Philipps-Universität Marburg

Fachbereich 19 Geographie

Sommersemester 2017

Projektseminar
Fernerkundung und Maschinenlernverfahren zur Erfassung von Waldstrukturen im Universitätsforst Caldern

Leitung:
Hanna Meyer
Alice Ziegler
Insa Otte

Konzeption eines maschinellen Lernverfahrens zur Prognose des Leaf Area Index für den Universitätsforst Caldern

vorgelegt von:
Monkos, Martin; B. Sc. Geographie, 4. Fachsemester, Matrikelnummer 2743361;
Kontakt: Monkos@students.uni-marburg.de
Navel, Ruth; B. Sc. Geographie, 6. Fachsemester, Matrikelnummer 2550970;
Kontakt: Navel@students.uni-marburg.de
Schönberg, Andreas; B. Sc. Geographie, 8. Fachsemester, Matrikelnummer 2579634;
Kontakt: Schoenb4@students.uni-marburg.de

Abgabedatum: 15.10.2017

Abstract

For the widely used approach to analyse forest structures Leaf Area Index can be used.
The intention of this research is to create a proceeding for analysing Leaf Area Index in similar areas by using the University Forest Caldern, which is located in Hessen, Germany. This goal was achieved by utilising the machine learning algorithm Random Forest with the verification by a k-fold cross-validation.
The research showed that random forest is able to predict Leaf Area Index and can be reproduced for similar areas using similar predictors. It also showed that band 4 and band 8 of a Sentinel-2 hyperspectral image, as well as the Normalized Difference Vegetation Index, have the greatest influence on predicting Leaf Area Index in this conception.

Keywords: *Leaf Area Index (LAI), machine learning, Random Forest, k-fold cross-validation, University Forest Caldern, LI-COR LAI-2200*

1. Einleitung

Der Universitätsforst Caldern, welcher im Bundesland Hessen (Deutschland) liegt, wird seit April 2017 von der Philipps-Universität Marburg als Testareal für diverse praxisnahe Modelluntersuchungen der Geographischen Fakultät genutzt. Für dieses Gebiet gibt es bisher keine Untersuchungen zum *Leaf Area Index* (LAI), welcher ein Kennwert zur Charakterisierung der Bewuchsdichte eines Ökosystems ist und einer dimensionslosen Quantifizierung entspricht (Bréda, 2003). Er ist nach Watson (1947) definiert als Blattfläche pro Bodenoberfläche und kann beispielsweise zur Berechnung der photosynthetischen Fläche eines Ökosystems, der Evapotranspiration oder zur fernerkundlichen Bestimmung der Reflexionsoberfläche genutzt werden (Watson, 1947; Chen et al., 1997).
Durch *maschinelle Lernverfahren* (MLV) ist eine Transformation von punktuellen Messwerten in die Fläche kosteneffizient möglich. Ziel der Arbeit ist die Entwicklung eines Konzepts zur Untersuchung des LAI mittels MLV für den Universitätsforst Caldern. Zudem soll dieses Konzept auf andere Untersuchungsgebiete angewendet werden können.
Deshalb wurde für die Erfassung des LAI im Gesamtareal eine induktive Methode, unter Verwendung des *LI-COR LAI-2200 Plant Canopy Analyser* (LI-COR, 2012), gewählt. Der Zeitraum der Aufnahme lag zwischen dem 10. und 17. Mai 2017, nahe dem maximalen Vegetationsaustrieb. Anschließend wurde mit MLV die Transformation der LAI-Stichprobenwerte in die Fläche ermöglicht. Hierzu wurde das Klassifikationsverfahren des *Random Forest* (RF) genutzt und die Resultate mit einer *K-Fold Cross-Validation* (CV) verifiziert.

2. Methoden

Die methodische Durchführung folgt einem linearen Konzept (vgl. Abb. 1). Zu Beginn wurde ein geeignetes Erhebungsverfahren für das Untersuchungsgebiet gewählt und aus dem resultierenden Umfang die LAI-Messungen in situ durchgeführt.
Anschließend erfolgte eine Auswahl der Prädiktoren, welche eine Korrelation zwischen Erhebungsdaten und Untersuchungsumgebung vermuten ließen. Nachfolgend wurde auf Basis der Daten aus der Erhebung und den Prädiktoren ein geeignetes MLV gewählt. Abschließend wurden die Resultate validiert.

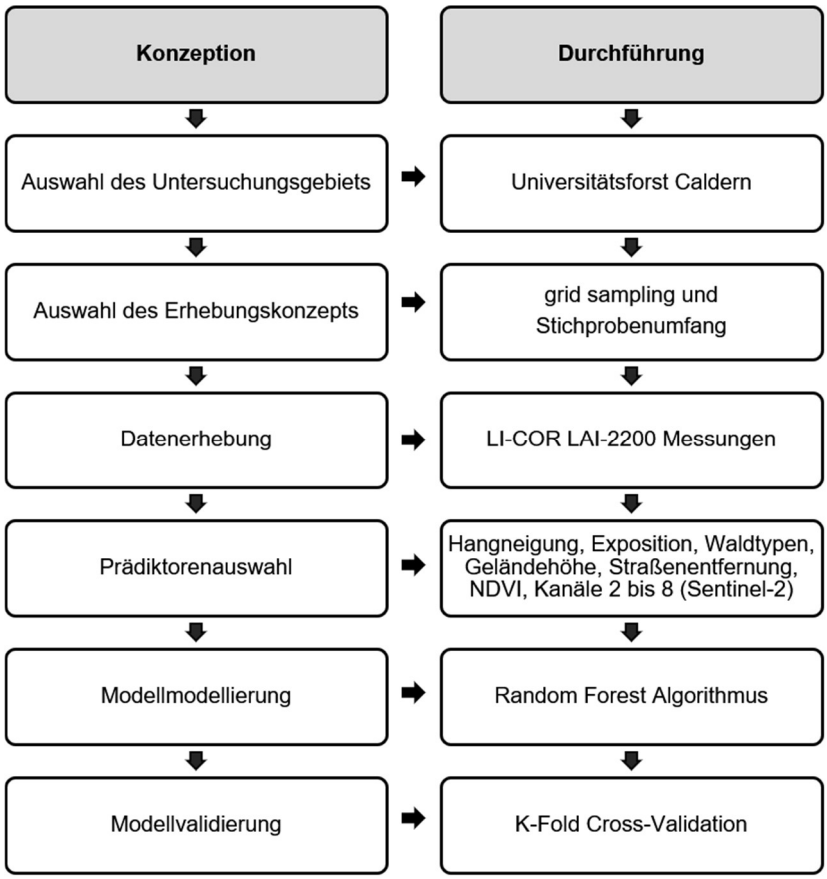

Abb. 1: Schematische Darstellung der Konzeption und Durchführung.

2.1. Untersuchungsgebiet

Der Universitätsforst Caldern (32N 477250 5632150) befindet sich im Landkreis Marburg-Biedenkopf. Das Untersuchungsgebiet umfasst eine Fläche von ca. 1,6 km². Die überwiegenden Baumarten sind Fagus und Quercus. Das Untersuchungsareal befindet sich in der Vegetationszone der sommergrünen Laubwälder, in den gemäßigte Breiten. Die Tagesmitteltemperatur im Untersuchungsgebiet beträgt 8,8°C und der Jahresniederschlag liegt zwischen 500 mm und 1100 mm (HLNUG, 2010).

2.2. Erhebungskonzept

Die Stichprobenpunkte zur Messung des LAI wurden über das Untersuchungsgebiet als *grid sampling* angelegt, welches bereits u.a. bei Law et al. (2001) und bei Sprintsin et al. (2009) im Zusammenhang mit LAI-Messungen Anwendung fand.

Aufgrund der Homogenität des Forsts, wurde auf eine große Stichprobe verzichtet. Sie wurde auf 70 Messpunkte festgelegt, welche in einem Abstand von 150 Metern zueinander liegen, wobei einige Punkte aufgrund von topographischen Unzugänglichkeiten versetzt wurden (vgl. Abb. 2).

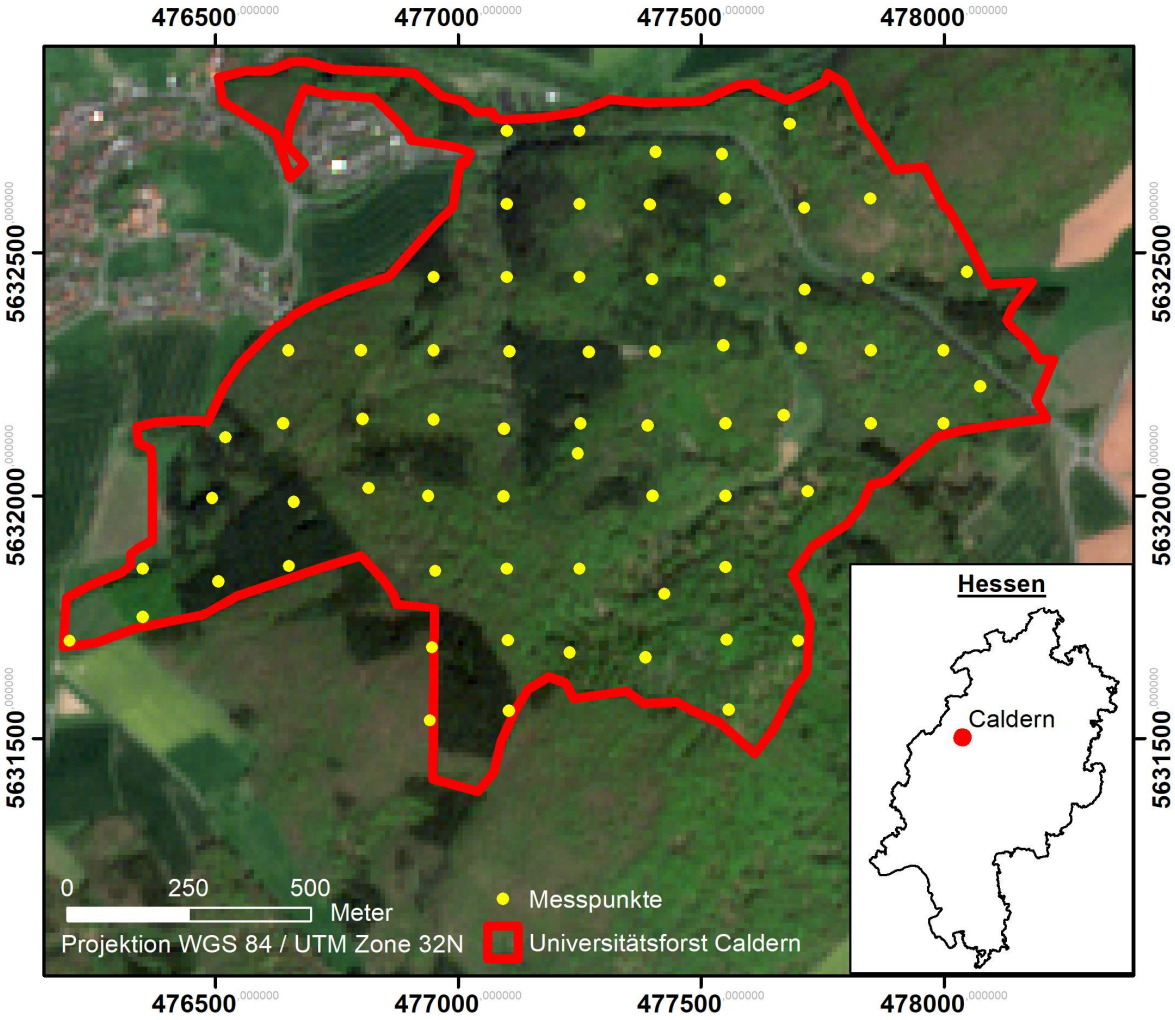

*Abb. 2: Untersuchungsgebiet und Messpunkte – Kartengrundlage: Sentinel-2 Echtfarbkomposit (*Copernicus Sentinel Data, 2017).

2.3. Datenerhebung

An jedem Standort wurden in einem zehn Meter Radius 20 Einzelwerte mit dem *LI-COR LAI-2200* gemessen. Aus den Messwerten wurde anschließend das arithmetische Mittel berechnet, um ein repräsentatives Ergebnis für den jeweiligen Standort zu generieren.
Der Zeitraum der Datenerhebung wurde so gewählt, dass eine bewölkungsfreie Sentinel-2 Multispektral-Aufnahme (Copernicus Sentinel Data, 2017) als Prädiktor berücksichtigt werden konnte. In diesem Zusammenhang wurde angenommen, dass es keine bedeutende Veränderung in der Wachstumsphase der Flora gab.

2.4. Prädiktorenauswahl

Die nachfolgenden Vorhersagevariablen dienen als statistische Merkmale zur Prognose der punktuell erhobenen LAI-Messdaten für die Gesamtfläche des Untersuchungsgebiets. Diese Prädiktoren wurden unter Berücksichtigung der räumlichen Variabilität des LAI und den Standortfaktoren, die das Untersuchungsareal beeinflussen, gewählt. Ihre Merkmalsausprägungen können kategorial, kontinuierlich oder metrisch sein (Kuhn & Johnson, 2016).

2.4.1. Hangneigung

Die Hangneigung als mittelbarer sekundärer Standortfaktor für Vegetation (Glaser et al., 2017) wurde gewählt aufgrund der Auswirkungen auf die individuelle Kronenform (Mitscherlich, 1978). Dieser Prädiktor wurde aus einem *Digitalen Geländemodell mit zehn Meter Auflösung* (DGM 10) (ATKIS, 2017) erstellt.

2.4.2. Exposition

Als mikroklimatischer Standortfaktor, welcher die Vegetationsstruktur und das Pflanzenwachstum beeinflusst, wurde die Exposition berücksichtigt. Untersuchungen in Makino (1997) und Monsi (2005) ergaben, dass die Strahlungsintensität Auswirkungen auf den Pflanzenbestand haben. Dieser abiotische Prädiktor wurde ebenfalls aus dem DGM 10 hergeleitet.

2.4.3. Waldtypen

Die Ergebnisse der Waldtypenkartierung des Universitätsforsts Caldern (Universität Marburg, 2017) wurden mit in das Modell eingepflegt, da sie das Areal auf Basis der Baumarten klassifiziert.

2.4.4. Geländehöhe

Der direkte Einfluss der Geländehöhe auf die lokale Vegetation wirkt sich auf die für das Untersuchungsareal relevanten Standortfaktoren, wie beispielsweise die Intensität und Dauer der Sonneneinstrahlung, die Windgeschwindigkeit und die Temperatur aus (Holden, 2017). Hierfür wurde zur Herleitung der Geländehöhe erneut das DGM 10 verwendet.

2.4.5. Straßenentfernung

Da es sich beim Untersuchungsgebiet um einen Forst handelt und es somit von Straßen und Wegen durchzogen ist, wurden diese auch als ein die Umgebung beeinflussender Faktor berücksichtigt. Hieraus wurde als Vorhersagevariable die Entfernung der LAI-Messpunkte zu Straßen und Wegen ermittelt.

2.4.6. NDVI

Der *Normalized Difference Vegetation Index* (NDVI) eignet sich zur spektralen Analyse von Pflanzenwachstum (Rouse et al., 1973) und wurde als Prädiktorvariable generiert. Die Datengrundlage hierfür war eine Sentinel-2 Multispektral-Aufnahme vom 10. Mai 2017. Der NDVI wurde mit folgender Formel berechnet:

$$NDVI = \frac{NIR - red}{NIR + red}$$

wobei NIR = Nahes Infrarot (Band 8) und red = Rot (Band 4) ist (vgl. Tab. 1).

Kanal	Bezeichnung	Wellenlänge (µm)
Band 2	Blau	0,490
Band 3	Grün	0,560
Band 4	Rot	0,665
Band 5	Vegetation rote Kante	0,705
Band 6	Vegetation rote Kante	0,740
Band 7	Vegetation rote Kante	0,783
Band 8	Nahes Infrarot	0,842

Tab. 1: Genutzte Kanäle der Sentinel-2 Multispektral-Aufnahme (European Space Agency, 2015).

2.4.7. Sentinel-2 Kanäle 2 bis 8

Für die Berücksichtigung von anderen Lichtspektren wurden sieben Kanäle der Sentinel-2 Multispektral-Aufnahme als Prädiktoren verwendet. Dies ermöglicht eine einseitige Vorhersagevariable hinsichtlich des Lichtspektrums durch den NDVI zu vermeiden.

2.5. Modellmodellierung

Die Prognose der punktuellen LAI-Messdaten für die Gesamtfläche des Untersuchungsgebiets wurde mit dem MLV des RF durchgeführt.

Da die Prädiktoren dieser Studie unterschiedliche Skalen besitzen und manche der Prädiktoren möglicherweise einen geringen Einfluss auf die Vorhersage des LAI für das Untersuchungsareal besitzen könnten, wurde aufgrund nachfolgender Kriterien der RF gewählt.

Der RF ist ein Algorithmus mit Baumstruktur und wurde ausgewählt, da er bessere Leistungsmerkmale als lineare Funktionen aufweist und in einer Studie von Appelhans et al. (2015), welche vergleichbare Prädiktoren verwendet, als ein sehr effizienter Algorithmus bewertet wurde (Appelhans et al., 2015).

Zudem ist der RF für die Verwendung von unterschiedlich skalierten Daten geeignet und robust gegenüber nicht aussagekräftigen Prädiktoren (Kuhn & Johnson, 2016). Weiterhin erwies sich der RF stabil gegenüber *overfitting* (Breiman, 2001). Eine ausführliche Beschreibung des RF ist in James et al. (2014) und Kuhn & Johnson (2016) auffindbar.

2.6. Modellvalidierung

Für die Verifizierung der Ergebnisse aus dem Modell, wurde eine CV durchgeführt. Ziel der CV ist es Aussagen über die Repräsentativität des erstellten Modells und der Ergebnisse treffen zu können.

Hierzu wird als Kennzahl zur Beurteilung des Modells der *Root Mean Square Error* (RMSE) berechnet. Dieser gibt an, wie gut die Funktionskurve an die vorliegenden Daten angepasst ist (Statistica, o.J.). Dabei gilt, je kleiner der RMSE, desto höher die Übereinstimmung.

Bei der CV wurden 10% der Ergebnisse als Trainingsdatensatz (Trainingsmenge) definiert und mit den übrigen 90% der Ergebnisse (Testmenge) auf Übereinstimmungen geprüft. Die CV folgt einem Zufallsmuster und somit ergeben sich bei jeder Iteration andere Ergebnisse für den RMSE und das *Bestimmtheitsmaß* R^2, welches angibt, wie stark die abhängige Variable, hier der LAI, von den Prädiktoren determiniert wird (Ernste, 2011).

Somit muss für ein aussagekräftiges Ergebnis das arithmetische Mittel über n-Iterationen gebildet werden; in diesem Fall wurden 20 Iterationen durchgeführt.

3. Ergebnisse

Bei der Prognose des LAI für das Untersuchungsgebiet, durch den RF, sind die Prädiktoren Band 4 und Band 8 der Sentinel-2 Aufnahme als Hauptdeterminanten identifiziert worden (vgl. Abb. 3). Der NDVI und die Hangneigung haben im Vergleich dazu einen geringeren Einfluss. Die Geländehöhe, sowie die Entfernung zu Straßen haben die geringste Bedeutung für die Prognose.

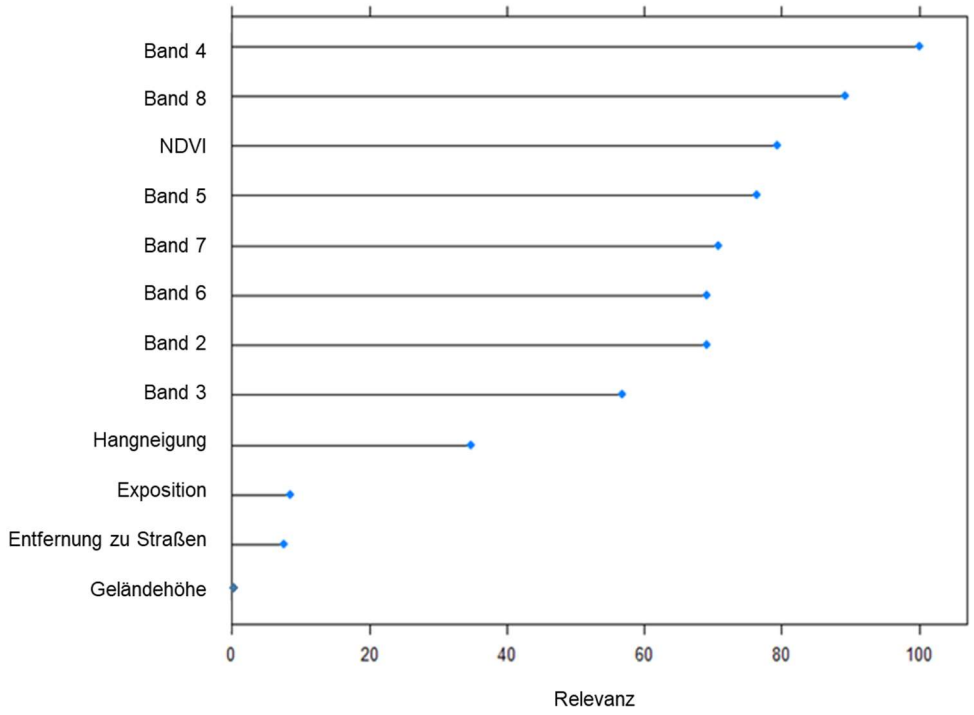

Abb. 3: Relative Variablenrelevanz in Prozent.

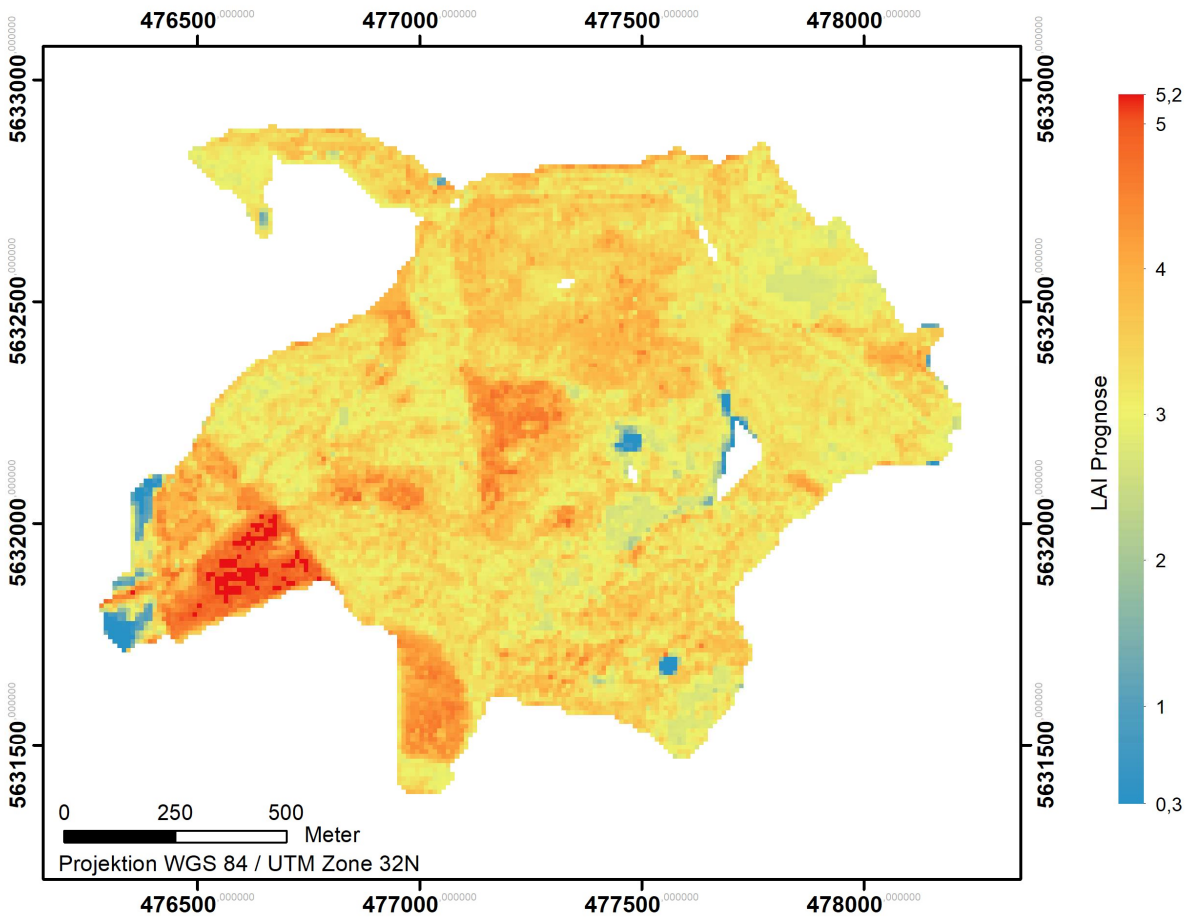

Abb. 4: Prognose des LAI für den Universitätsforst Caldern.

In Abbildung vier ist die Prognose des LAI auf das gesamte Untersuchungsgebiet visualisiert. 20 Iterationen des CV ergaben einen durchschnittlichen RMSE von 0,8841 und ein R^2 von 0,3952, mit dem häufigsten Durchlauf des RF über zwei Äste (mtry = 2) (Abb. 5).

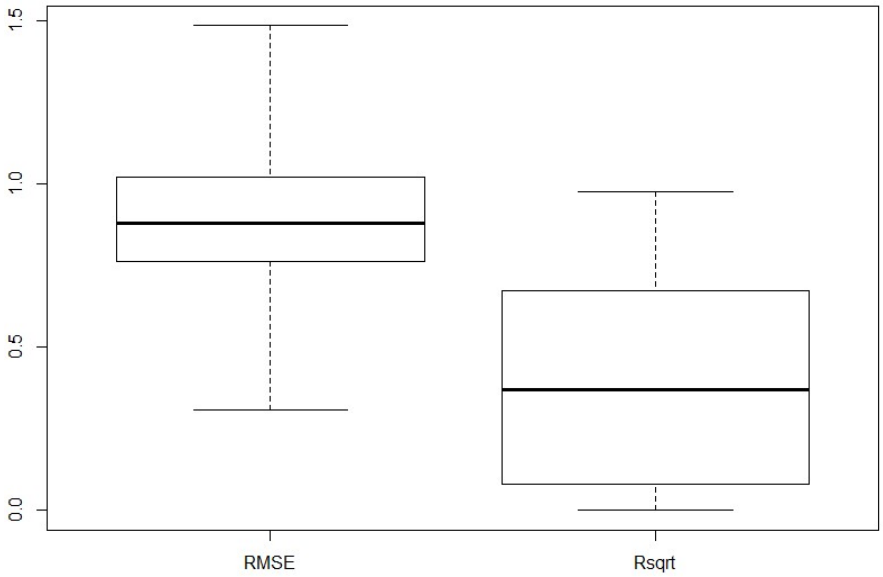

Abb. 5: Boxplot des RMSE und R^2.

4. Diskussion und Konklusion

Die Entwicklung eines Konzepts zur Bestimmung des LAI konnte mittels MLV unter Anwendung des RF positiv umgesetzt und mit der CV verifiziert werden. Die Prognosequalität des LAI für den Universitätsforst Caldern könnte jedoch noch verbessert werden. Dies ist mit einem RMSE von 0,8841 zu begründen, welcher nahe Eins liegt.

Dennoch wird die Prognose des LAI mittels der gewählten Prädiktoren, mit ca. 40% Einfluss auf das Modell, relativ gut erklärt. Dies wird durch ein R^2 von 0,3952 deutlich. Hierbei erwiesen sich die Prädiktoren Band 4, Band 8 und NDVI als Hauptdeterminanten des prognostizierten LAI. Um die Qualität des Ergebnisses zu erhöhen, könnte der Stichprobenumfang der LAI-Messungen im Untersuchungsgebiet ausgeweitet werden. Weiterhin könnten zusätzliche Standortvariablen als Prädiktoren in das Modell aufgenommen werden.

Die Übertragung des Konzepts auf andere Areale ist möglich. Ebenfalls können die Prädiktoren Hangneigung, Exposition und Geländehöhe verwendet werden, da es sich bei diesen um die Vegetation bestimmende, mittelbare Standortfaktoren handelt. Multispektral-Aufnahmen können als indirekte Indikatoren für Vegetation verwendet werden. Die Nutzung weiterer Prädiktoren muss jedoch spezifisch an das Areal angepasst werden.

Das hier durchgeführte Konzept ließe sich, durch eine Standardisierung mit fest definierten Prädiktoren und Messorten, reproduzieren und dadurch in eine Zeitreihenanalyse in zukünftigen Studien überführen.

Literaturverzeichnis

APPELHANS, T., MWANGOMO, E., HARY, D.R., HEMP, A. AND NAUSS, T. (2015), Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania, Spatial Statistics, 14: 91-113.

ATKIS (2017), Produktkatalog, Digitale Geobasisdaten, 4.

BREIMAN, L. (2001), Random Forests, Machine Learning, 45: 5-32.

BRÉDA, N.J.J. (2003), Ground-based measurements of leaf area index: a review of methods, instruments and current controversies, Journal of Experimental Botany, 54: 2403-2417.

CHEN, J.M. AND RICH, P.M (1997), Leaf area index of boreal forests: Theory, techniques, and measurements, Journal of Geophysical Research, 102: 29,429-29,443.

COPERNICUS SENTINEL DATA (2017)

ERNSTE, H. (2011), Angewandte Statistik in Geographie und Umweltwissenschaften, vdf Hochschulverlag AG, Zürich, Schweiz: 51, 73.

GLASER, R., HAUTER, CH., FAUST, D., GLAWION, R., SAURER, H., SCHULTE, A., UND SUDHAUS, D. (2017), Physische Geographie Kompakt, Springer, Berlin, Deutschland: 105.

HAWKINS, D. (2004), The Problem of Overfitting, Journal of Chemical Information and Computer Sciences, 44(1): 1-12.

HOLDEN, J. (2017), An Introduction to Physical Geography and the Environment, 4th edition, Pearson, Harlow, UK: 230, 715.

JAMES, G., WITTEN, D., HATTIE, T. AND TIBSHIRANI, R. (2014), An Introduction to Statistical learning: With Applications in R, 4th edition; Springer, New York, USA: 12, 303, 316, 320f., 328-332.

KUHN, M., AND JOHNSON, K. (2016), Applied Predictive Modeling, 5th edition, Springer, New York, USA: 8, 46, 384.

KUHN, M. (2017), Package ´caret`: Classification and Regression Training, R package version 6.0-77, CRAN, Wien, Austria.

LAW, B.E., TUYL, S.V., CESCATTI, A. AND BALDOCCHI, D.D. (2001), Estimation of leaf area index in open-canopy ponderosa pine forests at different successional stages and management regimes in Oregon, Agricultural and Forest Meteorology, 108: 1-14.

LI-COR (2012), LAI-2200 Plant Canopy Analyzer Instruction Manual, Li-Cor.

MAKINO, A., SATO, T., NAKANO, H. AND MAE, T. (1997), Leaf photosynthesis, plant growth and nitrogen allocation in rice under different irradiances, Planta, 203: 390-398.

MITSCHERLICH, G. (1978), Wald, Wachstum und Umwelt, 1. Band: Form und Wachstum von Baum und Bestand, 2. überarbeitete Auflage, J.D. Sauerländer's Verlag, Frankfurt a. M.: 10f..

MONSI, M. AND SAEKI, T. (2005), On the Factor Light in Plant Communities and its Importance for Matter Production, Annals of Botany, 95 (3): 549-567.

Philipps-Universität Marburg (2017), Waldtypenkartierung, unveröffentlichte Erhebung.

ROUSE, J. W., HAAS, R. H. , SCHELL J. A. AND DEERING, D. W. (1973), Monitoring vegetation systems in the Great Plains with ERTS, Third ERTS Symposium, NASA SP-351 I, 309-317.

SPRINTSIN, M., COHEN, S., MASEYK, K., ROTENBERG, E., GRÜNZWEIG, J., KARNIELI, A., BERLINER, P., AND YAKIR, D. (2011), Long term and seasonal courses of leaf area index in a semi-arid forest plantation, Agricultural and Forest Meteorology, 151: 565-574.

STATISTA GMBH (o.J.), Statistik-Lexikon: Definition Root Mean Square Error (RMSE; dt.: Wurzel der mittleren Fehlerquadratsumme), https://de.statista.com/statistik/lexikon/definition/303/root_mean_square_error/ (Zugriff: 05.10.2017)

WATSON, D.J. (1947), Comparative Physiological Studies on the Growth of Field Crops: I. Variation in Net Assimilation Rate and Leaf Area between Species and Varieties, and within and between Years, Annals of Botany, 11: 41-76.