



Flächennutzungsmonitoring IX Nachhaltigkeit der Siedlungs- und Verkehrsentwicklung?

IÖR Schriften Band 73 · 2017

ISBN: 978-3-944101-73-6

Workflows für Bilddaten und Big Data Analytics – Das Potenzial von Sentinel-2-Daten zur Baumartenklassifizierung

Melanie Brandmeier, Mathias Wessel

Brandmeier, M.; Wessel, M. (2017): Workflows für Bilddaten und Big Data Analytics – Das Potenzial von Sentinel-2-Daten zur Baumartenklassifizierung. In: Meinel, G.; Schumacher, U.; Schwarz, S.; Richter, B. (Hrsg.): Flächennutzungsmonitoring IX. Nachhaltigkeit der Siedlungs- und Verkehrsentwicklung? Berlin: Rhombos, IÖR Schriften 73, S. 135-142.

Workflows für Bilddaten und Big Data Analytics – Das Potenzial von Sentinel-2-Daten zur Baumartenklassifizierung

Melanie Brandmeier, Mathias Wessel

Zusammenfassung

In diesem Beitrag werden erste Ergebnisse eines Forschungsprojekts zur akkuraten Klassifikation von Baumarten anhand von Sentinel-2-Daten vorgestellt. Basierend auf der ArcGIS-Plattform wird über Rasterfunktionsketten ein optimaler Workflow erstellt, um einzelne Baumarten hierarchisch zu klassifizieren und die Übertragbarkeit auf ein anderes Gebiet zu testen. Dieser Workflow kann dann über den ArcGIS-Image-Server auf große Datenmengen angewendet werden.

Sowohl klassisch als auch multitemporal wurden verschiedene Bandkombinationen, Vegetationsindizes und Hauptkomponenten in pixel- vs. objektbasierten Klassifikationsansätzen verglichen. Als Klassifikatoren wurden zwei Machine-Learning-Algorithmen gewählt: ein Support Vector Machine Classifier (SVM) und ein Random Trees Classifier (RTC). Die besten Ergebnisse wurden über einen objektbasierten, hierarchischen Ansatz erzielt. Laub- und Nadelwälder wurden zunächst mit einer Gesamtgenauigkeit von 97,2 % über eine multitemporale, objektbasierte Klassifikation, basierend auf Kanal 8, erreicht. Ähnliche Genauigkeiten lieferte auch die SVM-Klassifikation, basierend auf den vier Hauptkomponenten der Augustszene mit vorangestellter Segmentierung der Maiszene. Eine anschließende Klassifikation der Laubbäume zwischen Buche, Eiche und restlichen Laubbäumen erzielte in einem multitemporalen Ansatz eine Gesamtgenauigkeit von 91 %. Es wurde an einer Szene im Mai, basierend auf den Kanälen 8,3 und 2, segmentiert und dann an den Hauptkomponenten 1 bis 12 der Augustszene über einen SVM-Algorithmus klassifiziert. Diese vorläufigen Ergebnisse sind vor dem Hintergrund der mittleren räumlichen Auflösung der Sentinel-2-Daten als sehr vielversprechend zu bewerten. Weitere Ziele des Projekts liegen in der Klassifikation einzelner Nadelbaumarten und in einer Übertragbarkeitsstudie auf ein zweites Gebiet.

1 Einführung

Bilddaten unterschiedlicher Sensoren kommen in immer mehr Forschungsbereichen zum Einsatz und stellen uns, gerade im Hinblick zunehmend größerer Datenmengen, vor neue Herausforderungen. Dies betrifft nicht nur die Analytik, sondern auch das Datenmanagement und die Bereitstellung von Image-Services.

Die Forschungsaktivität zum Thema Baumartenklassifikation hat in den vergangenen Jahren durch neue Technologien wie beispielsweise LIDAR-Systeme oder hyperspektrale Scanner signifikant zugenommen (Fassnacht et al. 2016). Auch hier ist das Thema „Big Data“ von zunehmender Bedeutung. Die Motivation liegt an der großen Anzahl möglicher Anwendungen. Diese sind neben Fragen der Waldinventur (van Aardt, Wynne 2007) auch Biodiversitätsabschätzungen und Monitoring (Shang, Chisholm 2014), Habitatmodellierungen (Jansson, Angelstam 1999) oder auch das übergeordnete Ziel des nachhaltigen Waldmanagements (EEA 2007). In der vorliegenden Studie evaluieren wir Sentinel-2-Multispektraldaten des Copernicus-Programms zur Baumartenklassifikation im Kontext großer Datenmengen.

Das Copernicus-Programm der ESA ist mit den Sentinel-1- bis -5-Satelliten eines der ambitioniertesten Erdbeobachtungsprogramme unserer Zeit. Die (zukünftige) Verfügbarkeit dieser Daten eröffnet sehr viele Möglichkeiten, stellt uns jedoch auch vor die oben genannten Herausforderungen. Der erste Sentinel-2-Satellit wurde im Juni 2015 gestartet, der zweite im März 2017. Zusammen liefern die beiden Satelliten in der optimalen Konstellation am Äquator alle fünf Tage eine Aufnahme und bieten neben der hohen zeitlichen Auflösung auch eine gute räumliche und spektrale Auflösung, welche gerade im „Red Edge“-Bereich sehr hoch ist (Abb. 1).

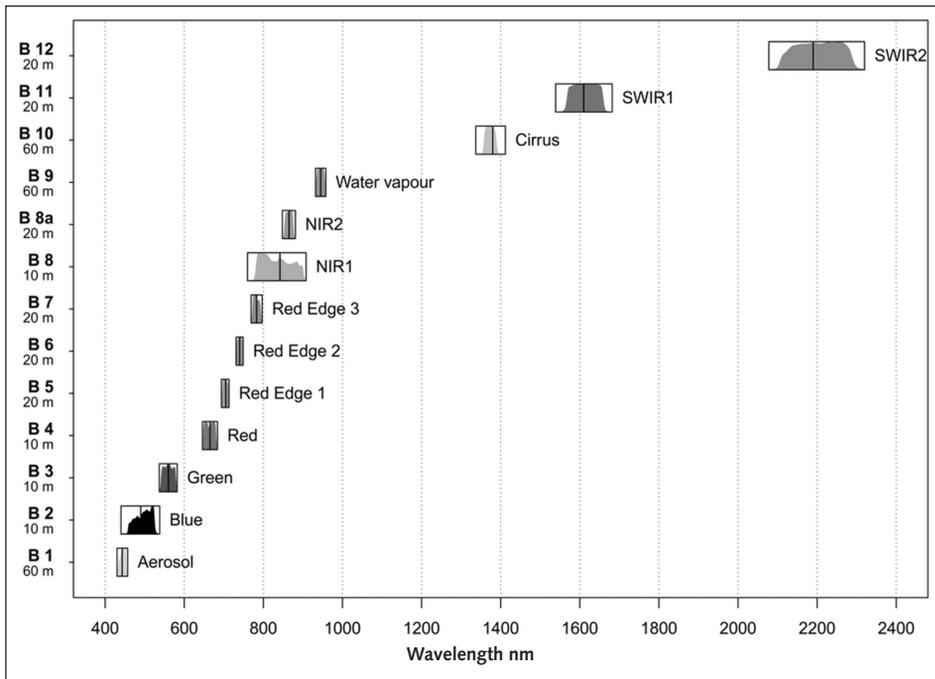


Abb. 1: Spektrale und räumliche Auflösung der Sentinel-2-Daten (Quelle: Immitzer et al. 2016)

Die Kombination der hohen zeitlichen Wiederholrate mit der spektralen Charakteristik ist für unsere Fragestellung sehr gut geeignet, da sich Vegetation gerade im „Red Edge“-Bereich gut klassifizieren lässt und ein multitemporaler Ansatz gewählt werden kann. Im Gegensatz zu Befliegungen oder zu anderen räumlich hochauflösenden Daten wie beispielsweise World-View2 oder GeoEye-1, bietet die freie Verfügbarkeit der Sentinel-Daten einen großen Kostenvorteil. Nachteilig ist die mittlere räumliche Auflösung, welche sich problematisch auf die Einzelbaumerkennung auswirken kann (Wulder et al. 2004). Eine genaue Evaluierung der Daten und das mögliche Potenzial der großflächigen Kartierung mittels Big Data Analytics ist Ziel unseres Projekts.

2 Methodik

Die Studie dient der Evaluierung der Sentinel-2-Daten für die Baumartenklassifikation im Sinne der Big-Data-Analyse. Methodisch wird anhand von zwei Testgebieten ein Workflow erstellt, der skalierbar ist und die dynamische Prozessierung großer Datenmengen erlauben soll. Technologisch wird dies über die ArcGIS-Plattform umgesetzt. Die Speicherung der Daten erfolgt innerhalb einer Geodatenbank in einem Mosaik-Dataset. Dabei handelt es sich um ein Datenmodell, mit dem Rasterdaten verwaltet werden. Es verfügt über erweiterte Funktionen der Raster-Abfrage und Verarbeitung und kann als Quelle für die Bereitstellung von Image-Services genutzt werden. Die Analytik baut auf Raster-Funktionsketten auf. Diese sind im Gegensatz zu normalen Geoverarbeitungswerkzeugen sehr effizient, da Berechnungen „on the fly“ durchgeführt werden. Die Verarbeitung großer Datenmengen mittels unseres optimierten Workflows kann dann zukünftig über den Image Server erfolgen.

2.1 Datengrundlage und Untersuchungsgebiete

Es wurden zwei Untersuchungsgebiete in Bayern ausgewählt für die wir Waldinventurdaten der Bayerische Staatsforsten (BaySF) und Orthofotos der Bayerischen Vermessungsverwaltung zur Verfügung hatten: Freisinger Forst und Ebersberger Forst. Die Waldinventurdaten werden als Trainingspunkte und zur Validierung der Klassifikation verwendet, die Orthofotos dienen der zusätzlichen Validierung. Der Workflow wird anhand des Ebersberger Forsts erstellt und am Freisinger Forst getestet. Als Klassifikationsgrundlage wurden drei Sentinel-2-Szenen (22. Mai, 9. August und 29. September 2016) nach dem Kriterium 0 % Wolkenbedeckung ausgewählt. Damit werden drei Jahreszeiten und insbesondere die Hauptblütephase von Laubbäumen im Mai abgedeckt. Dabei handelt es sich um frei verfügbare Level-1C-Daten.

2.2 Pre-Processing und Klassifikation

Da die Sentinel-2-Daten als L-1C-Produkt vorliegen, wurde zunächst eine Atmosphärenkorrektur mittels des Sen2Cor-Prozessors durchgeführt um L-2A-Daten (Bottom-Of-Atmosphäre-Reflektanz) zu erhalten. Die Korrektur basiert auf Look-Up-Tabellen, welche über ein Strahlungstransfermodell (libRadtran1) erstellt wurden (Müller-Wilm 2016).

Um einen optimalen Klassifikationsworkflow zu erstellen wurden zwei Machine-Learning-Algorithmen an verschiedenen Bandkombinationen, Hauptkomponenten und Vegetationsindizes getestet. Die verwendeten Rasterfunktionen in ArcGIS-Pro waren ein Support Vector Machine (SVM) Klassifikationsalgorithmus mit einer Radial-Basis-Kernel-Funktion sowie ein Random-Trees-Klassifikationsalgorithmus (RTC). Beide Algorithmen wurden pixelbasiert und objektbasiert getestet. Bei dem objektbasierten Ansatz ist der Klassifikation ein Mean-Shift-Segmentationsalgorithmus vorangestellt. Es wurde ein hierarchischer Ansatz gewählt: Zunächst wurde Laub- und Nadelwald optimal getrennt, dann die einzelnen Klassen weiter nach Baumarten unterteilt. In Abbildung 2 sind die Spektralprofile der einzelnen Baumarten für den Monat Mai dargestellt.

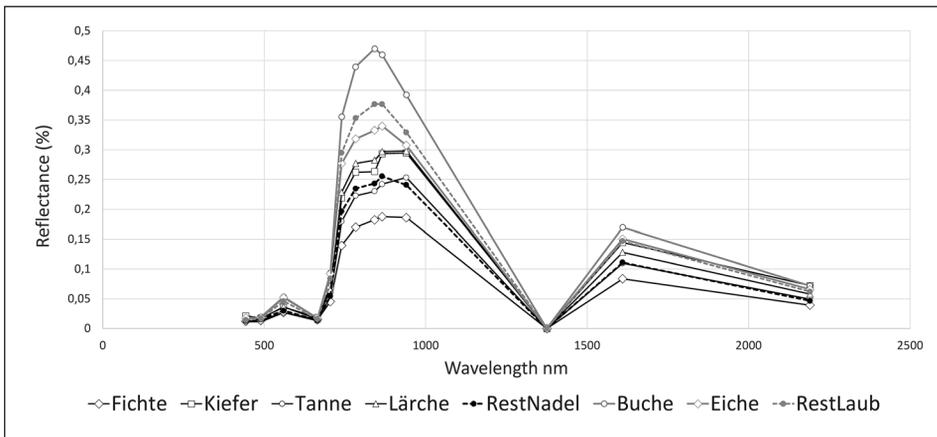


Abb. 2: Spektralprofile der einzelnen Baumarten (22. Mai) (Quelle: eigene Bearbeitung)

Man sieht deutliche Unterschiede zwischen Laub- und Nadelwald im Bereich von ca. 750 nm bis 1 000 nm. Die einzelnen Laub- und Nadelbäume weisen ebenfalls Unterschiede auf, jedoch deutlich weniger ausgeprägt.

Da die spektralen Unterschiede zwischen den einzelnen Laub- und Nadelbaumarten deutlich weniger ausgeprägt sind als zwischen Nadel- und Laubbaumarten, wurde zur Unterscheidung der einzelnen Arten auch ein multitemporaler Ansatz mit verschiedenen Bandkombinationen getestet. Dies ermöglicht Unterschiede im Jahresgang, wie beispielsweise in der Blütezeit oder im Herbst, zu nutzen (Abb. 3). Zusätzlich wurden die

Daten mittels einer Hauptkomponentenanalyse transformiert und ebenfalls für die Klassifikation getestet. Als Trainingsdaten wurden die Waldinventurdaten verwendet, für die anschließende Validierung der einzelnen Klassifikationen wurden jeweils unabhängige Testdaten aus den Waldinventurdaten verwendet.

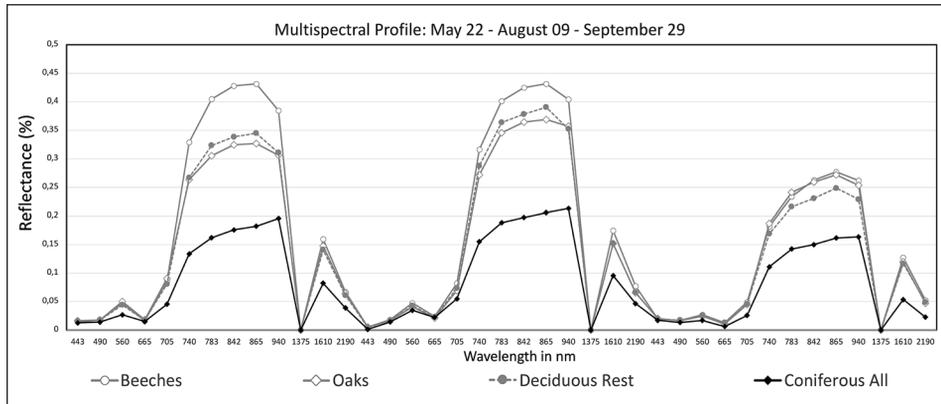


Abb. 3: Multitemporale Profile der Laubbaumarten und des Nadelwaldes (gesamt) (Quelle: eigene Bearbeitung)

Unterschiede im Jahresgang können für eine Verbesserung der Klassifikation der einzelnen Laubbaumarten genutzt werden.

3 Erste Ergebnisse

Insgesamt wurden 16 verschiedene Klassifikationen zur Trennung zwischen Laub- und Nadelwald getestet. Die Ergebnisse sind in diesem Schritt alle sehr gut und schwanken zwischen 81 % und 97 % Gesamtgenauigkeit. Das beste Ergebnis wurde über einen multitemporalen, objektbasierten Ansatz, basierend auf Kanal 8 der Szenen Mai, August und September und anschließender SVM-Klassifizierung, erzielt. Basierend auf diesem Ergebnis wurden 44 verschiedene Klassifikationen (Tabelle 1) zur Unterscheidung zwischen Eiche, Buche und restlichen Laubbäumen getestet. Die Genauigkeit schwankte zwischen 39 % und 91 %. Generell erzielten die objektbasierten Ansätze bessere Ergebnisse als die entsprechenden pixelbasierten Klassifikationen. Das beste Ergebnis wurde wieder in einem objektbasierten Ansatz erzielt. Einer vorangestellten Mean-Shift-Segmentierung, basierend auf den Kanälen 8, 3 und 2 der Maiszene (minimale Segmentgröße 5), wurde mit einer SVM-Klassifikation mit Radial-Basis-Funktion als Kernel, basierend auf den Hauptkomponenten 1 bis 12 der Augustszene, eine Genauigkeit von 95 % erzielt (Abb. 4). Man erkennt deutliche Aufforstungsmuster und eine Dominanz von Eichen und Buchen entlang der Waldwege. Der Nadelwald ist sehr stark von Fichten dominiert. Ein weiteres Ziel unserer Studie ist es, diese Klasse weiter zu unterteilen. Der

momentan limitierende Faktor ist ein Fehlen ausreichender Trainingsgebiete für Lärchen und Tannen, der durch die Omnipräsenz der Fichten bedingt ist.

Tab. 1: Vergleich der Ergebnisse der Klassifikationen der Laubbäume (Buche, Eiche, Rest); bestes Ergebnis mit 91 % Gesamtgenauigkeit hervorgehoben (Quelle: eigene Bearbeitung)

Accuracy	Method*	Classifier**	Input Image***	Segmentation Image	Segmentation criteria****
54,3	OBIA	SVM	May 2-3-4-5-6-7-8-9 + PCA 1-2	May 8-3-2	MSS: 5px, ACC, MDM
49,4	OBIA	SVM	May 2-3-4-5-6-7-8-9 + PCA 1-2	May 8-3-2	MSS: 5px, ACC, MDM, SD, COP, C, R
71,7	OBIA	SVM	May 2-3-4-5-6-7-8-9 + PCA 1-4	May 8-3-2	MSS: 5px, ACC, MDM
67,3	OBIA	SVM	May 2-3-4-5-6-7-8-9 + PCA 1-4	May 8-3-3	MSS: 5px, ACC, MDM, SD, COP, C, R
59	OBIA	SVM	May 2-3-4-5-6-7-8-9 + PCA 1-4	May 8-3-4	MSS: 5px, ACC, MDM, C
61,6	OBIA	RTC	May 2-3-4-5-6-7-8-9 + PCA 1-4	May 8-3-5	MSS: 5px, ACC, MDM
63,3	PB	SVM	May 2-3-4-5-6-7-8-9 + PCA 1-4		
76,2	PB	RTC	May 2-3-4-5-6-7-8-9 + PCA 1-4		
80,7	OBIA	SVM	May 8-3-2 + PCA 1-4 + NDVI	May 8-3-2	MSS: 5px, ACC, MDM
76	OBIA	SVM	May NDVI + PCA 1-4	May 8-3-3	MSS: 5px, ACC, MDM
65	OBIA	SVM	May 8-3-2 + NDVI	May 8-3-4	MSS: 5px, ACC, MDM
75,7	OBIA	SVM	May 9-8-7-3-2	May 8-3-5	MSS: 5px, ACC, MDM
74,3	OBIA	SVM	May 9-8-7-3-2 + PCA 1-4 + NDVI	May 8-3-6	MSS: 5px, ACC, MDM
55	PB	SVM	May 9-8-7-3-2 + PCA 1-4 + NDVI		
75,1	PB	RTC	May 9-8-7-3-2 + PCA 1-4 + NDVI		
87,2	OBIA	SVM	May 7-6-5	May 8-3-2	MSS: 5px, ACC, MDM
59,9	OBIA	SVM	May 7-6-5 + PCA 1-4	May 8-3-3	MSS: 5px, ACC, MDM
78,8	OBIA	SVM	May 7-6-5	May 8-3-4	MSS: 5px, ACC, MDM, SD, COP
87,2	OBIA	SVM	May 7-6-5	May 8-3-5	MSS: 5px, ACC, MDM
53,7	PB	SVM	May 7-6-5		
67,1	OBIA	RTC	May 7-6-5	May 8-3-2	MSS: 5px, ACC, MDM
87,2	OBIA	SVM	May 7-6-5 + NDVI	May 8-3-3	MSS: 5px, ACC, MDM
87,2	OBIA	SVM	May 7-6-5-4	May 8-3-4	MSS: 5px, ACC, MDM
62,5	OBIA	RTC	May 7-6-5-4	May 8-3-5	MSS: 5px, ACC, MDM
80,3	OBIA	SVM	May 7 Aug 6 May 5	May 8-3-6	MSS: 5px, ACC, MDM
61,8	OBIA	SVM	May - all bands	May 8-3-7	MSS: 5px, ACC, MDM
89,1	OBIA	SVM	August - all bands	May 8-3-8	MSS: 5px, ACC, MDM
63,1	OBIA	RTC	August - all bands	May 8-3-9	MSS: 5px, ACC, MDM
46	PB	SVM	August - all bands		
65,2	OBIA	SVM	August - all bands + PCA + May 1-4	May 8-3-2	MSS: 5px, ACC, MDM
80,5	OBIA	SVM	August 7-6-5	May 8-3-3	MSS: 5px, ACC, MDM
80,7	OBIA	SVM	August 8-3-2	May 8-3-4	MSS: 5px, ACC, MDM
60,9	OBIA	SVM	September - all bands	May 8-3-5	MSS: 5px, ACC, MDM
63,3	OBIA	SVM	May PCA 1-12	May 8-3-6	MSS: 5px, ACC, MDM
60	OBIA	SVM	May all bands + May PCA 1-4	May 8-3-7	MSS: 5px, ACC, MDM
90,9	OBIA	SVM	August PCA 1-12	May 8-3-8	MSS: 5px, ACC, MDM
86,3	OBIA	SVM	August PCA 1-4	May 8-3-9	MSS: 5px, ACC, MDM
60,4	OBIA	RTC	August PCA 1-4	May 8-3-10	MSS: 5px, ACC, MDM
38,7	PB	SVM	August PCA 1-4		
74,6	OBIA	SVM	May PCA 1-4 + August PCA 1-4	May 8-3-2	MSS: 5px, ACC, MDM
77,2	OBIA	SVM	August PCA 1-4 + September PCA 1-4	May 8-3-3	MSS: 5px, ACC, MDM
66	OBIA	SVM	May PCA 1-4 + August PCA 1-4 + September PCA 1-4	May 8-3-4	MSS: 5px, ACC, MDM
67,7	OBIA	SVM	September PCA 1-12	May 8-3-5	MSS: 5px, ACC, MDM
80,8	OBIA	SVM	August NDVI	May 8-3-6	MSS: 5px, ACC, MDM
86,9	OBIA	SVM	August PCA 1-4 + NDVI	May 8-3-7	MSS: 5px, ACC, MDM

* PB: Pixel based, OBIA: Object based Image Analysis

** SVM: Support Vector Machine, RTC: Random Trees Classifier

*** NDVI: Normalized Difference Vegetation Index, PCA: Principal Component Analysis

**** MSS: minimum segment size, ACC: Active chromaticity color, MDN: Mean digital number, SD: Standard deviation, COP: Cout of pixel, C: Compactness, R: Rectangularity

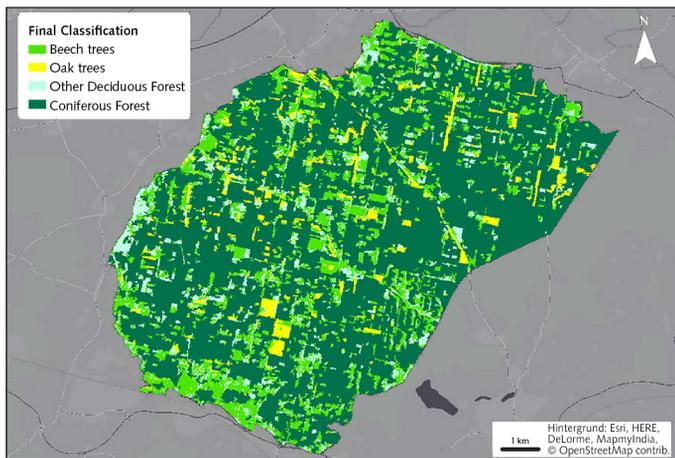


Abb. 4: Ergebnisse der Klassifikation von Buchen, Eichen, restlichem Laubwald und Nadelwald (Quelle: eigene Bearbeitung)

4 Diskussion und Ausblick

Unsere vorläufigen Ergebnisse zeigen sehr gut das Potenzial der Sentinel-2-Daten zur Klassifikation der Hauptbaumarten. Momentan erzielen wir bei der Trennung von Laub- und Nadelwald eine Gesamtgenauigkeit von 97 %, bei der anschließenden Unterteilung der Laubwaldklasse in Buche, Eiche und restliche Laubbäume 91 %. Im Vergleich zwischen objektbasierten und pixelbasierten Ansätzen erzielten objektbasierte Ansätze für beide Machine-Learning-Algorithmen deutlich bessere Ergebnisse. Die Mean-Shift-Segmentierung lieferte beste Ergebnisse an der Mai-Szene, basierend auf den Kanälen 8-3-2. Dies liegt an den signifikanten spektralen Unterschieden in diesem Wellenlängenbereich im Frühjahr (Abb. 2). Als Parameter der Segmentierung erwiesen sich eine minimale Segmentgröße von 5 Pixeln und „Active Chromaticity Color“ sowie „Mean Digital Number“ als optimal. Die Erweiterung um die Parameter „Standard Deviation“, „Count of Pixel“, „Compactness“ und „Rectangularity“ führten eher zu einer Verschlechterung der Ergebnisse. Die Verschlechterung durch Formparameter könnte an der nur mittleren räumlichen Auflösung der Sentinel-Daten liegen, allerdings auch an der Struktur des Ebersberger Forstes. Die besten Ergebnisse zur Trennung der Laubbäume lieferten Klassifikationen mit multitemporalem Ansatz. Dies wurde über eine Segmentierung an der Maiszene und Klassifikation der Augustszene erreicht. Hier schwanken die Gesamtgenauigkeiten zwischen ca. 39 % und 91 % für den SVM-Klassifikator.

Die nächsten Schritte des Projekts sind die Klassifikation der einzelnen Nadelbäume sowie eine Übertragbarkeitsstudie auf den Freisinger Forst. Ein optimaler Workflow, basierend auf Rasterfunktionen, sollte auf große Datenmengen angewendet werden können.

Für dieses Ziel ist eine Übertragbarkeitsstudie unabdingbar um nicht Gefahr zu laufen, die Klassifikation zu spezifisch für eine Szene mit sehr hoher Gesamtgenauigkeit, aber geringer Generalisierbarkeit anzulegen.

5 Literatur

- EEA – European Environmental Agency (2007): European Forest Types: Categories and Types for Sustainable Forest Management Reporting and Policy. In: EEA Technical Report 9/2006. Copenhagen.
https://www.eea.europa.eu/publications/technical_report_2006_9
(Zugriff: 01.06.2017).
- Fasnacht, F. E.; Latifi, H.; Stereńczak, K.; Modzelewska, A.; Lefsky, M.; Waser, L. T.; Straub, C.; Ghosh, A. (2016): Review of studies on tree species classification from remotely sensed data. In: Remote Sensing of Environment 18: 64-87.
- Immitzer, M.; Vuolo, F.; Atzberger, C. (2016): First experience with Sentinel-2 data for crop and tree species classifications in central Europe. In: Remote Sensing 8.3: 166.
- Jansson, G.; Angelstam, P. (1999): Threshold levels of habitat composition for the presence of the long-tailed tit (*Aegithalos caudatus*) in a boreal landscape. In: Landscape ecology 14 (3): 283-290.
- Müller-Wilm, U. (2016): Sentinel-2 MSI – Level-2A Prototype Processor Installation and User Manual. Telespazio VEGA Deutschland GmbH: Darmstadt.
- Shang, X.; Chisholm, L. A. (2014): Classification of Australian native forest species using hyperspectral remote sensing and machine-learning classification algorithms. In: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7 (6): 2481-2489.
- Van Aardt, J. A. N.; Wynne, R. H. (2007). Examining pine spectral separability using hyperspectral data from an airborne sensor: An extension of field-based results. In: International Journal of Remote Sensing 28 (2): 431-436.
- Wulder, M. A.; Hall, J. R.; Coops, N. C.; Franklin, S. E. (2004): High spatial resolution remotely sensed data for ecosystem characterization. In: BioScience 54.6: 511-521.