PFG 2016 / 5–6, 285–299
Stuttgart, December 2016

Important Variables of a RapidEye Time Series for Modelling Biophysical Parameters of Winter Wheat

Thorsten Dahms, Sylvia Seissiger, Würzburg, Erik Borg, Hermann Vajen, Bernd Fichtelmann, Neustrelitz & Christopher Conrad, Würzburg

Keywords: biophysical parameter, RapidEye, vegetation indices, winter wheat, phenology, conditional inference forest

Summary: With the increasing availability of high resolution data, remote sensing is gaining importance for agricultural management. Sensor constellations such as RapidEye or Sentinel-2 have a strong potential for precision agriculture because they provide spectral information throughout the cropping season and at the subfield level. To explore this potential, methods are required that accurately transfer the spectral information into biophysical parameters which in turn permit quantitative assessments of plant growth on the field. Boundary condition for a successful monitoring, e.g., a repeated derivation of the biophysical parameters is to cope with the challenge of enormous data amounts, i.e. to select the input data that is most relevant.

In this study, biophysical parameters of winter wheat, namely the fraction of absorbed photosynthetic active radiation (FPAR), the leaf area index (LAI) and the chlorophyll content (expressed by SPAD), were modelled with RapidEye data in Mecklenburg-West Pomerania, Germany, using Random Forest based on conditional inference trees. Focus was set at the selection of the most important information out of spectral bands and indices for parameter prediction on winter wheat. Insitu and remote sensing observations were grouped into phenological phases in order to examine the importance of single spectral bands or indices for modelling biophysical reality in the several growing stages of winter wheat. The coefficient of determination for FPAR (LAI; SPAD) ranged between 0.19 and 0.83 (0.33 and 0.66; 0.21 and 0.45). Model accuracy was linked with the phenological phase. The results showed that for each biophysical parameter, different spectral variables become important for modelling and the number of important variables depends on the phenological time span. The prediction of biophysical parameters for short phenological groups often depends only on one to three variables. The results also showed that in the phenological phase of fruit development, the model accuracy is the lowest and the determination of the importance is comparatively vague.

Zusammenfassung: *Wichtige Variablen aus RapidEye-Zeitreihen für die Modellierung biophysikalischer Parameter von Winterweizen*. Hochaufgelöstes Monitoring agrarwirtschaftlicher Flächen gewinnt immer mehr an Bedeutung. Aus fernerkundlicher Sicht beruht dieses Monitoring auf der robusten Ableitung verschiedener biophysikalischer Parameter aus räumlich und zeitlich hoch aufgelösten Fernerkundungsdaten, z.B. RapidEye oder Sentinel-2. Ziel aktueller Forschung ist es, die biophysikalischen Parameter FPAR (Fraction of Absorbed Photosynthetic Active Radiation), LAI (Leaf Area Index) und den Chlorophyllgehalt aus fernerkundlichen Daten zu ermitteln. Hierbei reizen die großen Datenmengen häufig die Berechnungskapazitäten aus. Somit wird eine umsichtige Reduzierung der zu verarbeitenden Datenmenge die Anwendbarkeit dieser Methode verbessern.

In der vorliegenden Studie wurden conditional inference Random Forests eingesetzt, um zum einen die biophysikalischen Parameter unter Verwendung von RapidEye Szenen zu modellieren, und zum anderen die Bedeutung der einzelnen Eingangsparameter (Spektrale Bänder des RapidEye und Vegetationsindizes) zu quantifizieren. Die direkt auf dem Feld und die fernerkundlich erhobenen Beobachtungen des Winterweizens wurden in unterschiedliche Entwicklungsstadien (phänologische Gruppen) eingeteilt. Bei der Modellierung des FPAR (LAI; SPAD) wurden hierbei Bestimmtheitsmaße zwischen 0.19 und 0.83 (0.33 und 0.66; 0.21 und 0.45) erreicht. Dies zeigt, dass die Genauigkeit der Modellierung der jeweiligen biophysikalischen Parameter stark von der entsprechenden

© 2016 E. Schweizerbart'sche Verlagsbuchhandlung, Stuttgart, Germany www.schweizerbart.de
DOI: 10.1127/pfg/2016/0303 1432-8364/16/0303 $ 3.75

phänologischen Gruppe abhängt. Darüber hinaus zeigen die Ergebnisse, dass die Bedeutung der unterschiedlichen Eingangsparameter für die unterschiedlichen biophysikalischen Parameter und unterschiedlichen Entwicklungsstadien stark unterschiedlich ist. Häufig sind es nur bis zu drei spektrale Variable, die einen Parameter in den kurzen Entwicklungsphasen beschreiben. Die Ergebnisse zeigen auch, dass das Modellieren biophysikalischer Parameter im phänologischen Stadium der Fruchtreife am ungenauesten ist.

1 Introduction

Recently launched and upcoming satellite missions like the Sentinel systems will highly increase the amount of spatiotemporal data provided by remote sensing (Bontemps et al. 2015). This kind of high resolution data offers great opportunities among others in agriculture (Franke & Menz 2007). Remote sensing based information of high spatial and temporal resolution can for instance be beneficial for agricultural applications like precision farming and crop yield estimation (Haboudane et al. 2004, Ahmadian et al. 2016). These applications demand accurate and up to date information on the vegetation (Jin et al. 2013), e.g. on the phenological state and on vegetation growth such as biomass production, e.g. expressed by absorbed photosynthetically active radiation (FPAR), the leaf area index (LAI), or chlorophyll content. One example is the study of Eitel et al. (2007), where the nitrogen status of winter wheat was predicted to support farmers with the information whether to apply supplemental fertilizer during the growing period of the crop. However, such applications useful for precision agriculture are still rare.

In order to observe and analyse vegetation using biophysical parameters, several remote sensing approaches were proposed in the past (Hall et al. 1995, Mutanga & Skidmore 2004, Le Maire et al. 2011). One option is empirical modelling, i.e. the identification of an optimal statistical relation between spectral measurements, e.g. vegetation indices, and in situ observations. The suitability of empirical approaches varies among the biophysical parameters because they vary in their complexity. Linear statistical approaches may be sufficient for the derivation of FPAR at least for some crops (Myneni & Williams 1994, Lex et al. 2013). However, e.g. for the derivation of LAI, there are strong indications that one vegetation index or spectral band cannot explain the biophysical reality of the vegetation cover over the entire growing season (Viña et al. 2011, Lex et al. 2013), because the physical appearance of the crop and, moreover, canopy parameters like cover fraction and plant height vary with the phenological stages of crops. Thus and not exclusively for crops, different univariate and multivariate, linear and non-linear statistical methods have been applied for monitoring biophysical parameters of vegetation with high-resolution data. Machine learning algorithms such as the Random Forest algorithm (Breiman 2001) are typically able to cope with a strong non-linearity of the functional dependence between some biophysical parameters and the reflected spectra (Beckschaefer et al. 2014). Differentiation among different phenological stages could also improve empirical estimations of biophysical parameters of vegetation, at least for some growing stages of vegetation, as e.g. shown by Tillack et al. (2014) or Lex et al. (2015). Nevertheless, little attention has been put on the derivation of biophysical parameters using high resolution remote sensing data in combination with machine learning algorithms for crop monitoring at different stages of the vegetation period.

One challenge to increase the practical use of remote sensing based information products for precision agriculture is the enormous expenditure (e.g. data amount, storage space, processing time), which is necessary for the derivation of the relevant biophysical parameters. To minimize this aspect the reduction of the spectral resolution, e.g. by composing spectral indices or band selection can be useful and information is required, which indices and spectral bands have the most effect on modelling biophysical parameters at which growing stage. Machine learning methods provide an assessment of the so-called variable importance, which returns the relevance and suitability of certain spectral bands and indices for accurate modelling of biophysical parameters. Beckschaefer et al. (2014) demonstrated the usability of the variable importance when linking remote sensing observations with biophysical parameters for subtropical upland ecosystems.

Different remote sensing applications deal with the extraction of variable importance from Random Forests (Mutanga et al. 2012, Beckschaefer et al. 2014). However, Strobl et al. (2007) pointed out that an analysis of causal effects using the classical Random Forest approach can be biased in case of having correlated regressors. Against this background, Strobl et al. (2008) introduced the conditional variable importance method to determine the variable importance for correlated regressors. In cause-effect analyses based on Random Forest, in which remote sensing data is utilized, this conditional variable importance method is critical, because spectral bands or e.g. vegetation indices are commonly highly correlated.

The aims of this study are (i) to predict biophysical parameters, namely FPAR, LAI, chlorophyll content of winter wheat during the different growing stages using RapidEye time series and in-situ data, (ii) to identify the most important spectral bands or indices for modelling these biophysical parameters and (iii) to investigate how the indicator importance of these variables changes in the phenological cycle.

2 Study Area

The study area was located near the town Demmin in Mecklenburg-West Pomerania (Mecklenburg-Vorpommern) in Northeast-Germany (Fig. 1). The landscape was formed by glaciers and melting waters during the Weichsel glacial period, approximately 10,000 years ago. The northern part of the study area is characterized by low topographical variations between 5 m – 20 m a.s.l. whereas the south can be described as hilly to undulating. Due to significant differences in parent substrate material and topography, soils are primarily loamy sands and sandy loams alternating with pure sand patches or clayey areas (Gerighausen et al. 2009). The climate is moderate, with an average annual temperature of 8–8.5 °C and an average annual rainfall of 550 mm – 600 mm (Borg et al. 2009). The investigated fields were located within the test site DEMMIN (Durable Environmental Multidisciplinary Monitoring Information Network), one of four test areas of the TERENO lowland observatory (Borg et al. 2009, HGF 2015). The test site is an intensively used agricultural ecosystem.

3 Data and Methods

3.1 *RapidEye*

The RapidEye satellite system is a constellation of five identical earth observation satellites in one orbit with the capability to provide multi-spectral images over large areas with frequent revisits at high resolution (6.5 m at nadir). A detailed description of the RapidEye system can be found in Borg et al. (2013). In addition to the Blue (B) (440 nm – 510 nm), Green (G) (520 nm – 590 nm), Red (R) (630 nm – 685 nm) and Near-Infrared (NIR) (760 nm – 850 nm) bands, the sensor has a RedEdge (RE) (690 nm – 730 nm) band, especially suitable for vegetation analysis (Kross et al. 2015). The RapidEye level 3A standard product covers an area of 25 km × 25 km, is radiometrically calibrated to spectral radiance, as well as orthorectified and resampled to 5 m spatial resolution (Chander et al. 2013). In this study a time series of nine RapidEye images was available. It was recorded within the growing period of winter wheat in 2015. The acquisition dates are given in Fig. 3.

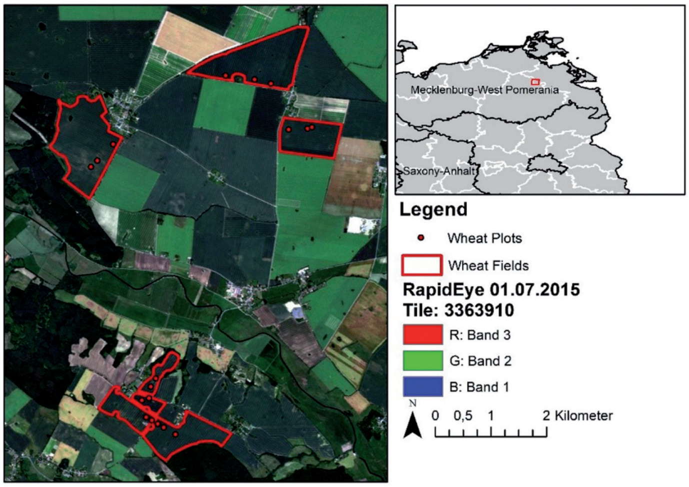


Fig. 1: Study area and location of the Environmental Sampling Units (ESU) in the winter wheat fields.

3.2 *In-situ Observations*

In diverse studies different biophysical key parameter of interest for precision farming applications were identified (Moran et al. 1997, Baret et al. 2007). Incoming Photosynthetic Active Radiation (PAR) is the primary driving force of photosynthesis and biological production. The Fraction of Photosynthetic Active Radiation (FPAR) resembles the fraction of absorbed incoming Photosynthetic Active Radiation (APAR) in relation to the available PAR and is a key input for light used efficiency modelling (LUE) (Seaquist et al. 2003). The LAI characterizes the leaf surface available for energy and mass exchange between surface and atmosphere (Carlson & Ripley 1997). Chlorophyll content can be considered as one of the main inputs in the vegetation models development. Thus, it is considered to be an indicator of the photosynthetic efficiency of the plant (Darvishzadeh et al. 2008). These three key biophysical variables were investigated in the presented study.

The field survey concept was to gather FPAR, LAI and chlorophyll information in a weekly to bi-weekly recurrence. FPAR and LAI were measured using a SunScan instrument (Delta-T Devices Ltd., Cambridge, England) and SPAD (Soil & Plant Analyzer Development) values were measured using a handheld chlorophyll meter (SPAD-502, Minolta Osaka Company, Ltd., Osaka, Japan). The data used in this study was collected on 18 Environmental Sampling Units (ESUs) (Baret et al. 2002) on seven winter wheat fields. The EUSs have an extent of 20 m × 20 m. Within each ESU, twelve measurement points were set within a rectangular cross. The twelve measurements over one ESU were averaged. FPAR and LAI were measured once on each point inside the ESU. The SPAD measurements were taken on each point ten times and averaged. A scheme of an ESU can be found in Fig. 2. The majority of the measurements were taken by the team of the calibration and validation site DEMMIN (Borg et al. 2009).

3.3 *Pre-processing*

An essential aspect, which substantially affects the accuracy of satellite-based remote sensing information, represents the pre-processing like e.g. geo- or atmospheric correction (Mannschatz et al. 2014). However, comparisons of the geographical coordinates of the ESUs recorded with a GPS during the field campaigns and the RapidEye data showed high accuracy in geolocation which in turn made further geo-corrections unnecessary. The RapidEye scenes were atmospherically

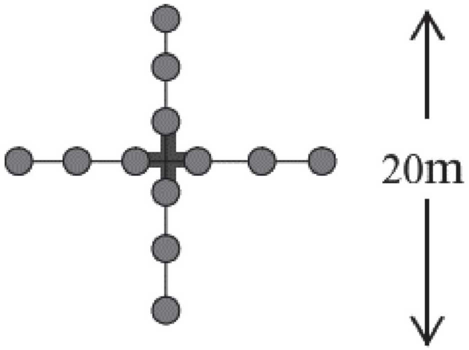


Fig. 2: ESU sampling scheme after GARRIGUES et al. (2002).

corrected and cloud masked using ATCOR2 (RICHTER 2010).

The reflectance spectrum of each RapidEye scene within all 18 ESUs was extracted by averaging RapidEye reflectance in a 20-meter radius around the centre of single ESU. This represents the spatial resolution of the new and upcoming Sentinel-2 Multi Spectral Instrument (MSI) data. Numerous vegetation indices comprising SR, NDVI, SAVI, RE_NDVI, RDVI and EVI were calculated (Tab. 1). In addition, the products of the tasselled cap transformation were included as they represent another important group of spectral indices in agriculture (SCHOENERT et al. 2014). With the additional RedEdge, the RapidEye system has been designed to derive information on the vegetation status (JUNG-ROTHENHÄUSLER et al. 2007). Thus, different vegetation indices, which consider the RedEdge (RE_NDVI, rel-Length, curve, Length; CONRAD et al. 2012) were also integrated in the feature set.

3.4 *Phenological Groups*

The second aim of this study was to investigate the variable importance of vegetation indices and single bands for the prediction of FPAR, LAI, and SPAD with respect to different phenological stages. Therefore, each dataset was grouped according to the phenology using the BBCH-characterization (Biologische **B**undesanstalt, **B**undessortenamt und **CH**emische Industrie) of the observation field data. The BBCH scale gives numeric information about the morphologic development stage of a plant (LANCASHIRE et al. 1991). The groups were named after the BBCH range they cover. The grouping of the data is illustrated in Fig. 4. This step ensured that the spectral behaviour was associated with the physical appearance of the plant and no longer with the data acquisition date. In other words, data pairs (field measurement and satellite observation at that point) were not analysed per data acquisition period, but within each phenological group.

3.5 *Conditional Inference Forest*

Random forests (BREIMAN 2001) are ensembles of classification and regression trees that operate on binary partitions of the feature space (drawn by the training samples). Each tree is built from nodes and leaves. Nodes consist of a predictor variable and a split val-

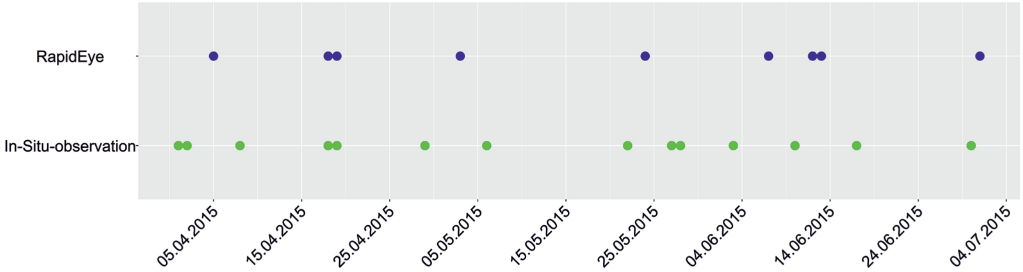


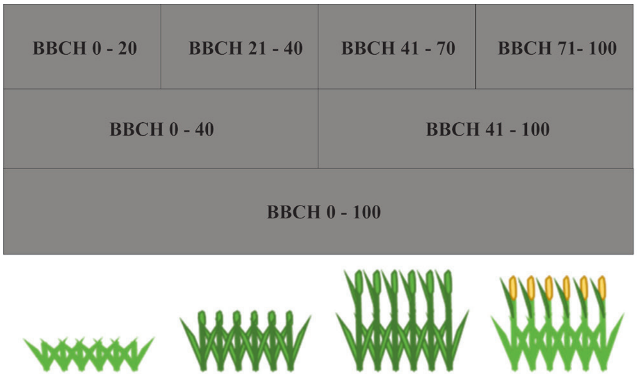
Fig. 3: Overview about the data acquisition times in the field and the available RapidEye observations.

Tab. 1: Overview of calculated vegetation indices. The bands are named according to the part of the spectra they represent. Note that λ refers to the central wavelength of the respective band, e.g. λ *Red* = 657.5 nm, λ *RedEdge* = 710 nm, λ *NIR* = 805 nm.

Index	Equation	Reference
TCT_B	$0.2435 * Blue + 0.3448 * Green + 0.4881 * Red + 0.4930$ $* RedEdge + 0.5835 * NIR$	Schoenert et al. 2014
TCT_G	$(-0.2216) * Blue + (-0.2319 * Green + (-0.4622) * Red$ $+(-0.2154)*RedEdge + 0.7981*NIR$	
TCT_Y	$(-0.7564) * Blue + (-0.3916) * Green + 0.5049 * Red$ $+ 0.1400 * RedEdge + 0.0064 * NIR$	
SR	$\frac{NIR}{Red}$	Jordan 1969
NDVI	$\frac{(NIR - Red)}{(NIR + Red)}$	Rouse et al. 1974
SAVI	$(1 + 0.5) * \frac{(NIR\text{-}Red)}{(NIR + Red + 0.5)}$	Huete 1988
RE_NDVI	$\frac{(NIR - RedEdge)}{(NIR + RedEdge)}$	Gitelson & Merzlyak 1996
RDVI	$\frac{(NIR - Red)}{\sqrt{(NIR + Red)}}$	Roujean & Breon 1995
EVI	$2.5 * \frac{(NIR - Red)}{(1 + NIR + 6 * Red - 7.5 * blue)}$	Huete et al. 2002
curv	$\frac{(\frac{(NIR - RedEdge)}{(\lambda NIR - \lambda RedRdge)}) - (\frac{(RedEdge - Red)}{(\lambda RedRdge - \lambda Red)})}{(\lambda NIR - \lambda Red)}$	Conrad et al. 2012
Length	$\sqrt{(NIR - RedEdge)^2 + (\lambda NIR\text{-}\lambda Rededge)^2 +}$ $\sqrt{(RedEdge - Red)^2 + (\lambda RedEdge - \lambda Red)^2}$	
relLength	$\frac{Length}{\sqrt{(NIR\text{-}Red)^2 + (\lambda NIR - \lambda Red)^2}}$	

ue (applied to that variable). Leaves comprise in case of using the regression variant samples of the predicted variable with at best very similar values to minimize the distribution of predicted values within one leaf is the major aim of the algorithm. Within random forest, bootstrapping is applied, i.e. each tree utilizes a subset of samples, and random selection of a limited number of features for generating the nodes that split the data into two groups each. Random forest has shown to be suitable to analyse enormous input datasets like multi-temporal satellite data (Rodriguez-Galiano et al. 2012). In this study, a further development of random forest, the so called conditional inference trees (cforest) were utilized. In cforest, the regression tree ensemble is built from conditional inference trees which are able to consider cause-effect relations during variable selection and to reduce bias in case of highly correlated variables (Strobl et al. 2008).

Classification and regression trees (and subsequently the random forest/cforest algorithms) select features which are optimal suited for modelling. The permutation of feature values allows for assessing the so called im-



0 - 100 BBCH	All stages
0 - 40 BBCH	Germination to Booting
41 - 100 BBCH	Booting to Senescence
0 - 20 BBCH	Germination to Tillering
21 - 40 BBCH	Tillering to Booting
41 - 70 BBCH	Booting to Development of fruit
71 - 100 BBCH	Development of fruit to Senescence

Fig. 4: Grouping of the dataset (field sampling and satellite observation at that point) into different classes of phenological appearance according to the BBCH-code.

portance of the feature. The variable importance is expressed using the difference between an internal prediction error of the random forest routine (based on the so called out-of-bag/OOB error; Breiman 2001) before and after the permutation of variable values in a predictor variable. If the permutation of variable values is reduced to those samples occurring in the branch of a tree (sub-tree) for which the variable is selected a more unbiased extraction of the variable importance becomes possible (Strobl et al. 2007).

Both, the cforest routine and the conditional variable importance algorithm as implemented in the 'party' package (Hothorn et al. 2010) of the statistic software R (R Core Team 2016) were utilized in this study. There is no effect of the number of trees on the average importance as long as the number of trees is sufficiently large to guarantee a stable estimate of the mean importance (Strobl & Zeileis 2008). Thus, and after explorative tests, the number of trees was set to 500.

The accuracy assessment for one cforest model run was conducted by calculating two statistical parameters. The root-mean-square error (RMSE) indicates the mean offset between the observed and the predicted data. The coefficient of determination (R^2) describes the percentage of variance that is explained by the model. In this study, a fivefold 10 times repeated cross validation was applied to calculate those two parameters. In contrast to the cross validation method using the OOB data, which is executed by the random forest algorithm internally, such an external cross validation is regarded to result in a more objective quality assessment of the model performance (Reunanen 2003).

The number of variables considered for each split within a single regression tree (mtry) can have great influence on the performance of cforest and the calculation of the variable importance (Díaz-Uriarte & De Andres 2006). Thus, each model was optimized. The CARET (Classification And REgression Training) package (Kuhn 2008) in the software R was used to tune the cforest models using 10 different mtry values (mtry = 2;3;5;7;8;10;12;13;15;17; note that 17 variables were totally available). Here, only the coefficient of determination (R^2) served as metric to compare the model qualities and hence to identify the optimal mtry value for modeling. The variable importance was calculated for the best performing model only. This procedure, i.e., tuning of cforest followed by the determination of variable importance for the optimal model, was repeated 100 times to assess the stability of the method in terms of absolute error and variable selection.

4 Results

4.1 *Prediction Accuracy*

Tab. 2 shows the cforest prediction accuracy of FPAR, LAI and SPAD in different phenological phases. The table depicts the average of the 100 best models, except for the mtry value, which represents the most often chosen value over the 100 runs. The highest R^2 value for FPAR ($R^2 = 0.83$) was achieved between 0 and 40 BBCH, while the lowest R^2 refers to the model between 41 and 70 BBCH ($R^2 = 0.19$). The best performance ($R^2 = 0.66$) for modelling LAI was also reached between 0 and 40 BBCH. The lowest accuracies ($R^2 = 0.33$) of the LAI models can be associated with the phenological groups of senescence (41 – 70 and 41 – 100 BBCH). Reduced accuracies were found for the SPAD models. They never outreached an R^2 value of 0.45 (21 – 40 BBCH). Due to limited in-Situ observations (N = 10) during the last BBCH stage (71 – 100 BBCH), the SPAD model for that period was not calculated.

4.2 *Variable Importance*

The cforest based variable importance for modelling the biophysical parameters (FPAR, LAI and SPAD) in every BBCH-based phenological group can be seen in Figs. 5 to 7. The boxplots show the unscaled variable importance for each of the 17 indices or bands received during all 100 model runs for one parameter and phenological group. The boxplots allow for comparing the variable importance of indices or bands used for modelling. High variable importance indicates an increase of the prediction error in cforest when the respective band or index is excluded. On the contrary, small or negative variable importance shows that omitting the tested band or index from cforest has none or negative impact on the model accuracy. The distribution of variable importance scores of the 100 model runs determines the size of the boxes, which in turn puts a light on the stability of the importance level of each index or band during modelling. For instance, a slim box indicates a more stable importance estimation of the respective index or band, a broad box suggest varying importance levels (relevance) of that variable over numerous runs.

Tab. 2: Cforest performance and final settings for FPAR LAI and SPAD according to the phenological groups (expressed by the BBCH-code).

	BBCH	**0 – 100**	**0 – 40**	**41 – 100**	**0 – 20**	**21 – 40**	**41 – 70**	**71 – 100**
FPAR	**RMSE**	0.16	0.12	0.04	0.16	0.12	0.04	0.16
	R^2	0.59	0.83	0.21	0.59	0.65	0.19	0.59
	mtry	12	5	17	12	5	17	12
	samples	124	68	56	25	43	36	20
LAI	**RMSE**	1.56	1.23	1.87	1.56	1.46	1.36	1.28
	R^2	0.41	0.66	0.33	0.41	0.57	0.33	0.41
	mtry	12	10	2	12	8	2	12
	samples	111	62	49	24	38	34	15
SPAD	**RMSE**	4.98	3.17	6.94	4.1	2.88	5.82	–*
	R^2	0.29	0.42	0.28	0.41	0.45	0.21	–*
	mtry	12	17	2	5	3	5	–*
	samples	161	103	58	34	69	48	–*

* Amount of in-situ data insufficient for modelling

4.2.1 FPAR

Fig. 5 depicts the variable importance of the single vegetation indices and spectral bands used for modelling FPAR. The distribution of the variable importance for the entire growing season (0 – 100 BBCH) in Fig. 5A shows that the most important indicators were the RedEdge band and the RE_NDVI. The variable importance plots for the phenological group 0 – 40 BBCH (Fig. 5B) indicate the RE_NDVI to be the most important variable followed by SAVI and NDVI. An atypical variable importance distribution occurred for the phenological group 41 – 100 BBCH (Fig. 5C). There, the SR and the EVI are listed as the most important variables. In this distribution nearly all indices associated with the RedEdge band have a negative impact on the model (variable importance below 0). The variable importance distribution of the shorter BBCH groups (0 – 20; 21 – 40; 41 – 70; 71 – 100 BBCH, Fig. 5D–G) resembles that of the longer phenological groups (0 – 40; 41 – 100 BBCH). Only, the boxplot referring to the 41 – 70 BBCH group exhibits an unusual distribution, compared to the other boxplots. This group is also the one with the lowest R² value.

4.2.2 LAI

The variable importance plots (Fig. 6) modelling the LAI showed that small groups of one to four indices are most important. For the entire growing season, the EVI showed the highest importance (Fig. 6A). The two most important variables for modelling the LAI in the first phenological group 0 – 40 BBCH (Fig.

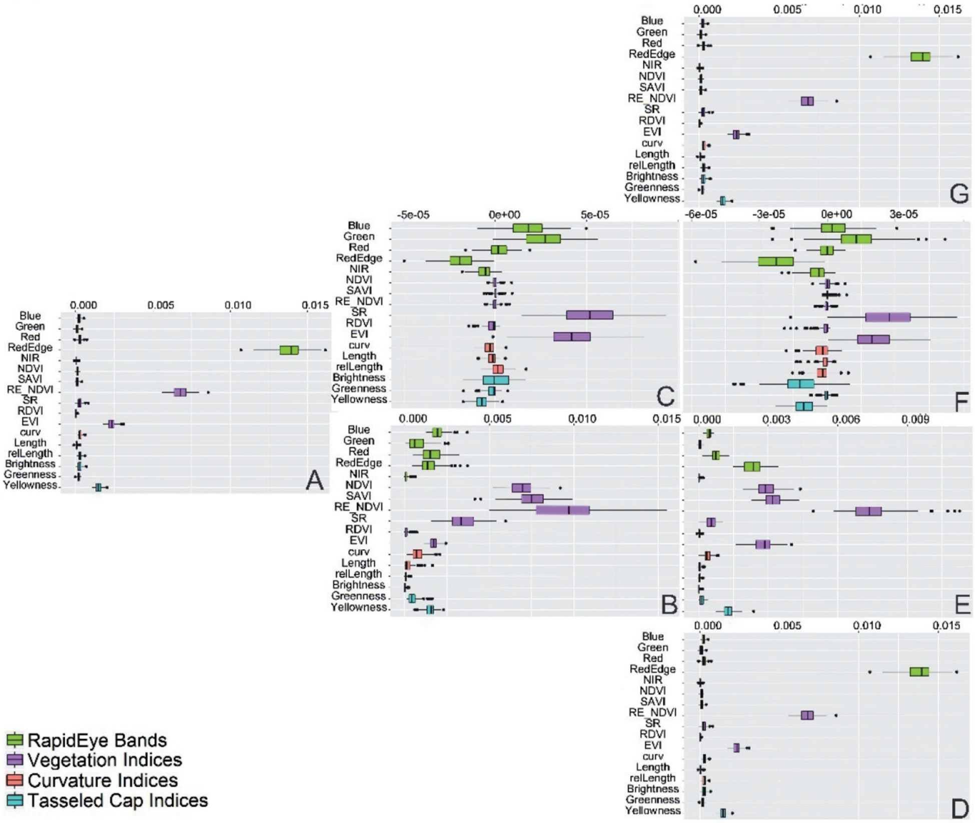


Fig. 5: FPAR variable importance distribution boxplots for different phenological groups (BBCH). A: 0 – 100; B: 0 – 40; C: 41 – 100; D: 0 – 20; E: 21 – 40; F: 41 – 70; G: 71 – 100.

6B) were RE_NDVI and EVI. The distribution of the variable importance for the phenological group 41 – 100 BBCH (Fig. 6C) highlights the TCT_Y index and the blue band as the most important variables. The plots of the shorter groups (Fig. 6D–G) show one to three vegetation indices (mainly EVI and RE_NDVI) to be the most important indices. An unusual variable importance distribution was received for the BBCH group 41 – 70 (similar to the group 41 – 100 BBCH).

4.2.3 SPAD

The RE_NDVI index is the most important variable to explain the SPAD in the 0 – 100 BBCH group (Fig. 7A) followed by the relLength. The RE_NDVI was the most important variable modelling the chlorophyll content in the phenological phases between 0 – 40 BBCH. Here, EVI and curv ranked on the second and third places, respectively. For modelling the SPAD value in the phenological group 41 – 70 BBCH (Fig. 6F), the SR index was identified to be the most important variable. The last phenological group (0 – 100 BBCH) could not be investigated, it was impossible to obtain enough field measurements in this group.

5 Discussion

Highest accuracies for modelling FPAR occurred during the vegetative phase (R^2= 0.83), whereas during the stages of the senescence reduced statistical relations were found. The phenological phase of fruit development was

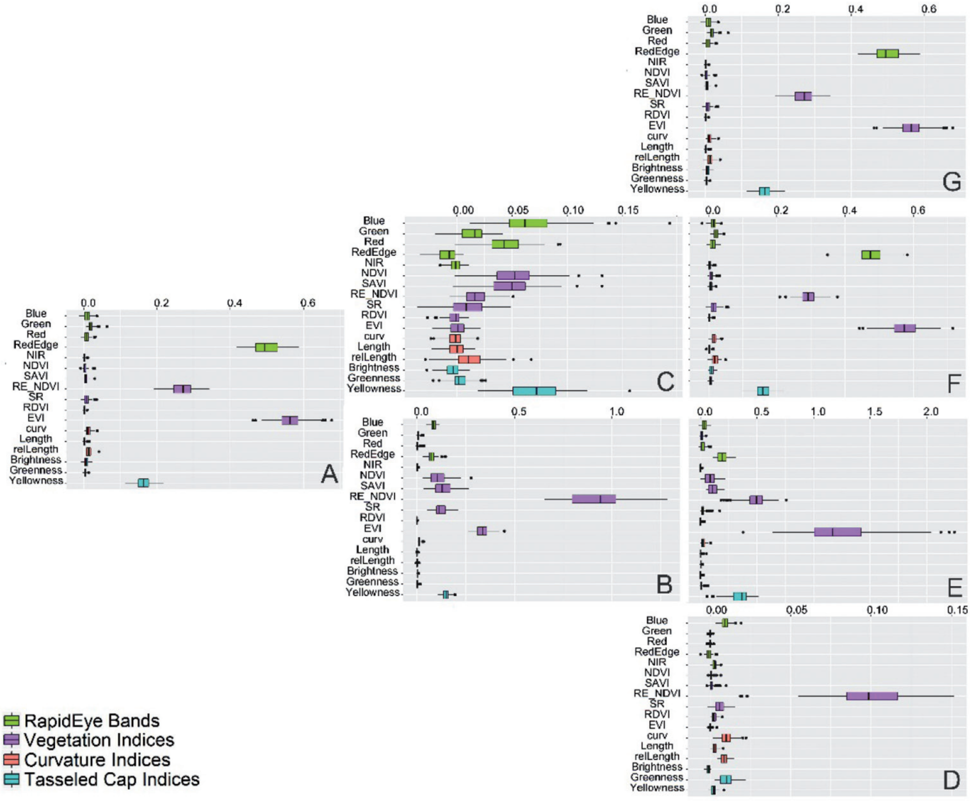


Fig. 6: LAI Variable importance boxplots for the different phenological groups (BBCH). A: 0 – 100; B: 0 – 40; C: 41 – 100; D: 0 – 20; E: 21 – 40; F: 41 – 70; G: 71 – 100.

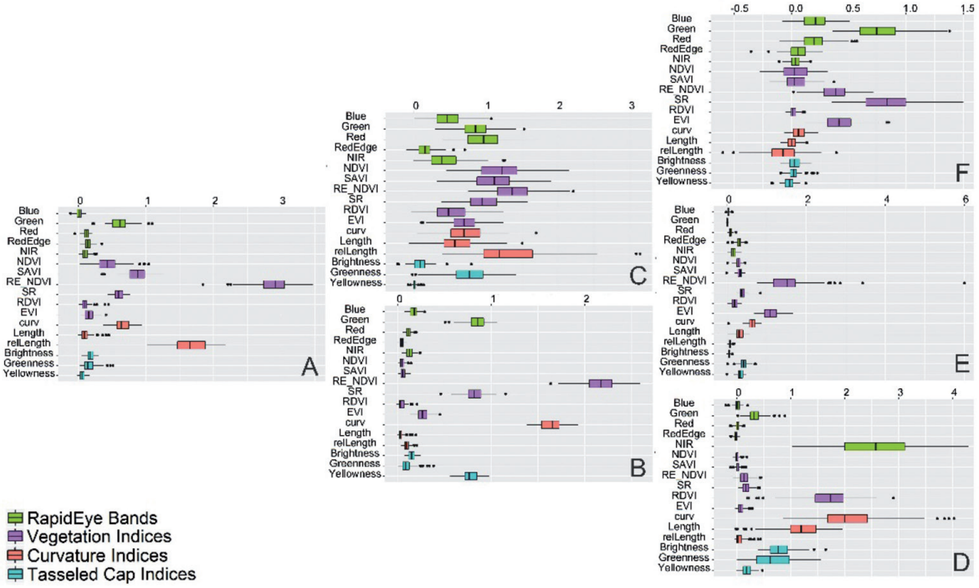


Fig. 7: SPAD variable importance distribution boxplots for the different phenological groups (BBCH). A: 0 – 100; B: 0 – 40; C: 41 – 100; D: 0 – 20; E: 21 – 40; F: 41 – 70.

nearby impossible to model with high accuracy ($R^2 = 0.19$), most likely due to canopy closure in combination with accompanied saturation effects in the RapidEye observations. The challenges modelling FPAR during the saturation or senescence phase is also highlighted by the by the variable importance distribution: While the variable importance of the initial growing stages shows three indices RE_NDVI, NDVI and SAVI to be the most important ones, the only vague patterns of variable importance were observed during the fruit development and the senescence phases. For the latter, the importance values were generally smaller and no group of important indices with distinct spectral properties emerged during analysis.

In comparison to FPAR only slightly reduced modelling accuracies were found when modelling the LAI. The accuracy levels were comparable with the accuracies Zhao et al. (2015), who modelled the LAI of wheat using univariate regressions and the HJ-1 sensor system and achieved a R^2 of 0.58 and RMSE values ranging from 0.7 to 0.89. However, in contrast to the FPAR results the variable importance plots for LAI indicate a more distinct distribution among the detailed phenological groups. There, a group of one to four indices were found to be most important for the cforest model. The observation that EVI is the highest ranked index among the phenological stages of growth, fruit development and senescence can be explained by a higher robustness of that index against saturation effects that occur in these phenological phases due to the closed canopy (Huete et al. 2002).

The SPAD cforest models reached lowest accuracy levels in this study ($R^2 \leq 0.45$). Again, the models of the phenological groups linked to early growth phases were the statistically best performing. In comparison, Schoenert et al. 2014 modelled the chlorophyll content using RapidEye with a R^2 of 0.77 on wheat using tasselled cap transformations. Eitel et al. (2007) used different vegetation indices calculated from spectrometer measurements to model SPAD values and reported R^2 values between 0.01 and 0.77. The comparatively low performance in this study may be explained by the temporal offset between the in situ observation and the satellite data acquisition of four days which may be too long for modelling a biophysical parameter like chlorophyll content. The RedEdge band and related indices always ranked under the most important variables for the prediction of SPAD values. This confirms the usefulness of analysing chlorophyll content in the RedEdge spectra as demonstrated previously (Eitel et al. 2007). The observation that spectral curvature indices can contribute to successful modelling of chlorophyll content is in line with the results presented by Eitel et al. (2007) based on simulated RapidEye and hyperspectral data for wheat.

6 Conclusion

Remote sensing applications for farmers like precision farming demand up to date information on the crop in specific phenological phases. Several field management methods like the application of fertilizers depend on the phenological phase of the plant. This study addressed the utility of RapidEye data and the use of machine learning for obtaining growth information about winter wheat in different phenological stages and to show how the variable importance changes along with the phenology. Thereby, the cforest was found suitable to model biophysical parameters for the entire growing season and to get an increased understanding about variables useful for predictions. Several vegetation indices were identified to be very important for the derivation of the biophysical parameters FPAR, LAI and chlorophyll content (approximated with the SPAD-value).

The model performance for the entire growing season outreached that for single phenological groups. There, the vegetative phase (0 – 40 BBCH) showed the best performance and more stable variable importance distribution, particular in contrast to the senescence phase (70 – 100 BBCH). Models with a high accuracy relied on a small set of input parameters only. The latter may allow for questioning the use of more complex approaches to model biophysical parameters of winter wheat and crops with similar physical appearance (e.g. other cereal crops).

The variable importance varied among the biophysical parameters and the phenological stages, which in turn indicates a link with the physical appearance of wheat during the cropping season. Nevertheless, altogether, the RE_NDVI or the RDVI were found to be the most important variables which in turns underlines the importance of RedEdge bands for modelling biophysical parameters of crops, at least those of winter wheat.

Cforest was applied to one ensemble of vegetation indices and single bands of the RapidEye system. Even though some features repeatedly showed high variable importance, the results may have varied in case other sensor systems, e.g. Sentinel-2, acquisition dates, or spectral features have been included. Such considerations have to be taken into account in further research and discussions about the transferability of the approach. Nevertheless, indication for the selection of the important features is given, because wheat represents a single plant type with a closed canopy. This information can in turn contribute to reduce the computation efforts, which is of paramount importance with a look on the continuously increasing data amount at the high resolution remote sensing sector.

Acknowledgements

We express our gratitude to the team of the Calibration and Validation Facility DEMMIN of the German Aerospace Center (DLR) who commonly conducted the field observations throughout the vegetation period accompanied by students of the University of Würzburg. The Blackbridge Company (RESA Project 0028) provided the RapidEye data for this study. The Federal Ministry of Economic Affairs and Energy funded this research (FKZ: 50 EE 1353).

References

Ahmadian, N., Ghasemi, S., Wigneron, J.P. & Zoelitz, R., 2016: Comprehensive study of the biophysical parameters of agricultural crops based on assessing Landsat 8 OLI and Landsat 7 ETM+ vegetation indices. – GIScience & Remote Sensing 53 (3): 337–359.

Baret, F., Houles, V. & Guerif, M., 2007: Quantification of plant stress using remote sensing observations and crop models: the case of nitrogen management. – Journal of Experimental Botany 58 (4): 869–880.

Beckschaefer, P., Fehrmann, L., Harrison, R.D., Xu, J. & Kleinn, C., 2014:. Mapping Leaf Area Index in subtropical upland ecosystems using RapidEye imagery and the randomForest algorithm. – iForest-Biogeosciences and Forestry 7 (1): 1.

Bontemps, S., Arias, M., Cara, C., Dedieu, G., Guzzonato, E., Hagolle, O., Inglada, J., Matton, N., Morin, D., Popescu, R., Rabaute, T., Savinaud, M., Sepulcre, G., Valero, S., Ahmad, I., Bégué, A., Wu, B., Abelleyra, D., Diarra, A., Dupui, S., French, A., Akhtar, I., Kussul, N., Lebourgeois, V., Le Page, M., Newby, T., Savin, I., Verón, S., Koetz, B. & Defourny, P., 2015: Building a data set over 12 globally distributed sites to support the development of agriculture monitoring applications with Sentinel-2. – Remote Sensing 7 (12): 16062–16090.

Borg, E., Lippert, K., Zabel, E., Loepmeier, F.J., Fichtelmann, B., Jahncke, D. & Maass, H., 2009: DEMMIN – Teststandort zur Kalibrierung und Validierung von Fernerkundungsmissionen. – Rebenstorf, R.W. (Hrsg.): 15 Jahre Studiengang Vermessungswesen – Geodätisches Fachforum und Festakt, Neubrandenburg, Eigenverlag: 401–419.

Borg, E., Daedelow, H., Missling, K.-D. & Apel, M., 2013: RapidEye Science Archive: Remote Sensing Data for the German Scientific Community. – Borg, E., Daedelow, H. & Johnson, R. (eds.): 5th RESA Workshop "Data for Science: From the Basics to the Service": 5–20, Neustrelitz, 20.–21.3.2012, GITO mbH Verlag, Berlin, ISBN 978-3-95545-002-1.

Breiman, L., 2001: Random forests. – Machine Learning 45 (1): 5–32.

Carlson, T. & Ripley, D., 1997: On the relation between NDVI, fractional vegetation cover, and leaf area index. – Remote Sensing of Environment 62 (3): 241–252.

Chander, G., Haque, M.O., Sampath, A., Brunn, A., Trosset, G., Hoffmann, D. & Anderson, C., 2013: Radiometric and geometric assessment of data from the RapidEye constellation of satellites. – International Journal of Remote Sensing 34 (16): 5905–5925.

Conrad, C., Fritsch, S., Lex, S., Loew, F., Ruecker, G., Schorcht, G. & Lamers, J., 2012: Potenziale des Red Edge Kanals von RapidEye zur Unterscheidung und zum Monitoring landwirtschaftlicher Anbaufrüchte am Beispiel des usbekischen wässerungssystems Khorezm. – Borg, E., Daedelow, H. & Johnson, R. (eds.): RapidEye Science Archive (RESA) – Vom Algorithmus zum Produkt, 4. RESA Workshop (pp. 203–217). – GITO, Berlin.

Darvishzadeh, R., Skidmore, A., Schlerf, M. & Atzberger, C., 2008: Inversion of a radiative transfer model for estimating vegetation LAI and chlorophyll in a heterogeneous grassland. – Remote Sensing of Environment 112 (5): 2592–2604.

Díaz-Uriarte, R. & De Andres, S.A., 2006: Gene selection and classification of microarray data using random forest. – BioMed Central Bioinformatics 7 (1): 1.

Dong, T., Meng, J. & Wu, B., 2012: Overview on methods of deriving fraction of absorbed photosynthetically active radiation (FPAR) using remote sensing. – Shengtai Xuebao/Acta Ecologica Sinica 32 (22): 7190–7201.

Eitel, J.U.H., Long, D.S., Gessler, P.E. & Smith, A.M.S., 2007: Using in- situ measurements to evaluate the new RapidEy satellite series for prediction of wheat nitrogen status. – International Journal of Remote Sensing 28 (18): 4183–4190.

Franke, J. & Menz, G., 2007: Multi-temporal wheat disease detection by multi-spectral remote sensing. – Precision Agriculture 8 (3): 161–172.

Garrigues, S., Allard, D., Weiss, M. & Baret, F., 2002: Comparing VALERI sampling schemes to better represent high spatial resolution satellite pixel from ground measurements: How to characterize an ESU. – http://w3.avignon.inra.fr/valeri/methodology/samplingschemes.pdf (19.10.2016).

Gerighausen, H., Borg, E., Fichtelmann, B., Guenther, A., Vajen, H.H., Wloczyk, C., Maass, H., 2009: Validation and calibration of remote sensing data products on test site DEMMIN. – 43. Ziolkowski Conference (pp. 18–33). – Russische Akademie der Wissenschaften.

Gitelson, A.A. & Merzlyak, M., 1996: Signature analysis of leaf reflectance spectra: algorithm development for remote sensing of chlorophyll. – Journal of Plant Physiology 148 (3): 494–500.

Haboudane, D., Miller, J.R., Pattey, E., Zarco-tejada, P.J. & Strachan, I.B., 2004: Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: Modeling and validation in the context of precision agriculture. – Remote Sensing of Environment 90 (3): 337–352.

Hall, F.G., Townshend, J.R. & Engman, E.T., 1995: Status of remote sensing algorithms for estimation of land surface state parameters. – Remote Sensing of Environment 51 (1): 138–156.

HGF, 2015: TERENO Northeastern German Lowland Observatory. – http://teodoor.icg.kfa-juelich.de/observatories/GL_Observatory?set_language=en (8.5.2016).

Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M. & Hofner, B., 2010: Model-based boosting 2.0. – Journal of Machine Learning Research 11: 2109–2113.

Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X. & Ferreira, L.G., 2002: Overview of the radiometric and biophysical performance of the MODIS vegetation indices. – Remote Sensing of Environment 83 (1): 195–213.

Huete, A., 1988: A soil-adjusted vegetation index (SAVI). – Remote Sensing of Environment 25 (3): 295–309.

Jin, X.L., Diao, W.Y., Xiao, C.H., Wang, F.Y., Chen, B., Wang, K.R. & Li, S.K., 2013: Estimation of wheat agronomic parameters using new spectral indices. – PLOS ONE 8: e72736 doi: 10.1371/journal.pone.0072736

Jordan, C.F., 1969: Derivation of leaf are index from quality of light on the forest floor. – Ecology 50: 663–666.

Jung-Rothenhäusler, F., Weichelt, H. & Pach, M., 2007: RapidEye. A novel approach to space borne geo-information solutions. – International Society for Photogrammetry and Remote Sensing, Hanover Workshop, 29.05. – 01.06.2017.

Kross, A., McNairn, H., Lapen, D., Sunohara, M. & Champagne, C., 2015: Assessment of RapidEye vegetation indices for estimation of leaf area index and biomass in corn and soybean crops. – International Journal of Applied Earth Observation and Geoinformation 34: 235–248.

Kuhn, M., 2008: Building predictive models in R using the caret package. – Journal of Statistical Software 28 (5).

Lancashire, P.D., Bleiholder, H., Langelueddecke, P., Stauss, R., Van Den Boom, T., Weber, E. & Witzenberger, A., 1991: An uniform decimal code for growth stages of crops and weeds. – Annals of Applied Biology 119: 561–601.

Le Maire, G., Marsden, C., Verhoef, W., Ponzoni, F.J., Seen, D.L., Bégué, A. & Nouvellon, Y., 2011: Leaf area index estimation with MODIS reflectance time series and model inversion during full rotations of Eucalyptus plantations. – Remote Sensing of Environment 115 (2): 586–599.

Lex, S., Conrad, C. & Schorcht, G., 2013: Analyzing the seasonal relations between in situ fpar/LAI of cotton and spectral information of RapidEye. – Borg, E., Daedelow, H. & Johnson, R. (eds.): RapidEye Science Archive (RESA) – From the Basics to the Service. – GITO, Berlin.

Lex, S., Asam, S., Löw, F. & Conrad, C., 2015: Comparison of two Statistical Methods for the Derivation of the Fraction of Absorbed Photosynthetic Active Radiation for Cotton. – PFG – Photogrammetrie, Fernerkundung, Geoinformation 2015 (1): 55–67.

Mannschatz, T., Pflug, B., Borg, E., Feger, K.H. & Dietrich, P., 2014: Uncertainties of LAI estimation from satellite imaging due to atmospheric correction. – Remote Sensing of Environment 153: 24–39.

Moran, M.S., Inoue, Y. & Barnes, E.M., 1997: Opportunities and limitations for image-based remote sensing in precision crop management. – Remote sensing of Environment 61 (3): 319–346.

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

Mutanga, O. & Skidmore, A.K., 2004: Narrow band vegetation indices overcome the saturation problem in biomass estimation. – International Journal of Remote Sensing 25 (19): 3999–4014.

Myneni, R.B. & Williams, D.L., 1994: On the relationship between FAPAR and NDVI. – Remote Sensing of Environment 49 (3): 200–211.

R Core Team (2014). R: A language and environment for statistical computing. – R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/ (20.10.2016).

Reunanen, J., 2003: Overfitting in making comparisons between variable selection methods. – Journal of Machine Learning Research 3: 1371–1382.

Richter, R. (2010). Atmospheric/Topographic Correction for Satellite Imagery (ATCOR-2/3 User Guide, Version 7.1, January 2008), 165.

Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M. & Rigol-Sanchez, J.P., 2012: An assessment of the effectiveness of a random forest classifier for land-cover classification. – ISPRS Journal of Photogrammetry and Remote Sensing 67: 93–104.

Roujean, J.L. & Breon, F.M., 1995: Estimating PAR absorbed by vegetation from bidirectional reflectance measurements. – Remote Sensing of Environment 51 (3): 375–384.

Rouse, J.W., Haas, R.H., Schell, J.A., Deering, D.W. & Harlan, J.C, 1974: Monitoring the vernal advancement of retrogradiation of natural vegetation. – NASA/GSFC, Type III, Final Report: 371, Greenbelt, MD, USA.

Schoenert, M., Weichelt, H., Zillmann, E. & Juergens, C., 2014: Derivation of tasseled cap coefficients for RapidEye data. – SPIE 9245, Earth Resources and Environmental Remote Sensing/GIS Applications V, 92450Q (October 23, 2014); doi: 10.1117/12.2066842.

Seaquist, J.W., Olsson, L. & Ardoe, J., 2003: A remote sensing-based primary production model for grassland biomes. – Ecological Modelling 169 (1): 131–155.

Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T. & Zeileis, A., 2008: Conditional variable importance for random forests. – BioMed Central Bioinformatics 9 (1): 25; doi: 10.1186/1471-2105-9-307.

Strobl, C., Boulesteix, A.L., Zeileis, A. & Hothorn, T., 2007: Bias in random forest variable importance measures – Illustrations, sources and a solution. – BioMed Central Bioinformatics 8 (1): 307; doi: 10.1186/1471-2105-8-25.

Strobl, C. & Zeileis, A., 2008: Danger: high power!-exploring the statistical properties of a test for random forest variable importance. – Technical Report Number 017, 2008, Department of Statistics University of Munich, Munich.

Tillack, A., Clasen, A., Kleinschmit, B. & Foerster, M., 2014: Estimation of the seasonal leaf area index in an alluvial forest using high-resolution satellite-based vegetation indices. – Remote Sensing of Environment 141: 52–63.

Viña, A., Gitelson, A.A., Nguy-Robertson, A.L. & Peng, Y., 2011: Comparison of different vegetation indices for the remote assessment of green leaf area index of crops. – Remote Sensing of Environment 115 (12): 3468–3478.

Zhao, J., Li, J., Liu, Q., Fan, W., Zhong, B., Wu, S. & Yin, G., 2015: Leaf area index retrieval combining HJ1/CCD and Landsat8/OLI data in the Heihe River Basin, China. – Remote Sensing 7 (6): 6862–6885.

Addresses of the authors:

Thorsten Dahms, Sylvia Seissiger & Christopher Conrad, Julius Maximilians University Würzburg, Institut für Geographie und Geologie, Lehrstuhl für Fernerkundung, Campus Hubland Nord 86, D-97074 Würzburg; e-mail: {thorsten.dahms}{sylvia.seissiger}{christopher.conrad}@uni-wuerzburg.de

Erik Borg, Vajen Hans-Hermann & Bernd Fichtelmann, German Remote Sensing Data Center, National Ground Segment, Kalkhorstweg 53, D-17235 Neustrelitz; e-mail: {erik.borg}{hans-hermann.vajen}{bernd.fichtelmann}@dlr.de

Manuskript eingereicht: September 2016
Angenommen: Oktober 2016