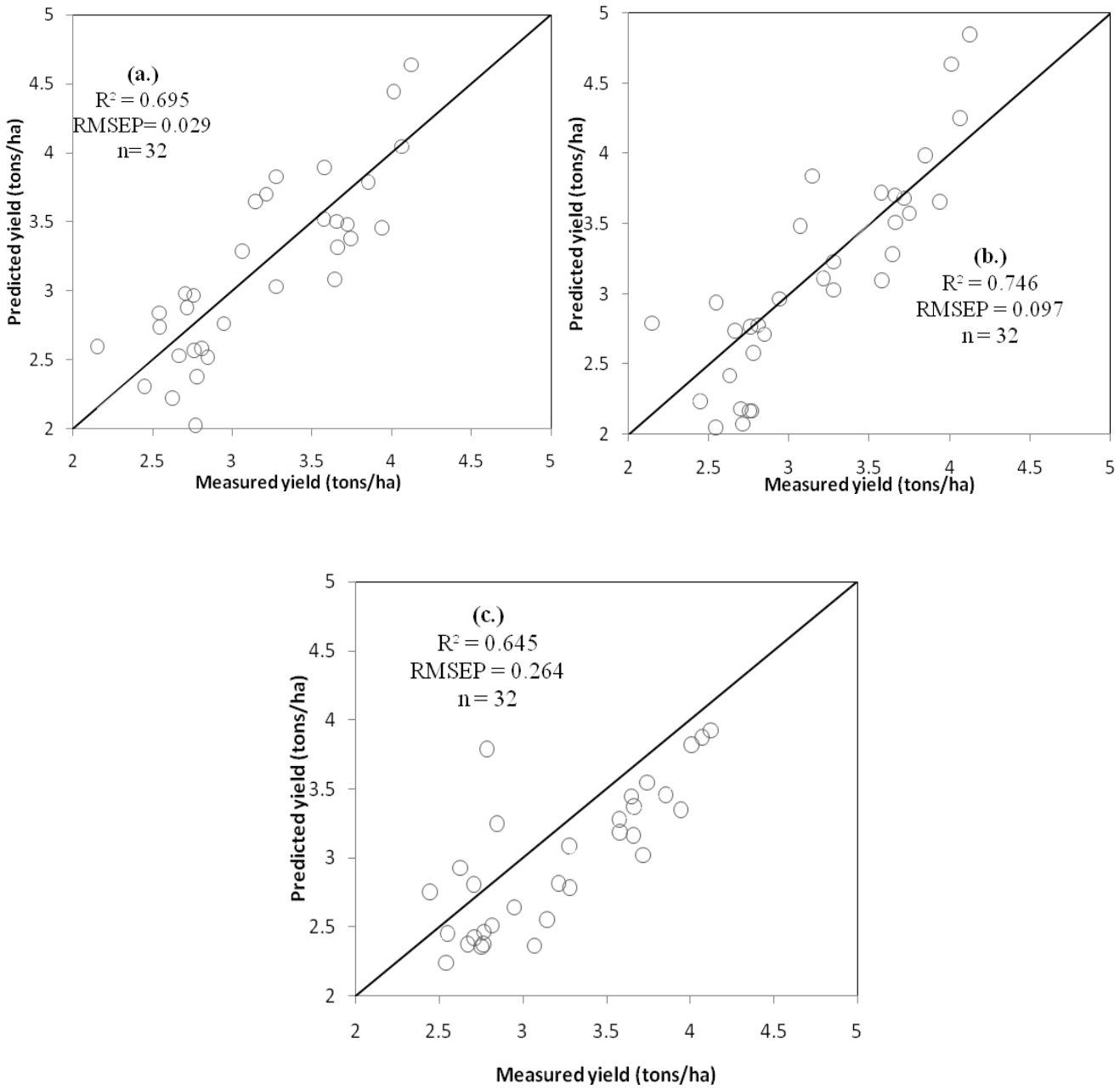


Figure 5: One-to-one relationship between predicted and measured maize yields from June images over Agnes using individual selected indices for Agnes (a) = NDVI, (b) = SAVI and (c) = VI

The June images resulted in maize grain yields that were either underestimating or overestimating at both the low and high productivity. The results illustrated that the June images could not be used to establish the vigour in the plants since the crops were at an advanced stage of maturity. Hence, the model would have required other parameters to strengthen its predictive ability.

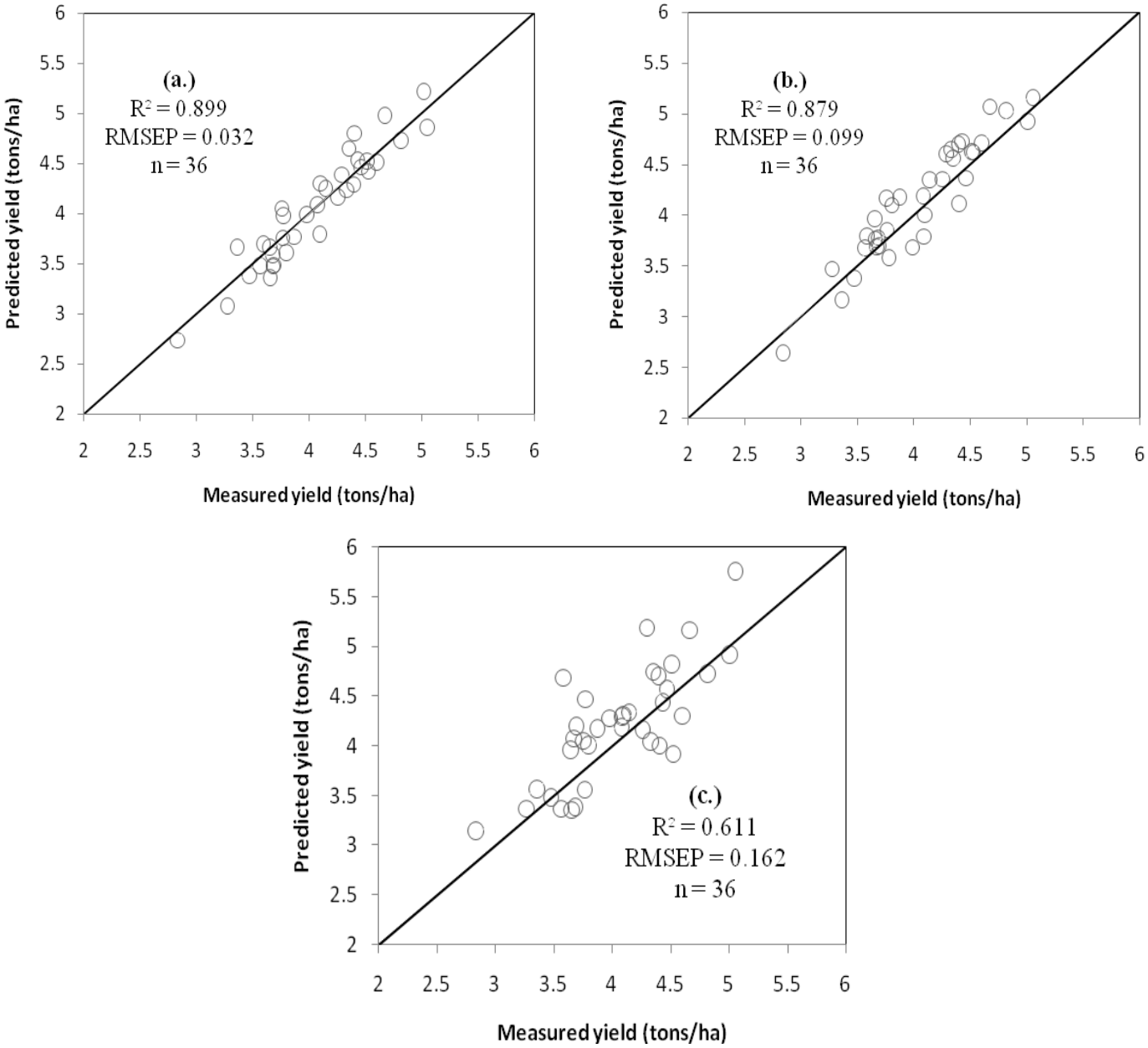


Figure 6: One-to-one relationship between predicted and measured maize yields from March images over Cairo using individual selected indices for Cairo in March (a) = NDVI, (b.) = SAVI and (c) = GVI

Once again in the Cairo field, the model was more precise with the NDVI and SAVI variables than the GVI which was challenged with overestimation as the productivity increased (Figure 6).

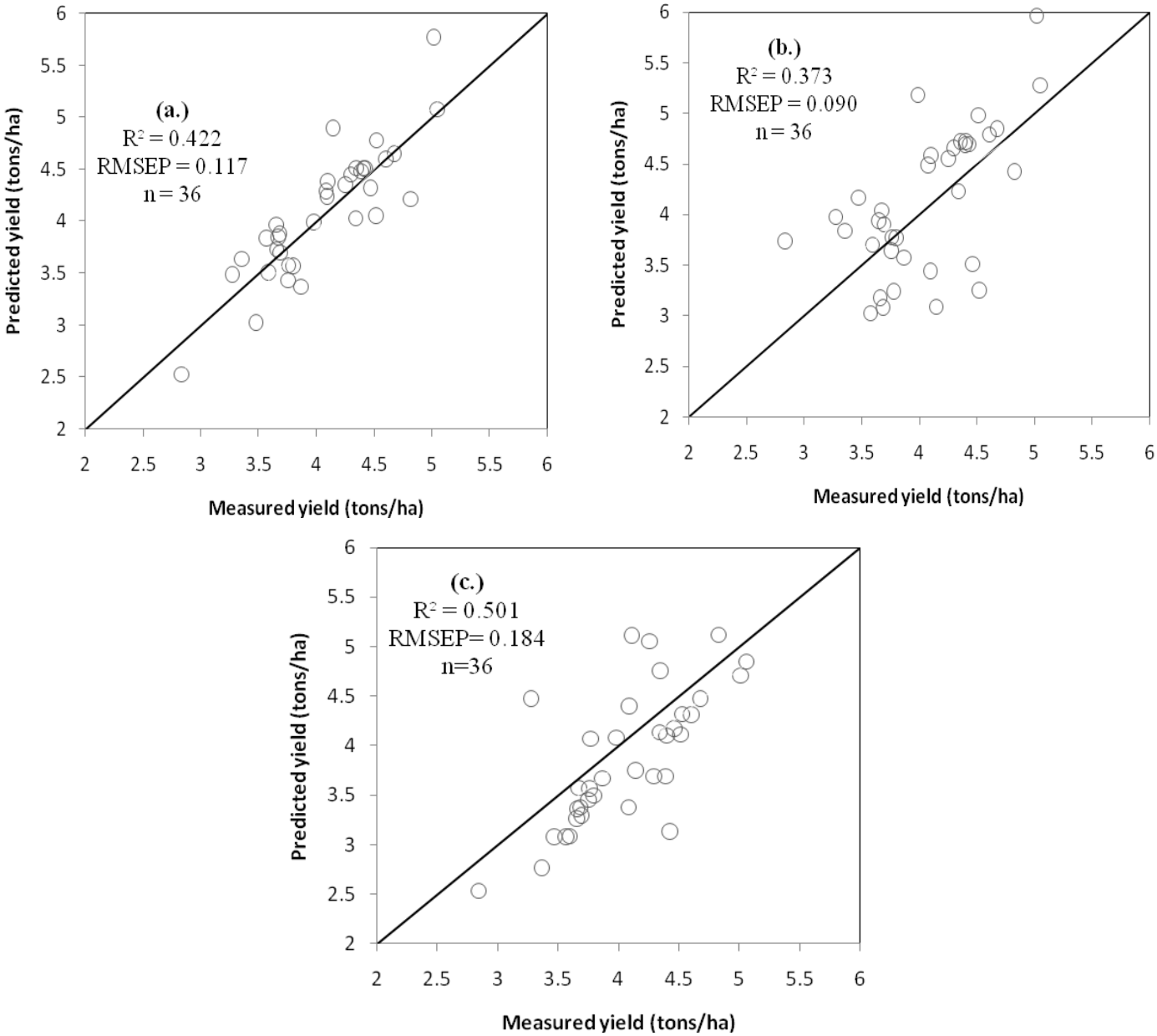


Figure 7: One-to-one relationship between predicted and measured maize yields from June images over Cairo using individual selected indices for Cairo (a) = NDVI, (b) = SAVI and (c) = VI

The Cairo field also experienced a poor performance of the model with variables (indices) developed from the June images which could be explained as above with the Agnes field. In validating the performance of the regression algorithms in maize grain predictions for the selected spectral vegetation indices, there was a general improvement in the accuracies from the ones ran from all the indices (Table 3). The evaluation of the accuracies was noted through the lower RMSEP and MBE, and the increased R^2 values for both fields as well as the growth stages (March and June).

There was an improvement in the performance of the yield prediction algorithms that was as a result of using the random forest selection function to identify relevant indices (variables of importance). These results illustrated a good performance of this algorithm in prediction and feature selection analysis as earlier noted by Prasad *et al.* 2006. The selected spectral vegetation indices related to the amount of green materials (NDVI, GVI or VI) in the maize plants (Rouse *et al.,* 1974;

Tucker, 1979). Spectral vegetation indices that are responsive to the green pigments are excellent indicators for vegetation quantity and status or vigour (Salazar *et al.*, 2008), and confirms the selection of the GVI in the top three indices as ranked by the OOB error estimates in this study as relevant to maize grain yield prediction.

The results from the two dates of data acquisition illustrated March as a period of better grain yield prediction for maize in this area of South Africa than the June. Even though a more robust study would need to be conducted to ascertain the optimum period of yield estimates or predictions with more monitoring dates throughout the growing season, the March period which was about four months to harvest (mid cropping season for maize in this region) was in conformity to another study in the United States of America (USA) where at 3-4 months before harvest predictions resulted in an estimation error of 3% (Kogan *et al.*, 2012). In another previous study, it was 2-3 weeks before and after tasseling (Kogan *et al.*, 2005) that was identified as optimum growth stage for yield predictions. Meanwhile an 8-leaf stage was optimal for predicting maize grain yields and the 3-leaf stage was optimal for biomass estimations for Islam *et al.* (2011). According to Panda *et al.* (2010), the mid cropping season of maize was ideal growth stage to predict grain yields of the crop in North Dakota. These discrepancies therefore suggest that optimal growth stage depends on the geographical region as well as the cultivar type being investigated as some might have a longer life span.

The observed inconsistency in maize grain yield predictability in the different fields (Agnes and Cairo in this study) (Appendix A) could also have been explained by the complexity in crop yield that depends on other non-imagery factors, such as nutrient stresses, or water availability. These factors were not considered in developing the random forest regression algorithms used in this study. High performance in crop yields could be obtained if their production parameters remain consistent throughout the season until harvest (Panda *et al.*, 2010) which is hardly the situation and therefore contribute to discrepancies in crop yield estimate results. Hence, these algorithms are site specific and applying them to other areas might not produce same accuracies but should still perform well considering all parameters.

5. Conclusions

In-field maize yield estimation or prediction using multispectral satellite imagery of medium resolution over rain-fed fields proved successful through the use of vegetation spectral indices. The spectral indices derived from SPOT 5 imageries were used as input variables into the random forest algorithm for regression analysis in predicting the grain yield by weight of maize across both fields with a good accuracy of high coefficient of determination (R^2) values and low RMSEP as well as MBE values. However, the prediction was more accurate earlier on in the season (vegetative growth stage in March) than later for the reproductive stage in June. This could be attributed to the fact that the green pigment in maize leaves is largely responsible for its yields through photosynthetic activities. The NDVI was amongst the most important indices relating to maize yields for this study

as proven through the random forest non-linear regression algorithm and selected by the *cforest* ranking them through the OOB errors.

It would be of interest for future research to engage in the optimisation of the grain prediction period in maize over the "Maize Triangle" of South Africa to ascertain in a timely manner the production of this important grain. Also as a result of the climatic challenges in obtaining real-time satellite optical data during the growing season, the potential of radar data such as Sentinel 1 could be exploited in monitoring crop productivity. There could also be the possibility to integrate climatic factors essential for maize productivity such as rainfall and temperature, and linking with soil nutrient data together with the vegetation spectral indices for a mechanical predictive model in future research. In that way, climatic-related periods of critical development that controls productivity would be established thereby providing information for decisive measures to be taken.

6. References

Al-Gaadi, K.A., A.A. Hassaballa, E. Tola, A.G. Kayad, R. Madugundu, B. Alblewi and F. Assiri, 2016: Prediction of potato crop yield using precision agriculture techniques. *PLoS ONE 11(9),* e0162219. doi:10.1371/journal.pone.0162219

Baez-Gonzalez, A.D., J.R. Kiniry, S.J. Maas, M.L. Tiscareno, J.C. Macias, J.L. Mendoza, C.W. Richardson, J.G. Salinas and J.R. Manjarrez, 2005: Large-area maize yield forecasting using leaf area index based yield model. *Agronomic Journal 97*, 418-425.

Baez-Gonzalez, A.D., P. Chen, M. Tiscareno-Lopez, and R. Srinivasan, 2002: Using satellite and field data with crop growth modeling to monitor and estimate corn yield in Mexico. *Crop Science 42*, 1943-1949.

Bognár, P., Cs. Ferencz, Sz. Pásztor, G. Molnár, G. Timár, D. Hamar, J. Lichtenberger, B. Székely, P. Steinbach and O.E. Ferencz, 2011: Yield forecasting for wheat and corn in Hungary by satellite remote sensing. *International Journal of Remote Sensing 32(17),* 4759-4767, DOI: 10.1080/01431161.2010.493566

Bouman, B.A.M., 1995: Crop modeling and remote sensing for yield prediction. *Journal of Agricultural Science 43*, 143-161.

Bullock, P.R., 2004: A comparison of growing season agro-meteorological stress and single-date Landsat NDVI for wheat yield estimation in west central Saskatchewan. *Canadian Journal of Remote Sensing 30*, 101-108.

Casanova, D., G.F. Epema and J. Goudriaan, 1998: Monitoring rice reflectance at field level for estimating biomass and LAI. *Field Crop Research 55*, 83-92.

Chen, J. M., 1996: Evaluation of vegetation indices and a modified simple ratio for boreal applications. *Canadian Journal of Remote Sensing 22*, 229-242.

Ferencz, Cs., P. Bognár, J. Lichtenberger, D. Hamar, Gy. Tarcsai, G. Timár, G. Molnár, S.Z. Pásztor, P. Steinbach, B. Székely, O. E. Ferencz and I. Ferencz-Árkos, 2004: Crop yield estimation by satellite remote sensing. *International Journal of Remote Sensing 25(20),* 4113-4149, DOI: 10.1080/01431160410001698870

Gianelle, D. and L. Vescovo, 2007: Determination of green herbage ratio in grasslands using spectral reflectance: Methods and ground measurements. *International Journal of Remote Sensing 28,* 931-942.

Gitelson, A., Y. Kaufman, and M. Merzlyak, 1996: Use of a green channel in remote sensing of global vegetation from EOSMODIS. *Remote Sensing of Environment 58,* 289-298.

Huete, A.R. and R.D. Jackson, 1988: Soil and atmosphere influences on the spectra of partial canopies. *Remote Sensing of Environment 25*, 89-105.

Huete, A.R., 1988: A soil adjusted vegetation index (SAVI). *Remote Sensing of Environment 25,* 295-309.

Inman, D., R. Khosla, R. M. Reich and D. G.Westfall, 2007: Active remote sensing and grain yield in irrigated maize. *Precision Agriculture 8,* 241-252

Islam, M.R., S.C. (Yani) Garcia and D. Henry, 2011: Use of normalised difference vegetation index, nitrogen concentration, and total nitrogen content of whole maize plant and plant fractions to estimate yield and nutritive value of hybrid forage maize. *Crop and Pasture Science 62(5),* 374-382.

Kastens, J. H., T.L., Kastens, D.L.A. Kastens, K.P. Price, E. A. Martinko and R. Lee, 2005: Image masking for crop yield forecasting using AVHRR NDVI time series imagery. *Remote Sensing of Environment 99,* 341-356.

Kogan, F. B. Yang, G. Wei, P. Zhiyuan and J., Xianfeng, 2005: Modelling corn production in China using AVHRR-based vegetation health indices. *International Journal of Remote Sensing 26,* 2325-2336.

Kogan, F., L. Salazar and L. Roytman, 2012: Forecasting crop production using satellite-based vegetation health indices in Kansas, USA. *International Journal of Remote Sensing 33(9),* 10 2798-2814.

Lewis, J.E., J. Rowland and A. Nadeau, 1998: Estimating maize production in Kenya using NDVI: some statistical considerations. *International Journal of Remote Sensing 19*, 2609-2617.

Li, A., S. Liang, A. Wang and J. Qin, 2007: Estimating crop yield from multi-temporal satellite data using multivariate regression and neural network techniques. *Photogrammetric Engineering & Remote Sensing 73(10),* 1149-1157.

Liaw, A. and M. Wiener, 2002: Classification and regression by random forest. *R News 2/3*, 18-22.

Mkhabela, M.S., M.S. Mkhabela and N.N. Mashinini, 2005: Early maize yield forecasting in the four agro-ecological regions of Swaziland using NDVI data derived from NOAA's-AVHRR. *Agricultural and Forest Meteorology 129*, 1-9.

Mo, X., S. Liu, Z. Lin, Y. Xu, Y. Xiang and T.R. McVicar , 2005: Prediction of crop yield, water consumption and water use efficiency with a SVAT-crop growth model using remotely sensed data on the North China Plain. *Ecological Modelling 183,* 301-322.

Ngie, A., F. Ahmed and K. Abutaleb, 2014: Remote sensing potential for investigation of maize production: review of literature. *South African Journal of Geomatics 3(2),* 163-184.

Noureldin, N.A., M.A. Aboelghar, H.S. Saudy, A.M. Ali, 2013: Rice yield forecasting models using satellite imagery in Egypt. *The Egyptian Journal of remote sensing and Space sciences 16*, 125-131.

Palmer, D.S., N.M. O'Boyle, R.C. Glen and B.O. Mitchell, 2007: Random forest models to predict aqueous solubility. *Journal of Chemical Information and Modelling 47,* 150-158.

Panda, S.S., D. P. Ames and S. Panigrahi, 2010: Application of vegetation indices for agricultural crop yield prediction using neural network techniques. *Remote Sensing 2*, 673-696; doi:10.3390/rs2030673.

Pontailler, J-Y., G.J. Hymus, and B.G. Drake, 2003: Estimation of leaf area index using ground-based remote sensed NDVI measurements: validation and comparison with two indirect techniques. *Canadian Journal of Remote Sensing 29(3),* 381-387.

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

Ren, J., Z. Chen, Q. Zhou and H. Tang, 2008: Regional yield estimation for winter wheat with MODIS-NDVI data in Shandong, China. *International Journal of Applied Earth Observation and Geoinformation 10,* 403-413.

Reynolds, M. and D.C. Yittayew, 2000: Slack, Estimation crop yields and production by integrating the FAO Crop Specific Water Balance model with real-time satellite data and ground-based ancillary data. *International Journal of Remote Sensing 21(18),* 3487-3508.

Roujean, J. and F. Breon, 1995: Estimating PAR absorbed by vegetation from bidirectional reflectance measurements. *Remote Sensing of Environment 51,* 375-384.

Rouse, J.W., R.H. Haas, J.A. Schell and D.W. Deering, 1974: Monitoring vegetation systems in the Great Plains with ERTS. In *Proceedings of Third ERTS Symposium,* Greenbelt, MD, December 1974; NASA SP-351-1, pp. 309-317.

Salazar, L., F. Kogan and L. Roytman, 2008: Using vegetation health indices and partial least squares method for estimation of corn yield. *International Journal of Remote Sensing 28,* 175-189.

Sau, F., K.J. Boote, W.M. Bostick, J.W. Jones and M.I. Minguez, 2004: Testing and improving evapotranspiration and soil water balance of the DSSAT crop models. *Agronomy Journal 96,* 1243-1257.

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

Singh, R., D.P. Semwal, A. Rai and R.S. Chhikara, 2002: Small area estimation of crop yield using remote sensing satellite data. *International Journal of Remote Sensing 23,* 49-56.

Strobl, C., A-L. Boulesteix, T. Kneib, T. Augustin and A. Zeileis, 2008: Conditional variable importance for random forests. BMC Bioinformatics, available online at: http://www.biomedcentral.com/1471-2105/9/307 Accessed 30/03/2015.

Strobl, C., Boulesteix, A., A. Zeileis and T. Hothorn, 2007: Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics, 8*: 1-21.

Strobl, C., T. Hothorn and A. Zeileis, 2009: Party on! A New, Conditional variable-importance measure for random forests available in the party package. The R Journal Vol. 1/2, December 2009 ISSN 2073-4859 pg 14-17. http://journal.r-project.org/archive/2009-2/RJournal_2009-2_Strobl~et~al.pdf, Accessed 15/04/2015.

Sun, J., 2000: Dynamic monitoring and yield estimation of crops by mainly using the remote sensing technique in China. *Photogrammetry Engineering & Remote Sensing 66,* 645-650.

Thenkabail, P. S., 2003: Biophysical and yield information for precision farming from near-real-time and historical Landsat TM images. *International Journal of Remote Sensing 24,* 2879-2904.

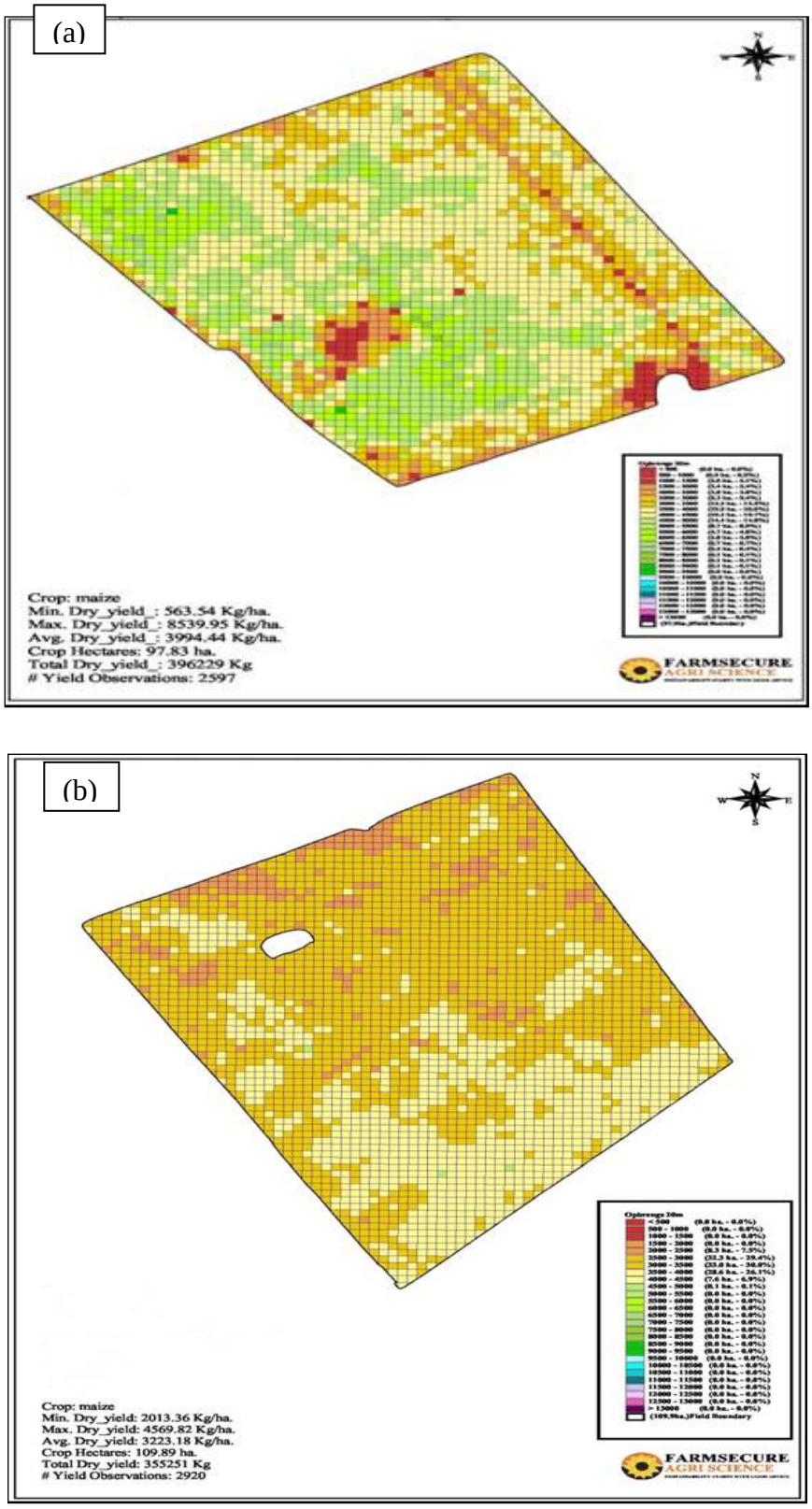
Tucker, C.J., 1979: Red and photographic infrared linear combinations for monitoring vegetation, *Remote Sensing of Environment 8(2),* 127-150.

You, J. X. Li, M. Low, D. Lobell and S. Ermon, 2017: Deep gaussian process for crop yield prediction based on remote sensing data. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17). February 4–9 at the Hilton San Francisco, San Francisco, California, USA. pg 4559-4565. https://cs.stanford.edu/~ermon/group/website/papers/jiaxuan_AAAI17.pdf Accessed 18/10/2017.

Zere, T.B., C.W. van Huyssteen and M. Hensley, 2004: Development of a simple empirical model for predicting maize yields in a semi-arid area. *South African Journal of Plant and Soil 22(1),* 22-27.

Zhao, J., K. Shi, and F. Wei, 2007: Research and application of remote sensing techniques in Chinese Agricultural Statistics. Paper presented at the Fourth International Conference on Agricultural Statistics, October 22-24, Beijing, China. www.stats.gov.cn/english/icas/papers/P020071017422431720472.pdf.

Appendix A



Spatial variation of maize grain yields across the (a) Cairo (b) Agnes fields (red is lowest and purple is highest)