PFG 2014 / 5, 0393–0407
Stuttgart, October 2014

Article

Improving Land Cover Maps in Areas of Disagreement of Existing Products using NDVI Time Series of MODIS – Example for Europe

Francesco Vuolo & Clement Atzberger, Vienna, Austria

Keywords: classification, land cover, random forest, accuracy / confidence, time series, NDVI

Summary: Regional to global scale land cover (LC) information is one of the most important inputs to various models related to global climate change studies, natural resource use and environmental assessment. This paper presents a methodology to derive land cover maps using time series of moderate-resolution imaging spectroradiometer (MODIS) 250 m normalized difference vegetation index (NDVI). An example for Europe is produced using the random forest (RF) classifier. For the accuracy assessment, the overall performance of our classification product (BOKU, Universität für Bodenkultur) is compared to the one of three existing LC maps namely GlobCover 2009, MODIS land cover 2009 (using the IGBP classification scheme) and GLC2000. Considered GlobCover and IGBP, the assessment is further detailed for areas where these two maps agree or disagree. The BOKU map reported an overall accuracy of 71%. Classification accuracies ranged from 78% where IGBP and GlobCover agreed to 63% for areas of disagreement. Results confirm that existing LC products are as accurate as the BOKU map in areas of agreement (with little margin for improvements), while classification accuracy is substantially better for the BOKU map in areas of disagreement. Two pixel-based measures of confidence of classification were derived, which showed a strong correlation with classification accuracy. The study also confirmed that RF provides an unbiased estimation of the error (out-of-bag) and therefore eliminates the need for an independent validation dataset.

Zusammenfassung: *Verbesserung von Landbedeckungskarten in Gebieten widersprüchlicher Grundlagen mit Hilfe der NDVI-Zeitreihe von MODIS – Beispiel für Europa.* Verlässliche regionale bis globale Informationen über die aktuelle Landbedeckung sind von größter Bedeutung für Fragen des Klimaschutzes, des Managements natürlicher Ressourcen sowie für Umweltbewertungen. Der vorliegende Beitrag beschreibt ein innovatives Verfahren, um Landbedeckungskarten aus Zeitreihendaten des Umweltsatelliten MODIS in 250 m Bodenauflösung zu generieren. Das überwachte Klassifikationsverfahren basiert auf multiplen Entscheidungsbäumen (Random Forest Classifier) und verwendet Informationen über den temporalen Verlauf der fernerkundlich erfassten Vegetationsdichte (NDVI). Um die Qualität unserer europäischen Landbedeckungskarte zu evaluieren, wird ein Vergleich mit drei existierenden globalen Landbedeckungskarten durchgeführt: GlobCover 2009, MODIS Land Cover 2009 (IGBP) und GLC2000. Als Referenz dient ein aus Google Earth generierter Referenzdatensatz basierend auf der visuellen Interpretation einer großen Anzahl von Referenzpunkten. Für die vergleichende Evaluierung wird zwischen Gebieten mit und ohne Übereinstimmung zwischen IGBP und GlobCover Produkten unterschieden. Die Landbedeckungskarte der Universität für Bodenkultur (BOKU) erreicht eine Gesamtgenauigkeit von 71%. Die Klassifikationsgenauigkeit variert dabei zwischen 78%, wenn IGBP und GlobCover übereinstimmen, und 63% in Gebieten ohne Übereinstimmung der zwei Vergleichskarten. Die BOKU Landbedeckungskarte zeigt in Gebieten ohne Übereinstimmung zwischen IGBP und GlobCover eine deutliche Verbesserung der Klassifikationsgenauigkeit. Die pixelweise generierten Konfidenzmaße zeigen darüber hinaus eine deutliche Korrelation mit der erzielten Klassifikationsgenauigkeit. Damit erhält der Nutzer ein detailliertes Bild über die zu erwartende Unsicherheitsmarge. Die Studie bestätigt, dass Random Forest eine ausgewogene (unbiased) Einschätzung der Fehler (out-of-bag) bietet.

www.schweizerbart.de
DOI: 10.1127/1432-8364/2014/0232
1432-8364/14/0232 $ 3.75

1 Introduction

Reliable and regularly updated land cover (LC) maps at medium spatial resolution and with regional-to-global coverage are required by land administrators, natural resource managers and scientists working in climate and environmental modelling (Hibbard et al. 2010). During the past 15 years, various international initiatives have been focused on the operational production of land cover products using satellite-based observations at a range of temporal and spatial resolutions. For instance, the European Commission's Joint Research Centre coordinated the realization of the global land cover 2000 (GLC2000) map using SPOT vegetation data at 1 km for the year 2000 (Bartholomé & Belward 2005). At higher spatial resolution, the European Space Agency (ESA) promoted the GlobCover project (Arino et al. 2008) to produce a global land cover map for the year 2005 and 2009 using MERIS data at 300 m pixel size. The land cover team at the Boston University recently released the Collection 5 of the MODIS global land cover type product (MCD12Q1) (Friedl et al. 2002, 2010). The product has a spatial resolution of 500 m (1 km in collection 4) and includes LC products obtained with five different classification systems, among them the International geosphere-biosphere programme (IGBP) and the University of Maryland (UMD) classifications. Of particular interest for the meteorological community is the ECOCLIMAP by Météo France (Masson et al. 2003), with a recent release of an improved version for Europe at 1 km resolution (ECOCLIMAP-II/Europe) (Faroux et al. 2013). ECOCLIMAP-II/Europe builds on existing land cover maps such as GLC2000 and CORINE land cover 2000 (Bossard et al. 2000) and on the analysis of leaf area index data from MODIS and of normalized difference vegetation index (NDVI) data from SPOT vegetation covering the period 1999 – 2005. Within the ESA's climate change initiative (CCI), a new set of global LC maps are also being generated using multi-sensor and multi-year observations (Landcover 2014).

Many studies have focused on dataset harmonization (Herold et al. 2006, 2009), intercomparison (Herold et al. 2008, Pflugmacher et al. 2011, Vintrou et al. 2012) and synergy to improve the consistency of existing products (See & Fritz 2006, Jung et al. 2006, Pérez-Hoyos et al. 2012, Vancutsem et al. 2013) or to extend their domain of application (Faroux et al. 2013). Despite the growing number of datasets, improved methods and joint efforts, land cover classification remains a challenging task due to the intrinsic high variability of class signatures at regional and global scales, together with the mixed pixel issues and class definition limitations. Most studies seem to agree that an improvement is possible by combining multiple input features (spectral- and temporal-observations), stratifications, e.g. based on climatological products, different classification approaches and by exploiting synergies between existing land cover products.

Interestingly, See & Fritz (2006) described a methodology to identify hotspots of disagreement between existing land cover products and to geographically focus reclassification efforts. Liu et al. (2004) noted that high agreement between two different classifiers can lead to a higher accuracy. This trend was confirmed by Vuolo & Atzberger (2012) where it was experimentally demonstrated that areas of thematic agreement between two existing land cover maps are more reliable in terms of classification accuracy compared to areas of disagreement. Moreover, the work of Vuolo & Atzberger (2012) showed that the classification performance of areas of disagreement could be notably improved using time series of NDVI from MODIS.

The scope of this paper is to extend the concept explored in Vuolo & Atzberger (2012) focusing on a larger dataset and including an additional land cover product for comparison. The paper is structured in three main components. First, we present the overall performance of our classification product (BOKU) and of three existing LC maps namely GlobCover 2009, MODIS land cover 2009 (using the IGBP classification scheme) and GLC2000. For this first comparison, validation focused on an independent dataset (visual interpretation of a large number of sample plots in Google Earth) not used during the training phase. BOKU map was derived using a smoothed and gap-filled time series of MODIS 250 m NDVI (average of multi-year

data centred on the year 2009) and the random forest (RF) classifier. Classification was stratified considering major European biogeographical regions. Then, we assessed the accuracy for the areas of agreement and of disagreement between the existing LC maps. For this detailed analysis we considered GlobCover and IGBP because they are produced with a similar spatial resolution and refer to the same year (2009) as the BOKU map. Finally, we produced a pan-European LC map and analysed the spatial pattern and the confidence of classification in relation to existing products. For this task, we exploited the entire reference dataset and based the accuracy assessment on the bootstrapped error estimates obtained during the training process.

2 Materials and Methods

A methodology is described for producing reliable land cover maps focusing on broad (here seven) LC classes. Only a few broad LC classes were chosen i) to provide a practical separation between managed vegetation and natural vegetation, and ii) to keep some flexibility and not preclude the possibility of comparisons with other LC schemes. The LC definitions used in this study and the corresponding MODIS IGBP (2009), GlobCover 2009 and GLC2000 class codes are provided in Tab. 1.

Multi-temporal MODIS NDVI observations were used to characterize the variations in growth patterns and phenology of different vegetation covers (Rodriguez-Galiano et al. 2012b). The image classification was performed by the machine learning random forest (RF) algorithm, which is an ensemble of decision trees.

2.1 *Satellite Data and Pre-Processing*

The input data used in this study consisted of 16-day NDVI composites from MODIS with a 250 m pixel size. The MODIS NDVI composite is a Level 3 product, calculated from the Level 2 daily surface reflectance product (MOD09 and MYD09 series) (Vermote et al. 2002). Data were aggregated using the constrained view angle – maximum value composite (CV-MVC) compositing method in a 16-day interval (Huete et al. 2002). The MODIS NDVI is currently available from Terra and Aqua platforms (MOD13Q1 and MYD13Q1) and the combined use of the two satellites provides time series with 8-day frequency.

In this study, we selected the year 2009 as the pivotal year and considered a MODIS time series of five years from the start of 2007 to the end of 2011. The CV-MVC time series were smoothed and interpolated to daily time steps based on a state-of-the-art Whittaker smoother (Atzberger & Eilers 2011a, 2011b, Atkinson et al. 2012). The smoothing/gap filling takes into account MODIS quality flags and the exact composite date of each pixel. For

Tab. 1: Land cover (LC) class codes and descriptions after aggregation of GlobCover 2009, GLC2000 and MODIS IGBP 2009 products. Water was not classified but taken from a water mask made for the MODIS satellite sensor data.

Generalized Land Cover Class	GlobCover 2009 Class	GLC2000 Class	MODIS IGBP 2009 Class	Description
Cropland	11,14,20,30	23,16,17,18	12, 14	Agriculture and managed vegetation
Deciduous forest	50,60	2,3	3, 4	Close to open deciduous broadleaf trees
Evergreen forest	70,90	4	1, 2	Close to open evergreen needleleaf trees
Mixed forest	100	6,9	5	Mixed broadleaf and needleleaf trees
Shrubland	110,130,150	11,12,14	6, 7, 8	Shrub and sparse herbaceous
Grassland	120,140	13	9, 10	Herbaceous vegetation, rangeland
Urban/built up	190	22	13	Urban, mixed urban or artificial land

purposes of filtering in near-real-time (NRT), shape constraints are used to minimize edge effects. The shape constraints are derived from the previous analysis of historical data. The processing ensures a significant reduction in high frequency noise and other artefacts resulting from undetected clouds and poor atmospheric conditions. The final product resulted in a filtered (smoothed and gap-filled) time series with a 7-day interval (52 observations per year) integrating data from MODIS Terra and Aqua platforms. The 5-year time series was summarized to provide 7-day inter-annual averages (n = 52) and the corresponding standard deviations (n = 52) for the period 2007 – 2011. The 104 multi-temporal observations representing the inter-annual averages and standard deviations centred on the year 2009 provided the main input for the land cover classification. The various steps of the time series data processing are exemplified in Fig. 1 for one arbitrarily selected MODIS pixel.

2.2 *Reference Dataset*

Similar to Vuolo & Atzberger (2012), the reference dataset was generated by visual interpretation of high spatial resolution images available in Google Earth (GE). The reference data was subsequently used to train the classification algorithm, to validate the results and to compare our map with existing LC products. A software toolbox was used to assist the display of GE images and to add the visually determined LC label to each of the surveyed point. This process was supported by the visualization of the temporal curves of NDVI values for each point under survey. Two quality indicators were also assigned: i) the confidence of interpretation, and ii) the homogeneity of the area under interpretation. The first index categorizes the uncertainties arising while interpreting the high spatial resolution images. The second index expresses the level of pixel homogeneity observed in the GE high spatial resolution images. A description of the toolbox and quality flags can be found in Vuolo & Atzberger (2012).

In our first study Vuolo & Atzberger (2012), we considered 1235 samples randomly selected over three core test sites in Europe (Austria, France and Macedonia). For the purpose of the present paper, we extended this reference dataset to a total number of 6383 points (1235 samples acquired from the previous study, 2526 acquired over seven additional core test sites and 2622 points to cover the remaining territory). The selection procedure was based on a random stratified sampling of Europe's main biogeographical regions (EEA 2012). Fig. 2 shows the extent of the study site. The regions considered were: *Alpine* (1), *Continental* (5), *Mediterranean* (7), *Pannonian* (8), *Atlantic* (10) and *Boreal* (11). The number of samples for each region and land cover class are presented in Tab. 2.

To reduce thematic errors in the reference dataset caused by the visual interpretation and spatial mismatch in the comparison of existing LC datasets, reference points that were qualified as 'Low homogeneity' and/or 'unsure' (n = 1717) were excluded from further analysis.

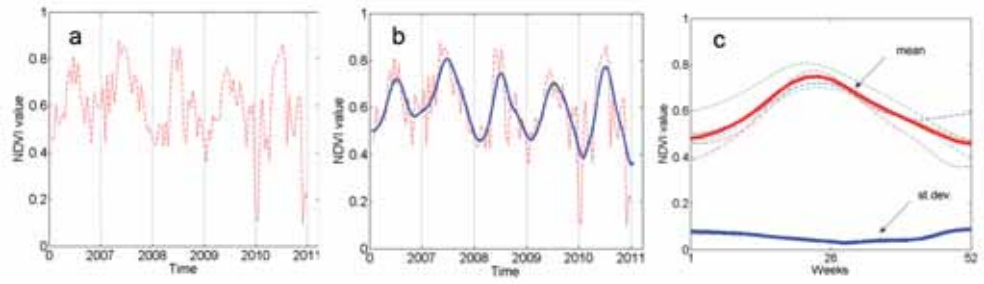

Fig. 1: a, b: Example of NDVI time series before and after filtering, c: temporal aggregation. Data smoothing was achieved in one step from year 2006 to 2011 using the Whittaker smoother as implemented at BOKU. The 104 features used for classification are shown in c: 52 mean values and 52 corresponding standard deviations.

The final dataset (n_{tot} = 4666) was randomly split into two subsamples (training and validation). The first was solely used to train the classification algorithm. Only the validation samples were used to evaluate the classification algorithm and to perform an intercomparison with existing LC products.

Tab. 2: Number of samples of the reference dataset for each biogeographical region and land cover type. The randomized division of reference samples into training (n = 2342) and validation data (n = 2324) was done by class for each region.

	Biogeographical region						Total no. samples
	1	**5**	**7**	**8**	**10**	**11**	
Cropland	78	695	450	59	186	68	**1536**
Deciduous F.	105	203	115	28	34	17	**502**
Evergreen F.	247	98	144	10	142	178	**819**
Mixed F.	124	231	67	7	49	122	**600**
Shrubland	103	25	194	1	42	14	**379**
Grassland	176	61	67	2	164	14	**484**
Urban	34	144	79	13	58	18	**346**
Total no. samples	**867**	**1457**	**1116**	**120**	**675**	**431**	**4666**
Area (km^2)	710	1959	1375	146	943	1016	× 1000
Sample / km^2	1.2	0.74	0.81	0.82	0.72	0.42	

2.3 *Classification Algorithm*

The land cover classification was performed by the machine learning random forest (RF) algorithm, which is an ensemble of decision trees. It uses bootstrap aggregating, i.e. bagging, to create different training subsets to produce a diversity of trees, each providing a classification result. The output class is obtained as the majority vote of the outputs of a large number of individual trees (Breiman 2001).

The algorithm produces an internal unbiased estimate of the generalization error, using the so-called 'out-of-bag' (OOB) samples (which are not included in the training subset) and provides a measure of the input features importance through random permutation. The randomized sampling leads to increased stability and better classification accuracy compared to a single decision tree approach. RF has been successfully applied in several classification problems achieving good results (Gislason et al. 2006, Rodriguez-Galiano et al. 2012a, Conrad et al. 2013, Toscani et al. 2013).

In this study we used a Matlab implementation of RF (Liaw & Wiener 2002, Abhishek

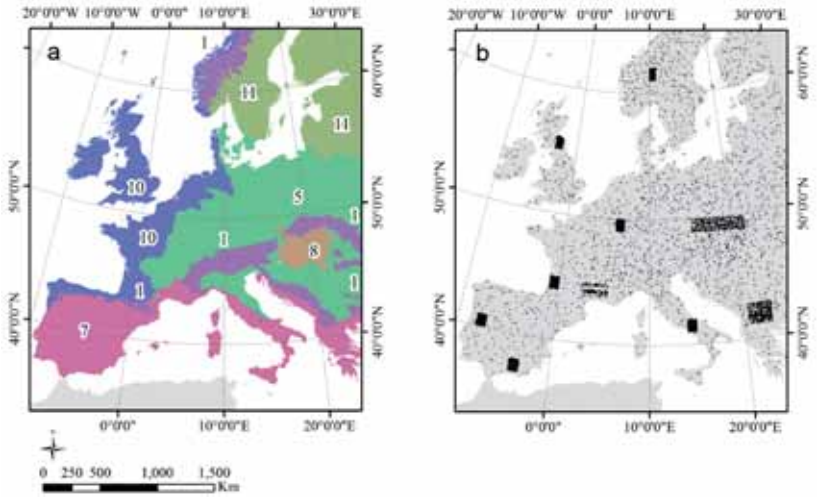

Fig. 2: a: Map of the biogeographical regions used for the stratification of the land cover classification, b: Reference dataset. The regions considered in this study were: Alpine (1), Continental (5), Mediterranean (7), Pannonian (8), Atlantic (10) and Boreal (11). The study area comprises most of the European area with a geographic extent from 10°00'W to 30°00'E and 35°00'N to 70°00'N. Regions with higher density of sampling points correspond to "core sampling sites".

2009). The software requires the setting of two parameters that are i) the number of trees to be grown in the run (*ntree*) and ii) the number of features used in each split (*mtry*). Several studies demonstrated that the default model parameters provide satisfactory results (Liaw & Wiener 2002, Díaz-Uriarte & Alvarez de Andrés 2006). We tested various combinations of *ntree* and *mtry*, which did not significantly affect the classification results. Therefore, as suggested in the software manual (Breiman & Cutler 2003), *mtry* was set equal to the square-root of the total number of input features (*mtry* = 10) using a sufficient number of trees (1000 in our case) so that adding more trees does not result in a significant performance gain.

2.4 *Accuracy Assessment and Classification Confidence Score*

The classification performance was based on common statistical measures (Foody 2002) derived from the classification error matrix, using solely validation samples. The selected statistical measures included the overall accuracy (OA), the producer's accuracy (PA) and the user's accuracy (UA). The two-side confidence intervals (CI) for the OA were calculated at 95% confidence level using the normal approximation method (Brown et al. 2001) including a continuity correction. The statistical significance of the differences between the pairs BOKU-GlobCover and BOKU-IGBP was evaluated with the McNemar's test with continuity correction (Siegel 1956).

The OOB error was used to assess the performance of the final pan-European LC map, which was generated using the entire reference dataset (training and validation samples). Two pixel-level measures of confidence of land cover classification were also obtained. A confidence score was calculated as the proportion of votes of the winning class to the total number of trees used in the classification. The higher the score, the more confident we are that a class is correctly classified. The second measure indicates the margin calculated as the proportion of votes for the winning class minus the proportion of votes of the second class.

3 Results

3.1 *Accuracy of Land Cover Classification*

The overall accuracies (OA) for the four LC maps produced by BOKU, IGBP, GlobCover and GLC2000 are presented in Tab. 3 for each biogeographical region and combined for all regions. Compared to the independent set of validation samples, BOKU achieved an OA of 71% (95% CI: 69% – 73%). The maps produced by IGBP, GlobCover and GLC2000 reported significantly lower OA ranging between 57% (GLC2000) and 63% (IGBP). From the numbers reported in Tab. 3 we calculated the percentage of increase or decrease of the overall accuracy of BOKU with respect to the other LC products. In this comparison, BOKU resulted 13%, 24% and 17% more accurate than IGBP, GlobCover and GLC2000, respectively. According to the two-tailed P-value, differences are considered to be extremely statistically significant ($p < 0.001$).

Regarding the six strata, and for all LC products, the *Boreal* biogeographical region was the least accurately classified. The *Alpine* biogeographical was inaccurately classified in all products, except BOKU. For the two mentioned regions GlobCover yielded the lowest accuracy. Results are not surprising since data samples in these two regions presented a lower level of confidence in the visual interpretation (due to the limited availability of high resolution images especially for the *Boreal* region) and a lower quality in the MODIS data input. The highest classification accuracies were ob-

Tab. 3: Overall accuracy for BOKU, IGBP (2009), GlobCover (2009) and GLC2000 for each biogeographical region and combined. All numbers refer to the independent validation dataset not used during training.

	BOKU	IGBP	GlobCover	GLC2000
Alpine (1)	74%	58%	47%	52%
Continental (5)	75%	70%	64%	61%
Mediterranean (7)	67%	61%	58%	63%
Pannonian (8)	83%	79%	63%	76%
Atlantic (10)	72%	56%	60%	71%
Boreal (11)	58%	58%	49%	52%
Combined regions	**71%**	**63%**	**57%**	**61%**

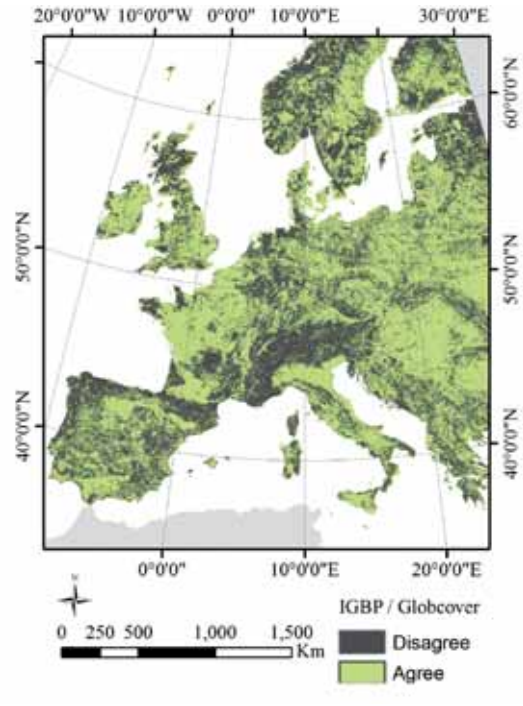

Fig. 3: Thematic agreement and disagreement between MODIS IGBP 2009 and GlobCover 2009.

tained for the *Pannonian* biogeographical region. The range of classification accuracies for the different LC products was larger for *Alpine* region (47% – 74%), compared to *Boreal* region (49% – 58%) region. For all regions, BOKU gave the highest OA.

The class-specific accuracies were analysed using the combined dataset. The error matrices are presented in Tab. 4. In details, *Cropland* reported a very high accuracy for most of the LC products, with producer's and user's accuracies greater than 70% in all cases. BOKU reported a very high producer's accuracy for *Cropland* (86%), indicating a good identification for all points visually interpreted as this class. Errors in the user's accuracy were often due to confusion with *Shrubland* and *Urban*. This result indicates that BOKU produces an overestimation of *Cropland* class with a commission error of 22%. A similar PA was observed in IGBP, with lower user's accuracies due to confusion of *Cropland* with *Deciduous Forest* and *Grassland* class. Regarding forest LC classes for BOKU, *Deciduous Forest* class was classified with producer's and user's accuracies of 67%, being often confused with *Mixed Forest* class. *Evergreen Forest* presented a higher accuracy with a producer's and user's accuracies greater than

Tab. 4: Error matrix and statistical measures for BOKU, IGBP 2009, GlobCover and GLC2000. All numbers refer to the independent validation dataset not used during training.

	Reference								
BOKU	C	DF	EF	MF	S	G	U	∑	UA
Cropland (C)	**660**	12	18	14	50	44	50	848	78%
Decid. F. (DF)	11	**166**	6	48	5	8	2	246	67%
Evergr. F. (EF)	22	15	**304**	56	19	15	1	432	70%
Mixed F. (MF)	23	47	51	**172**	2	12	0	307	56%
Shrubland (S)	18	2	13	1	**90**	17	3	144	63%
Grassland (G)	19	6	16	7	13	**145**	0	206	70%
Urban (U)	14	1	1	0	10	5	**127**	158	80%
∑	767	249	409	298	189	246	183	2341	
PA	86%	67%	74%	58%	48%	59%	69%		
OA	71% (95% C.I.: 69%–73%)								

	Reference								
IGBP (2009)	C	DF	EF	MF	S	G	U	∑	UA
Cropland (C)	**650**	83	15	38	18	77	25	906	72%
Decid. F. (DF)	3	**83**	3	24	1	2	0	116	72%
Evergr. F. (EF)	9	12	**219**	45	11	13	7	316	69%
Mixed F. (MF)	27	63	106	**176**	7	12	3	394	45%
Shrubland (S)	49	6	41	10	**143**	62	20	331	43%
Grassland (G)	13	1	17	2	5	**41**	1	80	51%
Urban (U)	13	1	1	0	0	1	**126**	142	89%
∑	764	249	402	295	185	208	182	2285	
PA	85%	33%	54%	60%	77%	20%	69%		
OA	63% (95% C.I.: 61%–65%)								

	Reference								
GlobCover	C	DF	EF	MF	S	G	U	∑	UA
Cropland (C)	**573**	28	12	17	20	52	41	743	77%
Decid. F. (DF)	43	**146**	50	103	8	19	6	375	39%
Evergr. F. (EF)	5	8	**215**	46	35	8	1	318	68%
Mixed F. (MF)	7	25	84	**96**	8	5	0	225	43%
Shrubland (S)	70	12	39	11	**99**	74	7	312	32%
Grassland (G)	55	28	6	23	13	**80**	3	208	38%
Urban (U)	8	0	2	0	0	4	**124**	138	90%
∑	761	247	408	296	183	242	182	2319	
PA	75%	59%	53%	32%	54%	33%	68%		
OA	57% (95% C.I.: 55%–60%)								

	Reference								
GLC2000	C	DF	EF	MF	S	G	U	∑	UA
Cropland (C)	**630**	51	34	28	23	55	62	883	71%
Decid. F. (DF)	33	**144**	26	58	12	18	5	296	49%
Evergr. F. (EF)	21	16	**260**	101	24	20	2	444	59%
Mixed F. (MF)	15	27	56	**98**	20	8	1	225	44%
Shrubland (S)	17	6	14	1	**92**	51	7	188	49%
Grassland (G)	46	5	15	10	10	**89**	3	178	50%
Urban (U)	2	0	1	2	1	3	**102**	111	92%
∑	764	249	406	298	182	244	182	2325	
PA	82%	58%	64%	33%	51%	36%	56%		
OA	61% (95% C.I.: 59%–63%)								

70%. *Mixed Forest* was often confused with *Deciduous* or *Evergreen Forest.* This trend was observed in all LC products, with additional confusion occurring between forest LC classes and *Cropland*, in particular in IBGP, with an omission error of 67%.

Shrubland presented the lowest PA in BOKU being often confused with *Cropland* and *Evergreen Forest* classes. This class was overestimated with a commission error of 37%. *Grassland* was classified with a producer's accuracy of nearly 60% and a user's accuracy of 70%. BOKU notably improved the classification accuracy for this class compared to the other LC products. Regarding *Urban* class, all LC products provided a good user's accuracy (> 80%), with GlobCover and GLC2000 achieving the best results. We observed an omission error of about 20% – 30% due to confusion of this class with *Cropland.*

3.2 *Accuracy for Areas of Agreement and Disagreement*

The classification accuracy was also evaluated separately for samples where GlobCover and IGBP maps agreed and where these maps disagreed. For this detailed analysis we considered only GlobCover and IGBP because they have a similar spatial resolution (300 m and 500 m pixel size) and refer to the same year (2009). The map of the areas of agreement and disagreement between these two products is presented in Fig. 3. The OA for BOKU, IGBP and GlobCover are presented in Tab. 5 for each

Tab. 5: The overall accuracy for IGBP, GlobCover and BOKU for areas of agreement and of disagreement between IGBP and GlobCover, respectively. The OA was assessed using the independent validation dataset (n = 2324 samples, of which 1230 were in agreement and 1094 in disagreement).

	IGBP = GlobCover			Total no. of samples	IGBP ≠ GlobCover			Total no. of samples
	IGBP	**GlobCover**	**BOKU**		**IGBP**	**GlobCover**	**BOKU**	
Alpine (1)	65%	65%	80%	226	50%	27%	68%	206
Continental (5)	88%	88%	84%	415	46%	31%	63%	311
Mediterranean (7)	78%	78%	73%	299	43%	36%	59%	257
Pannonian (8)	93%	93%	96%	28	60%	37%	70%	30
Atlantic (10)	80%	80%	72%	157	27%	43%	71%	180
Boreal (11)	76%	76%	68%	105	41%	22%	49%	110
Combined regions	**79%**	**79%**	**78%**	1230	**43%**	**33%**	**63%**	1094

Tab. 6: The overall accuracy (combined regions) for IGBP, GlobCover and BOKU for all possible combinations of agreement among the three land cover products.

	IGBP	GlobCover	BOKU	Total no. of samples
IGBP = GlobCover = BOKU	90%	90%	90%	941
IGBP = BOKU ≠ GlobCover	70%	16%	70%	449
IGBP ≠ BOKU = GlobCover	19%	72%	72%	316
IGBP = GlobCover ≠ BOKU	44%	44%	40%	289
IGBP ≠ GlobCover ≠ BOKU	26%	15%	44%	329

biogeographical region and combined. Results for areas of agreement and disagreement are presented separately.

In areas where GlobCover and IGBP agreed (n = 1230), we found an overall classification accuracy (combined regions) of 80% for IGBP / GlobCover and of 79% for BOKU. Percentage differences between BOKU and the combination of IGBP / GlobCover were not statistically significant ($p > 0.05$). Regarding the disagreement samples (n = 1094), we observed a general decrease in OA. However, this reduction was only modest for BOKU (from 71% to 63%) compared to the dramatic drop in

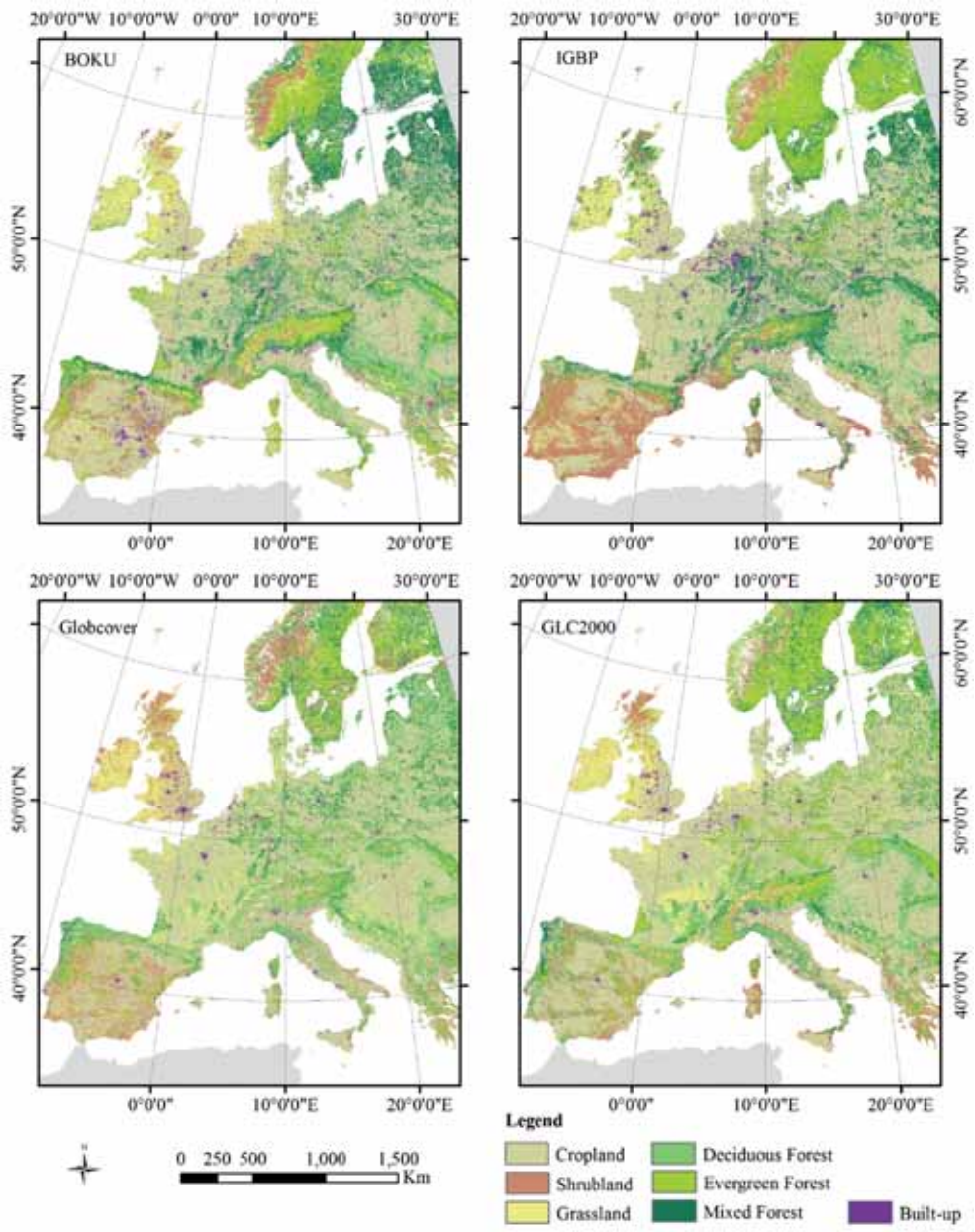

Fig. 4: BOKU, IGBP, GlobCover and GLC2000 land cover maps.

overall classification accuracy for GlobCover and IGBP. For example, in the case of IGBP, the overall classification accuracy dropped from 63% to 44%. For GlobCover the OA decreased from 57% to 33%. Hence, for the disagreement samples, BOKU resulted 43% and 95% more accurate compared to IGBP and GlobCover, respectively. Particularly remarkable improvements were achieved for *Atlantic*, *Alpine* and *Boreal* regions (Tab. 5).

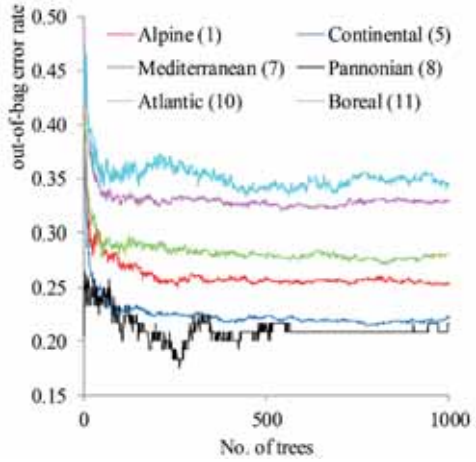

Fig. 5: Cumulative out-of-bag (OOB) error rate for the BOKU LC map for each biogeographical region. It shows how the OOB error changes while adding more trees to the ensemble classifier.

Classification results were further investigated considering all possible combinations of agreement (and disagreement) among the three LC products and they are presented in Tab. 6 for the combined biogeographical regions. In the case of a complete agreement (n = 941), we found an overall classification accuracy of 90%, which notably decreases where all products disagree (n = 329). Where BOKU agrees either with IGBP or with GlobCover, it achieved an OA greater than 70% indicating that our LC classification is more accurate (from 63% up to 72%) in the case of agreement with at least one of the two other LC products. Finally, the OA resulted slightly higher for the pair IGBP / GlobCover where they are in disagreement with BOKU. However, this latter combination presented the lower number of events (n = 289).

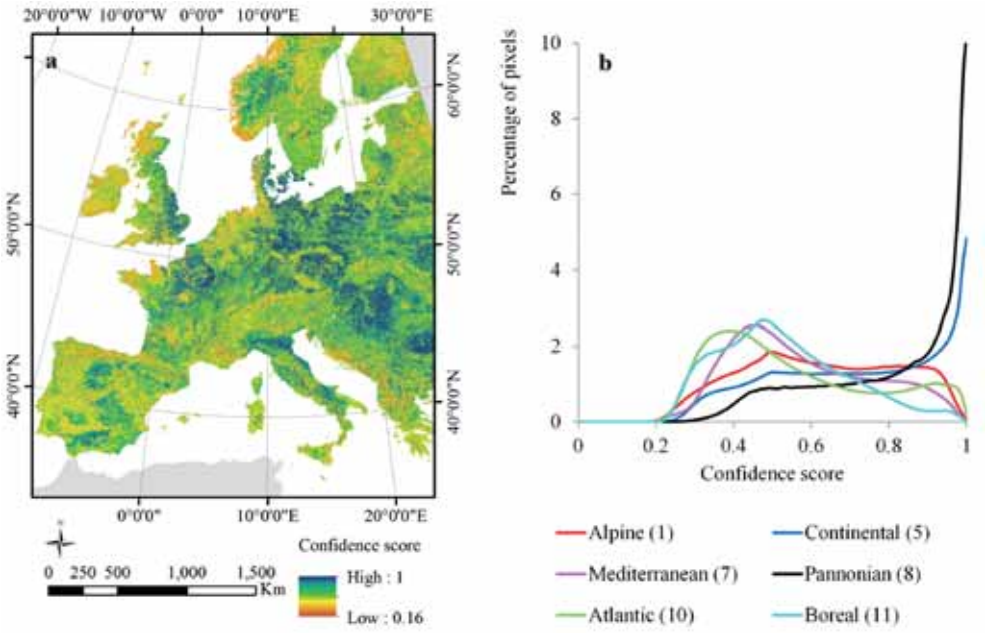

Fig. 6: a: Confidence scores for the BOKU LC map, b: Distribution of the confidence score values for the six biogeographic regions (number of bins = 100). A value of 1 indicates that all classification trees voted for the same class (maximum possible score). With seven LC classes, the theoretical minimum is 1/7.

3.3 *Map Production*

For producing the final map, we exploited the entire reference dataset (n_{tot} = 4666) without splitting data in training and validation samples. Fig. 4 illustrates the BOKU map along with IGBP, GlobCover and GLC2000 products. Fig. 5 shows the cumulative OOB error rate for each biogeographical region over the 1000 classification trees.

The total accuracy was obtained as the last element of the OOB error rate minus one. The combined region achieved an average accuracy of 72% (versus 71% with the independent dataset) and of 75% (Alpine), 78% (Continental), 67% (Mediterranean), 78% (Pannonian), 72% (Atlantic) and 65% (Boreal), respectively. It is interesting to notice that these figures were comparable to the accuracy measures obtained from the independent validation dataset (first column in Tab. 3). The analysis of our dataset confirmed that the OOB error rate provides an unbiased estimation of the error and therefore eliminates the need for an independent validation dataset (Breiman 2001). However, we acknowledge that these accuracy measures might not be directly comparable with the OA obtained with the independent validation dataset since a different number of samples was used.

Fig. 6 shows the confidence score map and spatial distribution of score values. A score of

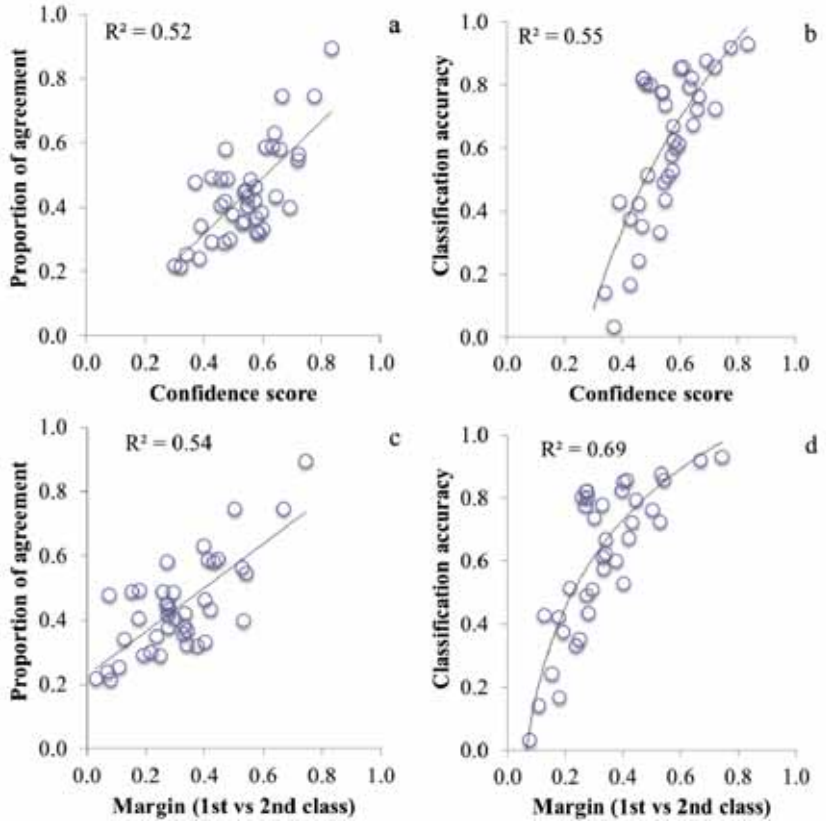

Fig. 7: a: Relationship of the confidence score with the proportion of agreement, b: Relationship of the confidence score with the classification accuracy, c: Relationship of the margin (defined as the difference in votes between the winning class and the second best class) with the proportion of agreement, d: Relationship of the margin with the classification accuracy. Each point represents the average value of one of the seven land cover classes within each of the six biogeographical regions (n = 42).

1 indicates that all classification trees voted for the same class, achieving the highest possible confidence.

Comparing Figs. 4 and 6, we notice that the pattern of the confidence score shows some spatial correlation with the land cover type and a clear positive trend with the class-specific accuracies. For instance, *Cropland* achieved a higher confidence score (0.70 ± 0.22 with a PA of 86%) compared to *Grassland* (0.49 ± 0.16 with a PA of 59%). A spatial correlation was also observed with the areas of agreement and disagreement between IGBP and GlobCover (Fig. 3).

Fig. 7a and b show the relationship between the confidence score and the proportion of pixels in agreement ($R^2 = 0.52$) and the classification accuracy ($R^2 = 0.55$), respectively. Similarly, Fig. 7c and d show the scatterplots between the classification margin and the proportion of pixels in agreement ($R^2 = 0.54$) and the classification accuracy ($R^2 = 0.69$), respectively. In all cases we observed a correlation showing that an increase in the confidence score corresponds also to an increase in the proportion of agreement and in the classification accuracy. For this comparison we used the producer's accuracy obtained from the independent validation dataset for each region.

4 Conclusions

This work evaluated the land cover (LC) classification performance of the random forest classifier using moderate-resolution imaging spectroradiometer (MODIS) 250 m normalized difference vegetation index (NDVI) time series at pan-European scale. The classification performance was compared to the overall accuracies of three existing products, GlobCover 2009, MODIS Land Cover Type IGBP 2009 and GLC2000. In particular, results were evaluated with respect to pixels in agreement and disagreement between GlobCover 2009 and MODIS IGBP 2009, i.e. 300 m and 500 m pixel size respectively. All results were obtained using harmonized legends with seven land cover classes. This disjoint analysis was performed to evaluate possible improvements in classification accuracy where existing maps are inconsistent (disagreement).

The results presented here expand our previous work (Vuolo & Atzberger 2012) and confirmed that there is a high potential in using multi-year (here 2007 – 2011, centred on the year 2009) time-series of MODIS 250 m NDVI for land cover classification. NDVI-derived BOKU LC maps achieved an overall accuracy of 71%. When comparing our results with the accuracy achieved by MODIS IGBP, GlobCover and GLC2000 aggregated to seven LC classes, we observed that BOKU LC map clearly outperforms these three (global) land cover products (see Tab. 4).

The findings at pan-European scale confirm that the classification accuracy of areas of agreement is systematically higher compared to areas where two (or more) maps disagree. Moreover, results demonstrated that in the case of the BOKU LC classification, the accuracy of data points for areas of disagreement was notably improved towards 43% and to 95% compared to IGBP and GlobCover, respectively (see Tab. 5). As expected, an improvement for agreement points was difficult to achieve, confirming that currently available LC products are already relatively accurate in areas of agreement (79% in our assessment and 90% where IGBP, GlobCover and BOKU agree).

Additionally, we analysed the consistency of two measures of confidence of classification, namely confidence score and margin. The two measures were obtained at pixel-level as direct outputs of the (ensemble) decision trees. We observed a clear relationship between the proportion of agreement/disagreement, accuracy and the confidence score/margin (see Fig. 7). Our results confirmed that these measures provide a reliable indication of confidence in classification for each pixel (Immitzer et al. 2012). For instance, the spatial pattern of the confidence score showed spatial correlation with the land cover type and – more importantly – a clear positive trend with the class-specific accuracies.

Regarding the bootstrapped error estimates obtained from the out-of-bag (OOB) samples, it is interesting to notice that these figures were comparable to the accuracy measures obtained from the independent validation dataset. Therefore, the analysis confirmed that the OOB error rate provides an unbiased es-

timation of the error and eliminates the need for an independent validation dataset (Breiman 2001).

Given the currently available global datasets, the users of LC products should in our opinion focus on combining existing maps and identify areas of agreement and disagreement. The accuracy of areas of spatial disagreement could then be improved based for example on the methodology proposed in this study. The proposed methodology can be easily generalized to different legend definitions or levels of land cover detail. This will help maximizing the overall accuracy of the resulting final land cover map as confirmed by our study.

Similar to several other studies (Jung et al. 2006, Kaptué Tchuenté et al. 2011, Fritz et al. 2011, Pflugmacher et al. 2011), our work confirmed that there is no clear preference of one LC product compared to others. A selection will always have to be based on a specific purpose or application.

Steps to further improve the accuracy of land cover maps include, for instance, ensembles of different algorithms for map production based on multi-source datasets (Benediktsson et al. 2007). Subsequently, these maps can be combined and synthetized in one product based on decision fusion rules (Waske & Benediktsson 2007, Udelhoven et al. 2009).

References

Abhishek, J., 2009: Classification and Regression by randomForest-matlab. – https://code.google.com/p/randomforest-matlab/ (12.12.2013).

Arino, O., Bicheron, P., Achard, F., Latham, J., Witt, R. & Weber, J.-L., 2008: GlobCover: The most detailed portrait of Earth. – European Space Agency Bulletin **2008** (136): 24–31.

Atkinson, P.M., Jeganathan, C., Dash, J. & Atzberger, C., 2012: Inter-comparison of four models for smoothing satellite sensor time-series data to estimate vegetation phenology. – Remote Sensing of Environment **123** (0): 400–417.

Atzberger, C. & Eilers, P.H.C., 2011a: A time series for monitoring vegetation activity and phenology at 10-daily time steps covering large parts of South America. – International Journal of Digital Earth **4** (5): 365–386.

Atzberger, C. & Eilers, P.H.C., 2011b: Evaluating the effectiveness of smoothing algorithms in the absence of ground reference measurements. – International Journal of Remote Sensing **32** (13): 3689–3709.

Bartholomé, E. & Belward, A.S., 2005: GLC2000: a new approach to global land cover mapping from Earth observation data. – International Journal of Remote Sensing **26** (9): 1959–1977.

Benediktsson, J.A., Chanussot, J., Fauvel, M., Haindl, M., Kittler, J. & Roli, F., 2007: Multiple Classifier Systems in Remote Sensing: From Basics to Recent Developments. – Springer, Berlin, Heidelberg.

Bossard, M., Feranec, J. & Otahel, J., 2000: CORINE land cover technical guide: Addendum 2000. – http://www.dmu.dk/fileadmin/Resources/DMU/Udgivelser/CLC2000/technical_guide_addenum.pdf (2.7.2014).

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

Breiman, L. & Cutler, A., 2003: Setting up, using, and understanding random forests V4.0. – University of California, Department of Statistics. – http://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf (2.7.2014).

Brown, L.D., Cai, T.T. & Dasgupta, A., 2001: Interval Estimation for a Binomial Proportion. – Statistical Science **16** (2): 101–133.

Conrad, C., Rahmann, M., Machwitz, M., Stulina, G., Paeth, H. & Dech, S., 2013: Satellite based calculation of spatially distributed crop water requirements for cotton and wheat cultivation in Fergana Valley, Uzbekistan. – Global and Planetary Change **110:** 88–98.

Díaz-Uriarte, R. & Alvarez de Andrés, S., 2006: Gene selection and classification of microarray data using random forest. – BMC bioinformatics **7** (1): 3.

EEA 2012: Biogeographic regions in Europe. – http://www.eea.europa.eu/data-and-maps/figures/biogeographical-regions-in-europe-1 (24.1.2014).

Faroux, S., Kaptué Tchuenté, A.T., Roujean, J.-L., Masson, V., Martin, E. & Le Moigne, P., 2013: ECOCLIMAP-II/Europe: a twofold database of ecosystems and surface parameters at 1 km resolution based on satellite information for use in land surface, meteorological and climate models. – Geoscientific Model Development **6** (2): 563–582.

Foody, G.M., 2002: Status of land cover classification accuracy assessment. – Remote Sensing of Environment **80** (1): 185–201.

Friedl, M., McIver, D., Hodges, J.C., Zhang, X., Muchoney, D., Strahler, A., Woodcock, C., Gopal, S., Schneider, A., Cooper, A., Baccini, A., Gao, F. & Schaaf, C., 2002: Global land cover mapping from MODIS: algorithms and early

results. – Remote Sensing of Environment **83** (1–2): 287–302.

Friedl, M.A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A. & Huang, X., 2010: MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. – Remote Sensing of Environment **114** (1): 168–182.

Fritz, S., See, L., McCallum, I., Schill, C., Obersteiner, M., Marijn Hannes Havlík, P., Achard, F.F., van der Velde, M. & Boettcher, H., 2011: Highlighting continued uncertainty in global land cover maps for the user community. – Environmental Research Letters **6** (4), http://iopscience.iop.org/1748-9326/6/4/044005/fulltext/ (8.7.2014).

Gislason, P.O., Benediktsson, J.A. & Sveinsson, J.R., 2006: Random Forests for land cover classification. – Pattern Recognition Letters **27** (4): 294–300.

Herold, M., Hubald, R. & Di Gregorio, A., 2009: Translating and evaluating land cover legends using the UN Land Cover Classification System (LCCS). – Workshop Report at FAO, GOFC-GOLD 43, Jena.

Herold, M., Mayaux, P., Woodcock, C.E.E., Baccini, A. & Schmullius, C., 2008: Some challenges in global land cover mapping: An assessment of agreement and accuracy in existing 1 km datasets. – Remote Sensing of Environment **112** (5): 2538–2556.

Herold, M., Woodcock, C.E., Mayaux, P., Belward, A.S., Latham, J. & Schmullius, C., 2006: A joint initiative for harmonization and validation of land cover datasets. – IEEE Transactions on Geoscience and Remote Sensing **44** (7): 1719–1727.

Hibbard, K., Janetos, A., van Vuuren, D.P., Pongratz, J., Rose, S.K., Betts, R., Herold, M. & Feddema, J.J., 2010: Research priorities in land use and land-cover change for the Earth system and integrated assessment modelling. – International Journal of Climatology **30** (13): 2118–2128.

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** (1–2): 195–213.

Jung, M., Henkel, K., Herold, M. & Churkina, G., 2006: Exploiting synergies of global land cover products for carbon cycle modeling. – Remote Sensing of Environment **101** (4): 534–553.

Kaptué Tchuenté, A.T., Roujean, J.-L. & De Jong, S.M., 2011: Comparison and relative quality assessment of the GLC2000, GlobCover, MODIS and ECOCLIMAP land cover data sets at the African continental scale. – International Journal of Applied Earth Observation and Geoinformation **13** (2): 207–219.

Immitzer, M., Atzberger, C. & Koukal, T., 2012: Tree Species Classification with Random Forest Using Very High Spatial Resolution 8-Band WorldView-2 Satellite Data. – Remote Sensing **4** (9): 2661–2693.

Landcover, 2014: http://www.esa-landcover-cci.org/ (7.7.2014).

Liaw, A. & Wiener, M., 2002: Classification and Regression by randomForest. – R news **2** (December): 18–22.

Liu, W., Gopal, S. & Woodcock, C.E., 2004: Uncertainty and Confidence in Land Cover Classification Using a Hybrid Classifier Approach. – Photogrammetric Engineering & Remote Sensing **70** (8): 963–971.

Masson, V., Champeaux, J.-L., Chauvin, F., Meriguet, C. & Lacaze, R., 2003: A Global Database of Land Surface Parameters at 1-km Resolution in Meteorological and Climate Models. – Journal of Climate **16** (9): 1261–1282.

Pérez-Hoyos, A., García-haro, F.C.C.F.J. & San-Miguel-Ayanz, J., 2012: A methodology to generate a synergetic land-cover map by fusion of different land-cover products. – International Journal of Applied Earth Observation and Geoinformation **19:** 72–87.

Pflugmacher, D., Krankina, O.N., Cohen, W.B., Friedl, M.A., Sulla-Menashe, D., Kennedy, R.E., Nelson, P., Loboda, T.V., Kuemmerle, T., Dyukarev, E., Elsakov, V. & Kharuk, V.I., 2011: Comparison and assessment of coarse resolution land cover maps for Northern Eurasia. – Remote Sensing of Environment **115** (12): 3539–3553.

Rodriguez-Galiano, V.F., Chica-Olmo, M., Abarca-Hernandez, F., Atkinson, P.M. & Jeganathan, C., 2012a: Random Forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. – Remote Sensing of Environment **121:** 93–107.

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

See, L. & Fritz, S., 2006: A method to compare and improve land cover datasets: Application to the GLC-2000 and MODIS land cover products. – IEEE Transactions of Geoscience and Remote Sensing **44** (7): 1740–1746.

Siegel, S., 1956: Nonparametric statistics for the behavioral sciences. – McGraw-Hill series in psychology McGraw-Hill, New York, NY, USA.

Toscani, P., Immitzer, M. & Atzberger, C., 2013: Wavelet-based texture measures for object-based classification of aerial images. – PFG – Photogrammetrie, Fernerkundung, Geoinformation **2013** (2): 105–121.

Udelhoven, T., van der Linden, S., Waske, B., Stellmes, M. & Hoffmann, L., 2009: Hypertemporal Classification of Large Areas Using Decision Fusion. – IEEE Geoscience and Remote Sensing Letters **6** (3): 592–596.

Vancutsem, C., Marinho, E., Kayitakire, F., See, L. & Fritz, S., 2013: Harmonizing and Combining Existing Land Cover/Land Use Datasets for Cropland Area Monitoring at the African Continental Scale. – Remote Sensing **5** (1): 19–41.

Vermote, E.F., El Saleous, N.Z. & Justice, C.O., 2002: Atmospheric correction of MODIS data in the visible to middle infrared: first results. – Remote Sensing of Environment **83** (1–2): 97–111.

Vintrou, E., Desbrosse, A., Bégué, A., Traoré, S., Baron, C. & Lo Seen, D., 2012: Crop area mapping in West Africa using landscape stratification of MODIS time series and comparison with existing global land products. – International Journal of Applied Earth Observation and Geoinformation **14** (1): 83–93.

Vuolo, F. & Atzberger, C., 2012: Exploiting the Classification Performance of Support Vector Machines with Multi-Temporal Moderate-Resolution Imaging Spectroradiometer (MODIS) Data in Areas of Agreement and Disagreement of Existing Land Cover Products. – Remote Sensing **4** (12): 3143–3167.

Waske, B. & Benediktsson, J.A., 2007: Fusion of Support Vector Machines for Classification of Multisensor Data. – Geoscience and Remote Sensing, IEEE Transactions on **45** (12): 3858–3866.

Address of the Authors

Dr. Francesco Vuolo & Prof. Dr. Clement Atzberger, University of Natural Resources and Life Sciences, Institute of Surveying, Remote Sensing and Land Information (IVFL), Peter-Jordan-Straße 82, A-1190 Wien, Tel.: +43-1-47654-5100, Fax.: +43-1-47654-5142, e-mail: {francesco.vuolo} {clement.atzberger} @boku.ac.at

Manuskript eingereicht: März 2014
Angenommen: Juli 2014