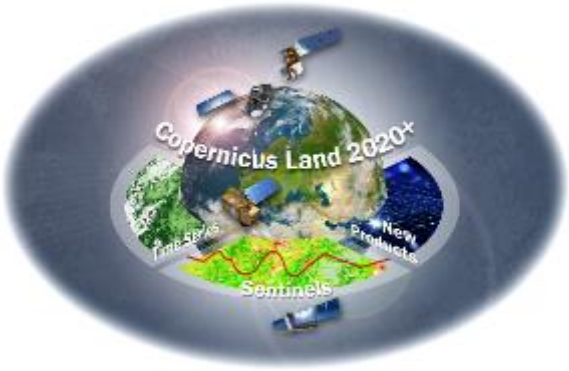
Horizon 2020

Space Call - Earth Observation: EO-3-2016: Evolution of Copernicus services

Grant Agreement No. 730008

ECoLaSS

Evolution of Copernicus Land Services based on Sentinel data



D8.1

"D33.1a – Time Series Analysis for Thematic Classification (Issue 1)"

Issue/Rev.: 1.0

Date Issued: 29.03.2018

submitted by:

UCL Université catholique de Louvain

in collaboration with the consortium partners:

SIRS JOANNEUM RESEARCH DIGITAL GAFAG DLR

submitted to:

European Commission – Research Executive Agency

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme, under Grant Agreement No. 730008.

CONSORTIUM PARTNERS

NO.	PARTICIPANT ORGANISATION NAME	SHORT NAME	CITY, COUNTRY
1	GAF AG	GAF	Munich, Germany
2	Systèmes d'Information à Référence Spatiale SAS	SIRS	Villeneuve d'Ascq, France
3	JOANNEUM RESEARCH Forschungsgesellschaft mbH	JR	Graz, Austria
4	Université catholique de Louvain, Earth and Life Institute (ELI)	UCL	Louvain-la-Neuve, Belgium
5	German Aerospace Center (DLR), German Remote Sensing Data Center (DFD), Wessling	DLR	Wessling, Germany

CONTACT:

GAF AG
Arnulfstr. 199 – D-80634 München – Germany
Phone: ++49 (0)89 121528 0 – FAX: ++49 (0)89 121528 79
E-mail: copernicus@gaf.de – Internet: www.gaf.de

DISCLAIMER:

DOCUMENT RELEASE SHEET

	NAME, FUNCTION	DATE	SIGNATURE
Author(s):	Moreau Inès (UCL) Jolan Wolter (UCL) Pierre Defourny (UCL) Heinz Gallaun (JR) Janik Deutscher (JR) Petra Miletich (JR) Klaus Granica (JR) Benedikt Zörfus (JR) Clémence Kenner (SIRS) Alexandre Pennec (SIRS) Sophie Villerot (SIRS) Benjamin Mack (GAF) Carolin Sommer (GAF) Christophe Rieke (GAF) Cornelia Storch (GAF) Katharina Schwab (GAF) Kathrin Schweitzer (GAF) Linda Moser (GAF) Martin Ickerott (GAF) Annekatrin Metz (DLR) Soner Üreyen (DLR) Igor Klein (DLR)	27/03/2018	
Review:	Inès Moreau (UCL) Monika Kovatsch (GAF) Katharina Schwab (GAF)	28/03/2018	
Approval:	Linda Moser (GAF)	29/03/2018	
Acceptance:	Massimo Ciscato (REA)		
Distribution:	Public		

DISSEMINATION LEVEL

DISSEMINATION LEVEL		
PU	Public	**X**
CO	Confidential: only for members of the consortium (including the Commission Services)	

DOCUMENT STATUS SHEET

ISSUE/REV	DATE	PAGE(S)	DESCRIPTION / CHANGES
1.0	29.03.2018	171	First issue of the Methods Compendium: Time series Analysis for Thematic Classification

APPLICABLE DOCUMENTS

ID	DOCUMENT NAME / ISSUE DATE
AD01	Horizon 2020 Work Programme 2016 – 2017, 5 iii. Leadership in Enabling and Industrial Technologies – Space. Call: EO-3-2016: Evolution of Copernicus services. Issued: 13.10.2015
AD02	Guidance Document: Research Needs Of Copernicus Operational Services. Final Version issued: 30.10.2015
AD03	Proposal: Evolution of Copernicus Land Services based on Sentinel data. Proposal acronym: ECoLaSS, Proposal number: 730008. Submitted: 03.03.2016
AD04	Grant Agreement – ECoLaSS. Grant Agreement number: 730008 – ECoLaSS – H2020-EO-2016/H2020-EO-2016, Issued: 18.10.2016
AD05	D21.1a - Service Evolution Requirements Report, Issue 1.0, Issued: 09.08.2017
AD06	D31.1 - Methods Compendium: Sentinel-1/2/3 Integration Strategies, Issue 1.0, Issued: March 2018
AD07	D32.1- Methods Compendium: Time Series Preparation, Issue 1.0, Issued: February 2018

EXECUTIVE SUMMARY

The Horizon 2020 (H2020) project, "Evolution of Copernicus Land Services based on Sentinel data" (ECoLaSS) addresses the H2020 Work Programme 5 iii. Leadership in Enabling and Industrial technologies - Space, specifically the Topic EO-3-2016: Evolution of Copernicus services. ECoLaSS will be conducted from 2017–2019 and aims at developing and prototypically demonstrating selected innovative products and methods for future next-generation operational Copernicus Land Monitoring Service (CLMS) products of the pan-European and Global Components. This will contribute to demonstrating operational readiness of the finally selected products, and shall allow the key CLMS stakeholders (i.e. mainly the Entrusted European Entities (EEE) EEA and JRC) to take informed decisions on potential procurement of the next generation of Copernicus Land services from 2020 onwards.

To achieve this goal, ECoLaSS will make full use of dense time series of Sentinel-2 and Sentinel-3 optical data as well as Sentinel-1 Synthetic Aperture Radar (SAR) data. Rapidly evolving scientific as well as user requirements will be analysed in support of a future pan-European roll-out of new/improved CLMS products, and the transfer to global applications.

This report constitutes a methods compendium for the investigated approaches of the work package 33 "Time Series Analysis for Thematic Classification" of ECoLaSS Task 3 (Automated High Data Volume Processing Lines). The objective of this WP to develop a framework for time series analysis for thematic classification based on Sentinel multi-sensor constellation. For this purpose, the WPP aims at developing and benchmarking (i) optical image compositing methods specifically dedicated to thematic classification, and (ii) time series classification methods for HR layers, crop type and new land cover/land use products. With the others WP of ECoLaSS Task 3 (Automated High Data Volume Processing Lines), it constitutes a basis for the demonstration activities of Task 4 (Thematic Proof-of-Concept/Prototype on Continental/Global Scale), i.e. High Resolution Layers (HRLs), Grassland, Crop type and new LC/LU products.

Section 1 of the document present the purpose and objectives of the WP, and the document structure. Section 2 describes the state-of-the-art methods and strategies for the selection of candidate methods for the benchmarking. It reviews the automated reference sampling methods and the image compositing methods needed for classification, and then provides state-of-the-art of time series classification methods for time series HRLs, agriculture and new land cover products. Based on this reviews and on the selection of candidate methods, section 3 concerns the testing and benchmarking of input data for classification and of time series classification approaches. For each benchmark, a conclusion explains the main outcomes and recommendations of the analysis. The benchmark of automated reference sampling is performed on two methods, five compositing methods are assessed and compared, and classification approaches are benchmarked separately for different thematic fields: (i) Imperviousness, (ii) Forest, (iii) Grassland, (iv) Agriculture, and (v) new land cover products. For each of the thematic classifications, different inputs, classification methods and parameters are assessed. Finally, section 4 concludes the document by summarising the main outcomes of the benchmarking.

The ECoLaSS project follows a two-phased approach of two times 18 months duration. This deliverable comprises the first issue, where preliminary results are presented. In the second 18-month project cycle, a second issue of this deliverable will be published, containing all relevant updates.

TABLE OF CONTENTS

List of Figures

List of Tables

Abbreviations

AL	Active Learning
ANNs	Artificial Neural Networks
AOI	Area Of Interest
AUC	Area Under the ROC Curve
AVG	Average
AVHRR	Advanced Very High Resolution Radiometer
BISE	Best Index Slope Extraction
BDC	Bi-Directional Compositing
BPA	Best Available Pixel
BRDF	Bidirectional Reflectance Distribution Function
CLC	Corine Land Cover
CLMS	Copernicus Land Monitoring Services
CORDA	Copernicus Reference Data Access
CORINE	Coordination of Information on the Environment
COV	Coefficient Of Variation
CT	Classification Trees
CYC	Cyclope
DAP	Differential Attribute Profiles
DEM	Digital Elevation Model
DFA	Discriminant Function Analysis
DLT	Dominant Leaf Type
DMP	Differential Morphological Profile
DMSP-OLS	Defense Meteorological Satellite Program's Operational Line-scan System
DST	Dempster–Shafer Theory
DWH	Data Warehouse
EC	European Commission
ECoLaSS	Evolution of Copernicus Land Services based on Sentinel data
EEA	European Environment Agency
EEE	Entrusted European Entities
EO	Earth Observation
ERS	European Remote-Sensing Satellite
ESA	European Space Agency
ESRI	Environmental Systems Research Institute
ETM+	Enhanced Thematic Mapper Plus
EU	European Union
EVI	Enhanced Vegetation Index
FAO	Food and Agriculture Organization
GC	Ground Cover
GHSL	Global Human Settlement Layer
GLC	Global Land Cover
GLCM	Grey Level Co-occurrence Matrix
GIO	GMES Initial Operations
GIS	GeoEye Imaging System Geographic Information System
GMES	Global Monitoring for Environment and Security
GRAVPI	Grass Vegetation Probability Index
GRD	Ground Range Detected
GUF	Global Urban Footprint
ha	Hectare
HH	Horizontal transmit/Horizontal receive (polarization)

HR	High Resolution
HRL	High Resolution Layer
HRSC	High Resolution Stereo Camera
HIS	Human Settlement Index
ID	Identifier
iForest	Isolation Forest
IMD	Imperviousness Density
IRECI	Inverted Red Edge Chlorophyll Index
IRS	Indian Remote-Sensing Satellite
ISA	Impervious Surface Area
iTree	Isolation Trees
IW	Interferometric Wide Swath Mode
JECAM	Joint Experiment for Crop Assessment and Monitoring network
JRC	Joint Research Centre
KC	Knowledge-Based Compositing
LC	Land cover
LCCS	Land Cover Classification System
LiDAR	Light Detection And Ranging
LISS	Linear Imaging Self-Scanning Sensor
LC	Land Cover
LCML	Land Cover Meta-Language
LU	Land Use
LUCAS	Land Use/Cover Area frame statistical Survey
MASD	Mean Absolute Spectral Dynamic
MC	Mean Compositing
MERIS	Medium Resolution Imaging Spectrometer
MGRS	Military Grid Reference System
MMU	Minimum Mapping Unit
MMW	Minimum Mapping Width
MNDWI	Modified Normalized Difference Water Index
MODIS	Moderate Resolution Imaging Spectroradiometer
MTV2	Modified Triangular Vegetation Index II
MVC	Maximum Value Composite
NDBI	Normalized Difference Built-Up Index
NDII	Normalized Difference Infrared Index
NDSVI	Normalized Difference Senescent Vegetation Index
NDVI	Normalized Difference Vegetation Index
NDWI	Normalized Difference Water Index
NGR	Natural Grassland
NIR	Near-InfraRed
NISI	Normalized Impervious Surface Index
NOAA	National Oceanic and Atmospheric Administration
NREVI	Normalized Red-Edge Vegetation Index
NUACI	Normalized Urban Areas Composite Index
NUTS	Nomenclature of territorial units for statistics
OA	Overall Accuracy
OLI	Operational Land Imager
OCSVM	One-Class Support Vector Machine
PSRI	Plant Senescence Reflectance Index
PSU	Primary Sampling Units
QC	Quantile Compositing
RBF	Radial Basis Function
RF	Random Forest

RGR	Red-Green Ratio
ROC	Receiver Operation Characteristic
ROI	Region Of Interest
RTM	Radiative Transfer Models
S-1	Sentinel-1
S-2	Sentinel-2
S-3	Sentinel-3
SAR	Synthetic Aperture Radar
SAVI	Soil Adjusted Vegetation Index
SE	Shannon Entropy
SFS	Structural Features Set
SIGEC	Système intégré de gestion et de contrôles
SMOTE	Synthetic Minority Over-Sampling Technique
SPOT	Satellite Pour l'Observation de la Terre/Satellite for observation of Earth
SRTM	Shuttle Radar Topography Mission
SVDD	Support Vector Data Description
SVM	Support Vector Machine
SWIR	Short Wavelength Infrared
TCD	Tree Cover Density
TCM	Tree Cover Mask
TM	Thematic Mapper
TP	True Positives
USA	United States of America
USGS	United States Geological Survey
V-I-S	Vegetation-Impervious-Soil
VANUI	Vegetation Adjusted Nighttime Light Urban Index
VHR	Very High Resolution
VH	Vertical transmit/Horizontal receive (polarization)
VV	Vertical transmit/Vertical receive (polarization)
VZA	View Zenith Angle
WAC	Weighted Average Compositing
WI	Wetness Index
WP	Work Package

1 Introduction

The Horizon 2020 (H2020) project, "Evolution of Copernicus Land Services based on Sentinel data" (ECoLaSS) addresses the H2020 Work Programme 5 iii. Leadership in Enabling and Industrial technologies - Space, specifically the Topic EO-3-2016: Evolution of Copernicus services. ECoLaSS will be implemented from 2017–2019 and aims at developing innovative methods, algorithms and prototypes to improve and invent future next-generation operational Copernicus Land services from 2020 onwards, for the pan-European and Global Components. ECoLaSS will make full use of dense Sentinel time series of optical (S-2, S-3) and Synthetic Aperture Radar (SAR) data (S-1). Rapidly evolving scientific as well as user requirements will be analysed in support of a future pan-European roll-out of new/improved Copernicus Land Monitoring services, and the transfer to global applications.

This report constitutes a methods compendium for the investigated approaches of the work package (WP) 33 "Time Series Analysis for Thematic Classification" of ECoLaSS Task 3 (Automated High Data Volume Processing Lines).

1.1 Purpose and objectives of the WP

The development of innovative Copernicus Land processing lines in Task 3 is first and foremost targeting the design of approaches for synergistic and integrated utilization of dense time series of high volumes of Sentinel-1/-2/-3 for mapping improved/new LC/LU products, variables and indicators. Therefore, the development work of Task 3 has been grouped into five methodological WP addressing methods development for time series integration, time series pre-processing, and development of methods for analyzing time series with respect to either thematic classification lines or change detection processes.

WP 33 aims to develop a framework for time series analysis for thematic classification based on Sentinel multi-sensor constellation. The objectives of the WP are

- to develop and benchmark optical image compositing methods specifically dedicated to thematic classification: adaptive compositing period, temporal resampling, feature based compositing, alternative time series classification methods over test sites
- to develop time series classification methods for HR layers, crop type and new land cover/land use products

The methods tested and algorithms described in this WP will support the demonstration activities for the development of various prototypes in ECoLaSS Task 4 (Thematic Proof-of-Concept/Prototype on Continental/Global Scale), i.e. High Resolution Layers (HRLs), Grassland, Crop type and new LC/LU products.

The ECoLaSS project follows a two-phased approach of two times 18 months duration. This deliverable comprises the first issue. In the second 18-month project cycle, a second issue of this deliverable will be published, containing all relevant updates concerning the benchmarking of input data for classification as well as the time series classification methods.

1.2 Document structure

After this introduction, the document is organized in three main sections:

- Section 2 provides a review on the inputs needed for classification (reference sampling and image compositing) and on the time series classification methods for HRLs, agriculture and new land cover products;
- Section 3 presents the testing and benchmarking of the candidate methods selected in the review for automated reference sampling, compositing methods and time series classification methods;
- Section 4 gives conclusions and an outlook.

2 Review (theory/state of the art)

The following subchapters describe the state-of-the-art of automated reference sampling (section 2.1), optical image compositing (section 2.2) and time series classification methods (section 2.3).

2.1 Automated reference sampling

The quality of the reference dataset, used for training or labeling, is the key for the accuracy of each classification result. Inappropriate training samples were indeed identified as the main source of errors in many classification processes (Pal et al., 2006). For instance, Foody and Arora (1997) showed that the choice of training samples had a significant effect on the classification results, whereas changing the classifier model (the number of layers in a neural network) was not significant. Nowadays, a lot of ancillary data is available that facilitates sample collection for training data (Gómez et al. 2016), e.g. field crop type data that is provided by European farmers in order to receive subsidies. Also, forest and leave type sample data can be derived from existing land cover maps. Although most land cover classes are relatively persistent over time, the sample quality can still be improved by suitable reference sampling techniques.

On a spatial basis, most approaches try to minimize the amount of outliers by applying a negative buffer before performing the spatial sampling and therefore, to avoid the selection of samples at LC class borders (according to the outdated map) and by excluding very small polygons (Blaes et al., 2005, Radoux et al. 2014, Inglada et al. 2017). For instance, the average per-field reflectance is extracted in Blaes et al. (2005) without the border pixels using a 15-m buffer zone and used for the parcel-based classifications.

Radoux et al. (2014) investigate operational methods for the automated classification of optical images, with the objective to establish that supervised classifiers can be trained from existing thematic maps. Their hypothesis is that the automated extraction of knowledge from existing maps is a sound alternative to the collection of highly reliable training samples from field surveys or from the most recent very high-resolution image interpretation. In order to mitigate the effect of potential errors in those maps, they proposes an approach for cleaning the training datasets by excluding outliers from the distribution of the spectral signatures. The proposed strategy made use of a probabilistic iterative trimming. This method has already been used in remote sensing for change detection (Radoux et al., 2010, Colditz et al., 2012). However, it has rarely been applied for training sample cleaning, which was its initial purpose. Iterative trimming consists of two iterative steps: (i) estimate the distribution of the spectral values within the training sample for a given land cover class and (ii) remove outliers from the sample based on a constant probability threshold. The iteration stops when no more outliers are detected. This study showed that the quality of the classification results based on local training set selection and self-cleaning could automatically yield a more accurate map than the original reference dataset. However, a major drawback of iterative trimming lies in the fact that it operates in a class-specific approach: in the case of a class dominated by mislabeled pixels, well-labeled pixels are consequently considered as outliers (Waldner et al., 2015).

Since outliers are a common problem in many real world datasets, several machine learning algorithms exist to solve the problem. For the problem of cleaning automatically generated training datasets for large area remote sensing classification problems, the algorithms should be efficient for large sample sizes, should work well for high-dimensional datasets and should deal with complex unknown distributions. The Isolation Forest (iForest) is a promising state of the art approach that fulfils all these properties (Liu et al. 2008). The iForest approach directly isolates outliers. This is in contrast to most other outlier detection methods which learn the structure of the normal instances and then identify outliers if they do not fit this structure. The direct outlier isolation takes advantage of two properties of outliers: i) they are less frequent than the normal instances and ii) their feature patterns are different from the normal instances' feature patterns. The iForest is an ensemble of Isolation Trees (iTree) which is a tree structure that isolates such few and different instances. The key isolation characteristic of an iTree is that anomalies are isolated closer

to the root of the tree and normal instances later in the tree. Apart of its properties to be optimal for high-dimensional datasets and large sample sizes, the iForest does not require the features to be scaled and is not very sensitive to parameters leading to overfitting or underfitting (Liu et al. 2008). It can be assumed that, as in the case of the frequently used Random Forest classifier (Breiman 2001), good results can be achieved with default parameters. The latter aspect is particularly important for the outlier detection because, in contrast to the case of a supervised classification task with reliable labels, tuning of parameters would be a non-trivial task.

The One-Class Support Vector Machine (OCSVM) (Schölkopf et al 1999) is a suitable approach for outlier detection with high dimensional datasets and complex non-linear class distributions. The OCSVM fits a maximal margin hyperplane to separate the training instances (all of the same class) from the origin of the feature space. To be able to model non-linear distributions, the kernel trick can be used to map the input data in a higher-dimensional feature space. As a result, the linear separating hyperplane in the higher-dimensional feature space corresponds to a non-linear plane in the input data space. The mapping in the higher-dimensional feature space is performed via a kernel (usually the radial basis function kernel) which has to be defined together with at least one parameters which can be sensitive with respect to the resulting model. Additionally, the OCSVM requires the nu parameter during training which tunes the upper bound of the fraction of outliers in the training dataset (Schölkopf et al 1999). It is worth mentioning that the Support Vector Data Description (SVDD), another frequently used method for outlier detection, is similar to the OCSVM an when used with a Radial Basis Function Kernel gives the same solution than the OCSVM (Tax & Duin 2004).

For imbalanced datasets, datasets for which the classification categories are not approximately equally represented, the Synthetic Minority Over-Sampling Technique (SMOTE) can be applied (Chawla et al., 2002). Often real-world data sets are predominately composed of “normal” examples with only a small percentage of “abnormal” or “interesting” examples. Under-sampling of the majority (normal) class has been proposed as a good means of increasing the sensitivity of a classifier to the minority class. This study shows that a combination of over-sampling the minority class and under-sampling the majority class can achieve better classifier performance than only under-sampling the majority class. It uses a bias to select more samples from one class than from another. This method can be used, for instance, to improve the minor classes accuracy in classification.

2.2 Optical image compositing

A challenge for large scale mapping is to achieve spatial continuity and consistency in the final map. There are two main sources of spatial inconsistencies: heterogeneity in the imagery (different orbits, acquisition dates, cloud/shadow contamination) and within-class spectral variability due to changes in environmental conditions, management decisions and practices (Waldner et al., 2017). To deal with the heterogeneity in the imagery, temporal synthesis of daily optical satellite observation has been applied for years to produce complete, cloud-free images over large areas and to reduce residual cloud contamination. These syntheses are also useful as they can be provided at the same date every year and do not depend on a cloud-free acquisition date. Compositing thus plays an important role in global and regional vegetation monitoring, land cover change analysis, and land cover mapping activities (Vancustem et al., 2009).

In addition, compositing enables a data volume reduction compared to the level 2A products, especially for moderate resolution near-daily coverage sensor data such as AVHRR, MODIS or SPOT-VEGETATION. Compositing of higher spatial but lower temporal resolution satellite data, such as Landsat, is not normally undertaken however because of high data costs and because the land surface state may change in the period required to sense several acquisitions (Hansen et al., 2008). With the five days revisit cycle of Sentinel-2 A/B, it is worth testing capabilities of classical compositing techniques on this high resolution sensor. Despite the lower revisiting frequency compared to medium resolution instruments, a better cloud screening is expected thanks to higher spatial resolution and to the large diversity of spectral bands

including the 1.38 µm band able to detect thin cirrus cloud (Hagolle et al., 2015). Directional effects are also largely reduced with Sentinel-2 thanks to the limited viewing angle of the acquisitions.

Various algorithms have been developed to produce a cloud-free synthesis from optical time series. Each compositing method corrects for angular effects and atmospheric variations differently. Two main categories of compositing are detailed in the following sections: time interval algorithms and feature-based algorithms.

2.2.1 Time interval algorithms

Traditional mapping efforts based on multi-spectral time series are preceded by compositing of spectral bands with image recorded within a relatively short time period.

The most popular compositing algorithm is the Maximum Value Composite (MVC) applied on Normalized Difference Vegetation Index (NDVI) (Holben, 1986). It was firstly created to produce continuous cloud-free images over large areas with Advanced Very High Resolution Radiometer (AVHRR) data to monitor green vegetation dynamics. On a pixel-by-pixel basis, each NDVI value of the compositing period is examined, and only the highest value is retained for each pixel location. The main advantage of this method is to select the date the most likely to be cloud-free among the list of available dates in the compositing period. Indeed, the selection of the maximum NDVI values minimizes clouds, aerosols and water-vapor effects, as well as bidirectional reflectance distribution function (BRDF) effects. In addition, this method does not require heavy computing resources. However, the composited reflectance bands may exhibit substantial radiometric variations, since composite radiances are generally recorded under varying atmospheric and geometric conditions (Cihlar et al., 1997). Particularly, the sensitivity of the NDVI to the sun-target-sensor geometry results in a biased selection of more off-nadir views in the forward scattering direction.

In order to select the best pixel values from the available observation set, various alternative criteria have been proposed and assessed, such as the minimum red value (D'Iorio et al., 1991; de Wasseige et al., 2000; Cabral et al., 2003), the minimum View Zenith Angle (VZA) (Cihlar et al., 1994a), the maximum Normalized Difference Water Index (NDWI) (Gao, 1996), the minimum Short Wave Infrared value (SWIR) (Stibig et al., 2001) used to map land cover in cloudy areas, and the third lowest value of an albedo-like index (Cabral et al., 2003). Some of these criteria reduce the artifacts observed on the MVC composites. However, the selection of a single extreme value, i.e. minimum or maximum, often favours specific atmospheric and geometric conditions, which may cause serious spatial inconsistencies in the composites and in the subsequent processing (Vancutsem et al., 2007a). Moreover, these single value selection criteria use a small part of the available information, even when several observations can be considered as cloud-free.

To avoid the drawback of the best pixel composites, for which the best pixel according to several criteria is selected among the available dates, average syntheses were explored. In these methods, the reflectance value is the average of surface reflectance of cloud-free pixels. The idea is to rely on the repetitivity of observations to statistically reduce errors that could happen due to undetected clouds or cloud shadows or atmospheric correction errors. Some algorithms such as the Best Index Slope Extraction (BISE) (Viovy et al., 1992) and the Average (AVG) (Qi and Kerr, 1995) make a better use of all the cloud free pixels. The BISE method greatly reduces the noise in time series and retains more cloud-free observation than MVC. However, the BISE algorithm requires additional information about the vegetation growth. From a statistical point of view, the AVG algorithm seems to be a more robust approach as it reduces the variability of the signal by averaging the highest 10% of the NDVI values within each compositing period. Nevertheless, the study achieved by Qi and Kerr (1995) could not conclude to any significant improvements compared to the MVC NDVI algorithm. The reason could be the low number of observations selected over some periods, i.e. one or two, because of poor atmospheric conditions, and a very restrictive threshold used in this study.

A more advanced approach to cope with the variability of the sun-target-sensor geometry of high temporal resolution sensors consists of normalizing the bidirectional reflectance by fitting a BRDF model to the available cloud-free observations. The reflectances are then standardized to the nadir view direction and to a specific solar zenith angle considered as representative for the observations. Some algorithms based on inversion of BRDF models have been developed for particular sensors; e.g. the bi-directional compositing (BDC) algorithm applied to SPOT-VEGETATION time series (Duchemin and Maisongrande, 2002). They lead to a great improvement with regards to previous compositing algorithms. Their operational implementation faces however some issues, i.e. the number of cloud-free observations required for the model adjustment.

To deal with the compositing issues of the best pixel composites that favour specific atmospheric and geometric conditions or BRDF normalization that faces implementation issues, a statistically sound alternative strategy called Mean Compositing (MC) (Vancutsem et al. 2007) has been proposed and tested successfully. The MC method treats all cloud-free reflectance values as estimates of the signal, and any remaining variability after cloud removal as an unpredictable noise. It consists of averaging all valid reflectance values for each pixel and each spectral band acquired during the chosen compositing period. Such an approach used under certain conditions reduces the variability induced by the directional effects and the possible remaining atmospheric perturbations after data pre-processing and cloud removal, to produce robust and consistent compositing output. The MC algorithm need to fulfill three conditions to be relevant from a statistical point of view: (i) an efficient quality control procedure able to discard any odd value, (ii) an accurate geometric correction, and (iii) a compositing period which is a multiple of the view zenith angle (VZA) cycle of the instrument.

This method was compared with three existing techniques (NDVI, MVC, AVG, BDC) (Vancutsem et al., 2007a). The results showed that the proposed strategy combined with an efficient quality control produces images with greater spatial consistency than currently available VEGETATION products but produces slightly more uneven time series than the most advanced compositing algorithm. Its performances were also assessed on Medium Resolution Imaging Spectrometer (MERIS) images in Vancustem et al. (2007b) against two other compositing methods: BISE and Cyclope (CYC) (Hagolle et al., 2004) which improves the BDC method. The optimal method was selected thanks to a qualitative examination of the temporal profiles, and a quantitative analysis of the noise introduced into composite images of the reflectance time series. The BISE algorithm is less effective in reducing time series noise than the MC and the CYC. Moreover, this method requires complementary information on the phenology of the region. MC and CYC provide very similar results. Owing to its performance and simplicity, the MC method was selected to process global MERIS time series.

Also using all cloud-free reflectance values acquired during the compositing period, the Weighted Average Compositing (WAC) (Hagolle et al., 2015) may be used to favour dates with low aerosol content, low cloudiness and pixels far from clouds. In order to enhance the fidelity to the central date, and to reduce artifacts due to undetected clouds or shadows, it gives more weight to the images closer to the middle of the compositing period, to the images with a low aerosol content, and to the pixels far from a cloud. However, the weighting must be light enough so that it does not finally select only one date, and finally looks like a best pixel method.

2.2.2 Feature-based algorithms

A more recent strategy of compositing to reduce the spectral variability is to derive temporal or spectral-temporal features from the time series. Compared to the time interval algorithms, feature-based algorithms do not present a fixed and regular compositing period. Spectral-temporal features are composites of the spectral reflectances measured at a specific stage in the season. They summarize events that did not necessarily co-occur in composite images. These composites facilitate the discrimination between classes by reducing the within-class heterogeneity. Drawbacks of spectral-temporal features are

related the amount of available cloud-free images and their quality. Dense time series are required to be able to extract stable spectral signatures at the key moments in the season. Besides, poor cloud/shadow screening results inevitably to noisy features.

The Knowledge-based Compositing (KC) is particularly designed for cropland mapping (Matton et al., 2015; Waldner et al., 2015; Lambert et al., 2016). It aims to extract relevant spectral and temporal features at specific events of the growing season to differentiate the cropland from the other land cover types. These features were defined according to generic characteristics of crop growth. Typically, the crop development cycle can be characterized by four key elements: (i) the growing of crops on bare soil after tillage and sowing; (ii) a higher growing rate than natural vegetation types; (iii) a well-marked peak of green vegetation; and (iv) a fast reduction of green vegetation due to harvest and/or senescence. Based on this conceptual framework, reflectances and Normalized Difference Vegetation Index (NDVI) time-series were analyzed to translate those characteristics into temporal features. Five distinct remote sensing stages in the crop cycle could be defined at the pixel scale (Figure 2-1): (i) the maximum value of red as bare soil has a high reflectance in the red (Tucker, 1979); (ii) the maximum positive slope of the NDVI time series; (iii) the maximum value of NDVI; (iv) the maximum negative slope of the NDVI time series; and (v) the minimum value of NDVI. The final spectral-temporal features corresponded to the reflectance values observed at these stages. These features are time independent, which allowed to deal with the cropland diversity and the agro-climatic gradient across the landscape. This compositing method requires an appropriate temporal distribution of observation, which can compensate for the low frequency of cloud-free observation for cropland mapping.

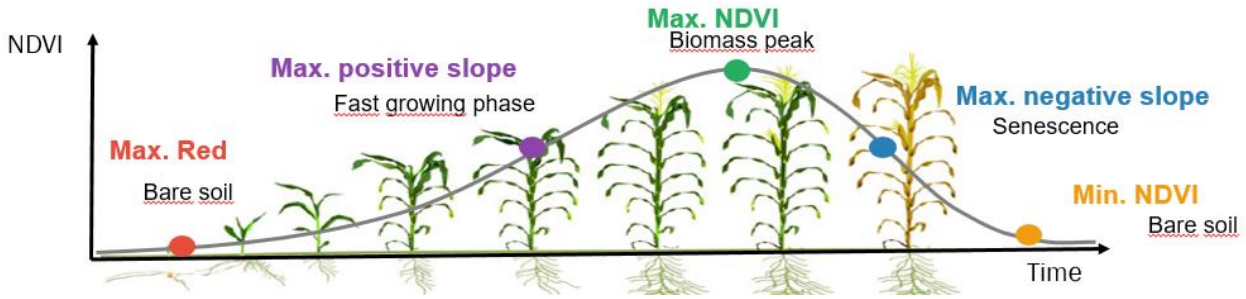


Figure 2-1. Representation of five temporal features of the Knowledge-based Compositing for cropland mapping (minimum NDVI, maximum NDVI, increasing slope, decreasing slope and maximum RED).

The KC method was successfully implemented in Matton et al., 2015 for automated annual cropland mapping along the season for various globally-distributed agrosystems, using high spatial and temporal resolution time series (SPOT-4 take 5 and Landsat-8). The methodology is based on cropland-specific temporal features, which are able to cope with the diversity of agricultural systems. Twenty features (four spectral bands of the five crop growth characteristics) are too numerous as input for most of the classifiers, as this can lead to a performance deterioration. Specific feature combinations were thus selected in order to create a relevant set for differentiating croplands from non-cropland. It was found that the SWIR band did not provide valuable enough information, and it was discarded. The final features were selected as the set of five features providing the best mean overall accuracy (OA) on all of the test sites. This included the red and NIR reflectances from the minimum NDVI stage and the green, red and NIR from the maximum NDVI stage.

In a similar study, Waldner et al. (2015) proposes a fully automated cropland classification method that complies with the requirements of operational agriculture monitoring. It relies on the knowledge of the expected cropland temporal trajectories to determine temporal features to be used in the classification. The overarching idea is to combine both the full discrimination potential proposed by the spectral bands

of a sensor with the synoptic interpretation capabilities of the NDVI. Hence, these features have a straightforward interpretation that consistent throughout the globe even if subjected to local variations. Cropland maps were generated with a support vector machine classifier trained on knowledge-based temporal features derived from SPOT-4 and Landsat-8 time-series and a baseline after a statistical cleaning. For large scale mapping, such features offer also a simple and comprehensive framework to integrate images from different orbits without losing temporal details. Indeed, as the method operates the compositing of the features at the pixel level, it tolerates time-series of different lengths which would increase temporal resolution and consequently the feature extraction.

A third example of KC application is the cropland mapping over Sahelian and Sudanian agrosystems with PROBA-V time series at 100m (Lambert et al., 2016). The methodology uses the five temporal features characterizing crop development throughout the vegetative season to optimize cropland discrimination. A feature importance analysis validates the efficiency of using a diversity of temporal features to complement them according to the cropland proportion. The maximum in red reflectance and the minimum NDVI seem the two most discriminant features in higher crop proportion regions. These features refer to the start of the growing period when differences between cropland and natural vegetation are high due to land preparation. All temporal metrics seem important in one or another crop proportion class without a specific distinction for any one of them. Regardless of the crop proportion classes, as expected, the blue band contributes the least in cropland discrimination due to the impacts of aerosols and atmospheric scattering. The SWIR band is of higher importance than the NIR band, while the red and the SWIR bands are the two most important bands for the classification.

A second way for extracting spectral-temporal features proposes statistical measures from a multi-temporal stack of good quality satellite observations. The advantage of these metrics is the creation of a standard feature space independent of specific time of year or number of input observations (Hansen et al., 2016). These characteristics allow generic models to be built and applied to large areas. Metrics consist primarily of measures derived from all input observations, for example the mean NDVI of all good observations during the study period. Metrics can also be calculated by interval quantile, for example the interquartile mean (mean of all observations between the 25th and 75th quartiles). Alternatively, metrics can be calculated for an individual band as a function of greenness or thermal rankings. For example, red reflectance is low at times of high greenness, and generally high for times of low greenness. A 90–100 interquantile mean of red reflectance ranked by NDVI typically yields a red reflectance value of <5% for forest cover for periods of one year or greater. This method is called in this study the Quantile Compositing (QC).

This QC method was applied to the humid tropical forest biome for a Landsat-based forest disturbance alert (Hansen et al., 2016). Metrics used consisted of individual ranks, means and regressions of red, near infrared, both shortwave infrared bands, as well as ranks of NDVI, near infrared and shortwave infrared (2.2 µm) (NBR), and near-infrared and shortwave infrared (1.65 µm) (NDWI). For this study, example composite metrics include median of first three good observations and median of last three good observations. For the purpose of the forest disturbance alert algorithm, the metrics are used largely as a reference in identifying stable forest pixels within the preceding four-year period.

In Waldner et al. (2017), spectral-temporal features from QC are used for national-scale cropland mapping of South Africa in the absence of within season ground truth data, based on Landsat time series and land cover information. To ensure spatial continuity and consistency in the final map, they reduce the data heterogeneity and spectral variability by deriving spectral-temporal features that capture the salient characteristics of crops. Three spectral-temporal features were derived from all exploitable pixels in the normalized Landsat time series, that is, pixels not affected by clouds, cloud shadows, adjacent clouds and quality flags: (i) the median reflectance value over the three-year time series, (ii) the average reflectance of all pixels belonging to the first decile of stacked NDVI values, (iii) the average reflectance of all pixels belonging to the last decile of stacked NDVI values. There were thus twelve input features for the classification (three temporal features of four spectral bands each). The feature importance analysis

underlined the importance of the SWIR band for crop classification as already reported by Lambert et al. (2016). The importance of the SWIR band ought to be related to a differential leaf water content between crops and natural vegetation (Tucker, 1980), especially in irrigated areas as well as to its specific links with canopy structure and crop residues. From a temporal perspective, three out of the top five spectral-temporal features come from the minimum NDVI which confirms that cropland is most separable when the soil is bare or prepared for sowing (Matton et al., 2015; Waldner et al., 2015).

The availability of 10-m satellite data such as Sentinel-1 and Sentinel-2 provides positive perspectives of improvement to increase the accuracy of the proposed classification scheme, especially in smallholder farming systems where a higher spatial resolution is required (Waldner et al., 2017). A higher density of images along the growing season would also allow to move toward annual cropland mapping, thereby reducing confusions due to land cover and land use change. The red-edge bands available with Sentinel-2 could be instrumental to enhance discrimination with grassland and wetland vegetation.

2.3 Time series classification methods

The following subchapters describe the state-of-the-art of time series classification methods for HRL Imperviousness, HRL Forest, HRL Grassland, Agriculture, and new land cover products.

2.3.1 HRL Imperviousness

Urbanisation is considered as a key driver in global environmental change (Schneider, Friedl and Potere 2010, Weng et al. 2014, Svirejeva-Hopkins, Schellnhuber and Pomaz 2004) and is accompanied by an ongoing consumption of land used for the construction of residential, industrial, and transportation-related areas. Here, continuing soil sealing, meaning the coverage of the soil surface by an impermeable material, leads to irreversible loss of biodiversity, fertile soil and valuable open areas. In this framework, it is of importance to map the extent of built-up areas as well as to derive more detailed information about the spatial distribution and density of impervious surface area (ISA). Currently, data representing the urban extent at global scale has been published, in particular the Global Human Settlement Layer (GHSL) and the Global Urban Footprint (GUF) (Pesaresi et al. 2013, Esch et al. 2017). However, besides providing spatial data on the location of built-up areas, the integration of spatial data on the density of ISA is valuable for a variety of applications, such as environmental monitoring, urban climate modelling, estimation of rainfall runoff in hydrological models, analysis of urban distribution and expansion as well as for population modelling (Yuan and Bauer 2007, Liu et al. 2015a, Zhou et al. 2010, Rodriguez, Andrieu and Morena 2008, Imhoff et al. 2010, Van de Voorde, Jacquet and Canters 2011).

In this connection, there are studies investigating the derivation of ISA by means of Earth observation data. Lu et al. (2013) present methods that are applied in the field of ISA mapping, including pixel-based, object-based, sub-pixel-based, spectral mixture analysis-based, regression-based, and threshold-based methods.

There are a number of studies which employed the Vegetation-Impervious-Soil (V-I-S) model to estimate ISA. This model analyses urban land cover composition and links the three components to spectral characteristics of remote sensing imagery (Ridd 2007). Lately, it was applied in the context of ISA estimation for a study area in India using Landsat imagery (Sarkar Chaudhuri, Singh and Rai 2017). In this study, single Landsat acquisitions for the years 2001, 2007, and 2014 were used as input. First, an exclusion of water surfaces was conducted by means of the Normalized Difference Water Index (NDWI). Afterwards, a minimum noise fraction transformation was applied to the calibrated spectral bands to determine the dimensionality of the image and to generate the eigenvalues and eigenimages. Based on these data, end members corresponding to vegetation, high and low albedo, as well as soil and impervious surfaces are identified and used for linear spectral mixing analysis to retrieve ISA.
Considering ISA estimation at larger extents, most of the studies use nighttime light data from the Defense Meteorological Satellite Program's Operational Line-scan System (DMSP-OLS) in combination with

multispectral imagery from the moderate resolution imaging spectroradiometer (MODIS) data. In this context, Guo, Lu and Kuang (2017) presented a new index called Normalized Impervious Surface Index (NISI), which is based on the combination of DMSP-OLS and MODIS NDVI data to overcome known issues of nighttime light data, such as saturation and blooming effects. Here, a maximum value composite was used for MODIS NDVI data including spectral information from 247 scenes. The study area includes several cities in China. At first, a training dataset is generated using Landsat-8 data, where a simple masking and clustering approach is performed to retrieve an ISA map. This ISA map is then resampled to a spatial resolution of 250 m to meet the spatial resolution of MODIS data. Next, the features NISI, Human Settlement Index (HSI), and Vegetation Adjusted Nighttime Light Urban Index (VANUI) are calculated. For ISA estimation a support vector regression model is employed using the Landsat-based training data and the calculated indices. Furthermore, Elvidge et al. (2007) computed an ISA map at global scale. To this end, they used Landsat-based ISA estimations to calibrate a linear regression model. As input to regression DMSP-OLS nighttime light data were used together with LandScan population data. The resulting data was the first global ISA map at a spatial resolution of 1 km. Moreover, Liu et al. (2015b) proposed a new index called Normalized Urban Areas Composite Index (NUACI). This index is calculated using DMSP-OLS nighttime light data along with MODIS Enhanced Vegetation Index (EVI) as well as MODIS NDWI data and is designed to overcome the limitations of nighttime light data. In this study, the MODIS 16-day composite was used for a period of one year to generate a maximum value composite for the EVI index. Comparable to other studies, a regression model was then applied to predict ISA for selected study areas.

Further studies derived ISA by means of spectral mixture analysis and the application of regression models using multispectral imagery at local to regional scale (Bauer, Loffelholz and Wilson 2007, Esch et al. 2009, Kaspersen, Fensholt and Drews 2015, Braun 2004). Bauer et al. (2007) used single acquisition Landsat images for the years 1990 and 2000 covering a study area in the United States. Training areas were manually digitised and an orthophoto at a spatial resolution of 1 m was used to determine ISA for the selected sites. Next, a tasselled cap transformation was applied on Landsat images and the greenness values were used as input for the regression model. A land cover classification was used to limit the region of interest to built-up areas. In a further study, ISA was modelled for a number of states in Germany using optical Landsat imagery. Here, an infrared aerial image at a spatial resolution of 40 cm was used for training and validation. To this aim, impervious surfaces were classified using a threshold-based approach and in a following step reference data (vector format) including land cover information were used to minimise classification errors. The derived impervious surface data at 40 cm resolution was aggregated to a grid size of 30 m (corresponding to the resolution of Landsat) to obtain an ISA map. This map was then employed to calibrate a support vector regression model. Afterwards, the calibrated model was applied on Landsat NDVI to retrieve ISA for the entire region of interest (Esch et al. 2009). Kaspersen et al. (2015) studied the usability of Landsat-based vegetation indices to estimate ISA for selected European cities. In particular, NDVI, Soil Adjusted Vegetation Index (SAVI) and fractional vegetation cover (FR) were used. Regression analyses were performed to predict ISA followed by an inverse calibration, using slope and intercept of predicted and observed (based on high resolution imagery) ISA, to minimise overestimation. Another study integrated a larger feature space including a vegetation index, all spectral bands, phenological information, and texture features to estimate ISA for study sites in the United States and China (Liu, Luo and Yao 2017). At this, the spectral bands are included from Landsat sensors, phenological features are extracted from combinational use of Landsat and MODIS data, and texture features obtained from gray level co-occurrence matrix approach are based on Landsat bands. Afterwards, feature selection, using the variable importance tool of the random forests algorithm is applied on the feature space to obtain a subset of features providing the highest discrimination. Then a subpixel mapping was conducted using a high resolution ISA map as training to model an ISA map. Moreover, Tsutsumida et al. (2016) monitored ISA development over 13 years for the study area of Jakarta (Indonesia) using MODIS EVI data. Training data was extracted from very high resolution images at Google Earth. For classification purposes, a sub-pixel random forests algorithm was applied. The results include annual ISA maps.

2.3.2 HRL Forest

In the following subchapters, first the production of the HRL Forest will be described, including the product definitions and methodology. Afterwards the state of the art, gives a general overview of the currently available forest maps on different regional scales.

2.3.2.1 HRL Forest production

The HRL Forest represents one out of five thematic layers of the Pan-European component coordinated by the EEA and is part of the Copernicus Land Monitoring Service (CLMS). It aims at mapping the status of tree-covered areas and its associated dominant leaf type at pan-European scale (EEA-39 member states) and in 20m spatial resolution using optical Earth Observation (EO) data for certain reference years in a 3-years update cycle.

It has been firstly produced in the frame of the GMES Initial Operations (GIO) phase 2011-2014 for the reference year 2012 (± 1 year) in 5 geographically splitted lots by different implementing European consortia, and with an involvement of EEA member states in a dedicated verification and enhancement phase. HRL Forest products 2012 have been produced based on mono-temporal High Resolution (HR) EO data coverages (HR_IMAGE_2012 with two pan-European coverages) provided by the ESA Data Warehouse (DWH), and in national projections. Additional EO data from other sources (e.g. Landsat 8 USGS) has been approved for gap-filling purposes only. Finally, national products have been re-projected and mosaicked to European lot-mosaics to serve two different service elements (service element 1 for EEA and service element 2 for JRC) with different specifications. This overall concept, together with considerable constraints of the data situation at that time (including a compounding access to national in-situ data), has led to significant differences in the product's specific patterns and thematic quality between the geographical lots. Due to several timely delays from production side and involvement of member states, the overall production time of the HRL Forest 2012 has exceeded the contractually specified 3 years in the end.

The second implementation of the HRL Forest for the reference year 2015 (± 1 year) has strongly benefitted from the lessons learned of the previous GIO phase, but has also made considerably higher requirements regards thematic accuracy and production time. The most important changes compared to 2012 are:

- production fully in European projection
- no split in geographical lots and service elements
- no country involvement through a separate verification and enhancement phase
- a generally increased product portfolio with additional change products, and corrected 2012 products to allow a full harmonisation across Europe
- a simpler workflow, implemented by an EIONET "production portal"
- an envisaged production time of 12 months (compared to 36 months in 2012)

However, the most noticeable change has been achieved by a drastically increased EO data situation. With the successful launch and operation of Sentinel-2A in 2015, the Copernicus community has got access to dense time-series data in an unprecedented detail and manner for the very first time. Together with the possibility to integrate freely available Landsat data for certain reference years, a completely new basis has been made available. Even the latest HR IMAGE dataset from ESA (HR_IMAGE_2015), representing one of the input datasets for the HRLs, has undergone a positive evolution with revised acquisition windows and a restriction to two primary satellites (ResourceSat-2 and SPOT-5), sharing almost the same radiometric characteristics.

These points paved the way from mono-temporal analysis and a single scene classification to multi-temporal analyses and time series classifications. Additionally, an improved access to national in-situ data

and ancillary datasets could be ensured through the Copernicus Reference Data Access (CORDA), further contributing to the production of consistent and harmonised HR Forest layers.

For the next implementation of the HRL Forest for the reference year 2018 in 10m spatial resolution, a further increasing data situation thanks to the availability of Sentinel-2B and the efforts of the Copernicus In-situ component has to be expected.

HRL FOREST PRODUCT DEFINITIONS

In the following, a brief overview on the HRL Forest product definitions will be given. A detailed HRL Forest Product Specifications Document will be made available for download at the CLMS website under https://land.copernicus.eu/user-corner/technical-library as soon as the HRL products are published.

Table 2-1 provides an overview of the Land Cover (LC) and Land Use (LU) features to be included/excluded in the tree cover mapping (if detectable from the 20m input satellite data), resulting in a binary Tree Cover Mask (TCM), being the baseline for the two 20m primary status layers Tree Cover Density (TCD) and Dominant Leaf Type (DLT). This mask is subsequently filled with the relevant leaf type information (broadleaved/coniferous) and tree cover density values. Both pixel-based layers represent the primary products from which all other layers (including change products) will be derived.

Table 2-1: LC/LU features to be included/excluded from the tree cover mask

Included Features (*if detectable from the 20m imagery)*	**Excluded Features** (*if detectable from the 20m imagery)*
• *Evergreen/deciduous broadleaved, sclerophyllous and coniferous trees of any use* • *Forests (grown-up and under development)* • *Orchards, olive groves, fruit and other tree plantations, agro-forestry areas* • *Transitional woodland, forests in regeneration* • *Groups of trees within urban areas (alleys, wooded parks and gardens)* • *Forest management/use features inside forests (forest roads, firebreaks, thinnings, forest nurseries, etc.) - if tree cover can be detected from the 20m imagery* • *Forest damage features inside forests (partially burnt areas, storm damages, insect-infested damages, etc.) - if tree cover can be detected from the 20m imagery*	• *Open areas within forests (roads, permanently open vegetated areas, clear cuts, fully burnt areas, other severe forest damage areas, etc.)* • *Dwarf shrub-covered areas, such as moors and heathland* • *Vineyards* • *Dwarf pine / green alder in alpine areas* • *Mediterranean shrublands (macchia, garrigue etc.)* • *Shrubland*

Tree Cover Density

The Copernicus HRL Forest defines Tree Cover Density as the „vertical projection of tree crowns to a horizontal earth's surface" and provides information on the proportional crown coverage per pixel. It is assessed by means of Very High Resolution (VHR) satellite data and/or aerial ortho-imagery and shows a natural sensitivity towards phenology and radiometric influences (e.g. haze). The Tree Cover Density represents a primary status layer derived from monotemporal satellite data and has the following main specifications:

- 20m spatial resolution
- Tree Cover Density range of 0-100%
- No Minimum Mapping Unit (MMU); pixel-based
- Minimum Mapping Width (MMW) of 20m

Dominant Leaf Type

The Dominant Leaf Type is another primary status layer of the HRL Forest, derived from multitemporal satellite image data and has the following main specifications:

- 20m spatial resolution
- Fully identical in its outline extent with the Tree Cover Density product
- Providing information on the dominant leaf type: broadleaved or coniferous
- No Minimum Mapping Unit (MMU); pixel-based
- Minimum Mapping Width (MMW) of 20m

Forest Type

The 20m Forest Type is produced by applying a minimum „Forest" definition, largely following the forest definition of the Food and Agriculture Organization (FAO), accessible under www.fao.org/docrep/006/ad665e/ad665e06.htm.

Tree cover in traditional agroforestry systems such as Dehesa/Montado is explicitly included for EEA purposes. The product is derived through a spatial intersection of the two primary status layers Tree Cover Density and Dominant Leaf Type and has the following main specifications:

- 20m spatial resolution
- Tree Cover Density range of ≥10-100%
- Minimum Mapping Unit (MMU) of 0.52 ha (13 pixels); applicable both for tree-covered areas and for non-tree-covered areas in a 4x4 pixel connectivity mode, but not for the distinction of dominant leaf type within the tree-covered area for which no such minimum is set.
- Minimum Mapping Width (MMW) of 20m

METHODOLOGY

Besides the pre-processing of the satellite data, the selection process of suitable data (EO data, ancillary data) formed a fundamental step in the production process. According to specific selection criteria (i.e. cloud/haze cover, acquisition dates) the best available satellite scenes have been selected and subsequently processed. Since Sentinel-2A represented the main input data source, the Military Grid Reference System (MGRS) has been defined as production unit system. The HRL Forest 2015 used a multi-temporal and multi-sensor approach for creation of the TCM and DLT.

- Multi-temporal in this context means a time series of classifications using EO data of the specified reference year 2015 +/-1 year. However, the largest part of satellite data is from 2016 (~82%).
- Multi-sensor implies the use of several optical sensors in order to fill data gaps and to increase the number of data coverages per MGRS tile, namely Sentinel-2A, Landsat 8 OLI, ResourceSat-2 and SPOT-5.

On average, about 18 multi-temporal scene coverages (Sentinel-2A, Landsat 8, see Figure 2-2) have been used for the per-pixel analysis per MGRS tile. An initial land cover classification has been performed for each MGRS tile using Support Vector Machines (SVM). Subsequently, a rule-based approach has been applied to generate the dominant leaf type and to derive a pre-final TCM. The latter one has undergone several revisions, including manual enhancement steps and plausibility analyses using existing Copernicus data (CLC 2012 and other thematic HRLs).

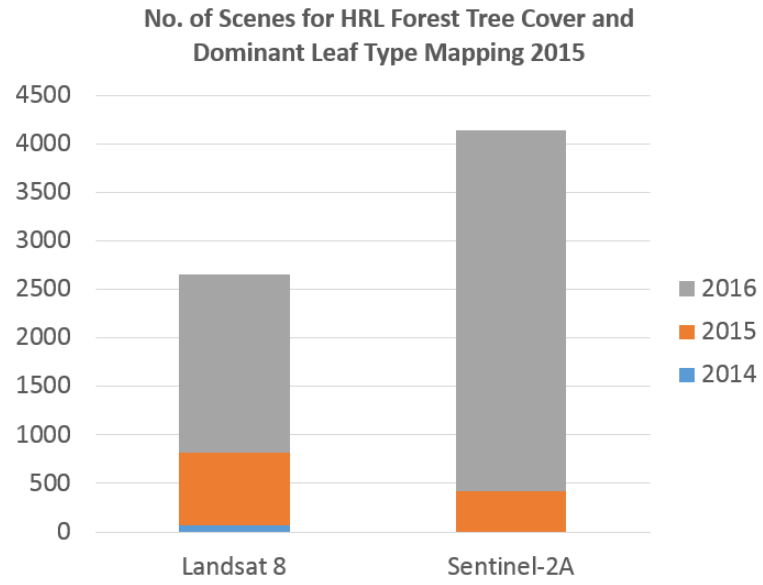
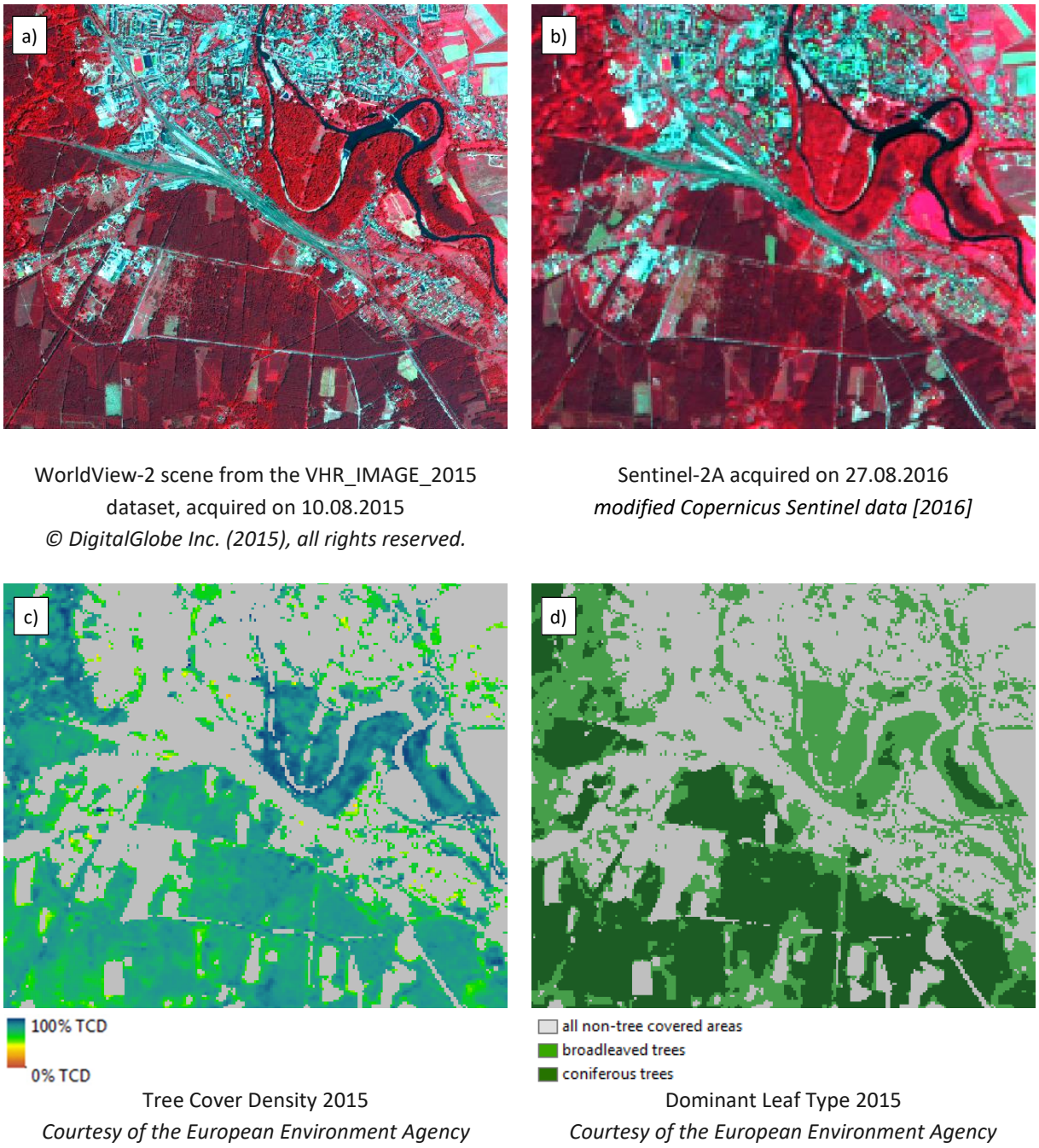


Figure 2-2. Number of Scenes for HRL Forest Tree Cover and Dominant Leaf Type Mapping 2015.

Contrary to the DLT 2015, the status of the TCD 2015 has been derived by classifying nearly 1,000 single satellite images (Sentinel-2, Landsat 8, ResourceSat-2, SPOT-5) from the 2015 reference year (± 1 year) on a mono-temporal basis, but within the confines of the multi-temporally derived Tree Cover Mask. Tree Cover density values have been calculated using a multiple linear regression estimator, fed by more than 150,000 automatically collected reference samples. In order to magnify the accuracy of the TCD product, more than 500,000 reference points have been interpreted visually based on existing VHR_IMAGE_2015 data as well as suitable ortho-imagery and subsequently integrated in the classification process.



WorldView-2 scene from the VHR_IMAGE_2015 dataset, acquired on 10.08.2015
© DigitalGlobe Inc. (2015), all rights reserved.

Sentinel-2A acquired on 27.08.2016
modified Copernicus Sentinel data [2016]

Tree Cover Density 2015
Courtesy of the European Environment Agency

Dominant Leaf Type 2015
Courtesy of the European Environment Agency

Figure 2-3. Example of used input data and resulting 20m products for a region in western Poland. a) VHR_IMAGE_2015, b) Sentinel-2A, c) TCD 2015, d) DLT 2015.

Figure 2-3 shows the outcome of both, the TCD 2015 (Figure 2-3c) and the DLT 2015 (Figure 2-3d) classification based on modified Copernicus Sentinel data (Figure 2-3b). Compared to the VHR_IMAGE_2015 (Figure 2-3a) the DLT distinguishes well between broadleaved and coniferous trees. The parts of lower and higher density visible in Figure 2-3a are well represented by the TCD in Figure 2-3c.

The pixel-based primary status layers Tree Cover Density and Dominant Leaf Type have been validated through a systematic stratified random sampling approach with more than 9,500 Primary Sampling Units (PSU) and exceeded an overall thematic accuracy of 90%.

2.3.2.2 Forest state of the art

The estimation of forest gain/loss is of great interest under different aspects (e.g. forest policy, nature protection, climate change, REDD) as the impacts are quite extensive. Since it is much more efficient to use remote sensing data for the analysis of changes in the forest compared to field studies, there are various (research) projects dealing with this subject in different regions and with different data and methods. Over the years much effort has been undertaken to develop and improve new and existing techniques, i.e. improved classification approaches considering a general increase in spatial and temporal resolution as well as the accuracy assessment of the applied methods (Hansen et al., 2003; Hansen et al., 2000; Kulkarni and Lowe, 2016; Healey et al., 2005; Zhu and Woodcock, 2014; Coppin and Bauer, 1996).

Regarding the methods for classification of forest the methodologies differ from each other. On a medium to high resolution scale the currently used methods based on remotely sensed data can be generally divided into two categories: a mono-temporal (potentially followed by post-classification time series analysis) and a multi-temporal approach (pre-classification time-series analysis) (Hirschmugl et al., 2017; Mitchell et al., 2017; Miettinen et al., 2014). The mono-temporal approach in this context describes the classification of each relevant image from the data stack, like it has been applied in the production of the HRL Forest 2015 and the TCD 2015, respectively, followed by integrating various classifications applying a rule-based approach (see chapter 2.3.2.1).

Multi-temporal approaches include specific analyses before the classification is carried out. The Best Available Pixel (BPA) approach for example comprises a per-pixel evaluation of certain parameters, which is applied to a stack of EO data. In a next step the resulting composite can be classified. Another multi-temporal way to prepare the EO data for the classification is the per-pixel derivation of specific metrics. In a first step an index like the NDVI for example is calculated for a certain amount of EO data. Afterwards, statistics of the data stack are calculated on a per-pixel level, e.g. mean, minimum, or maximum, which results in the final composite that the classification is based on. Multi-temporal approaches have been applied for forest classification by various authors (Enßle et al., 2016; Hansen et al., 2013; Kempeneers et al., 2011; Zhu, 2017).

Until now, most of the forest mapping products are based on optical Landsat data (e.g. Hansen et al., 2013; Potapov et al., 2015; Potapov et al., 2008; Cohen et al. 2002; Zhu and Woodcock, 2014; Healey et al., 2005). On a global scale Hansen et al. (2013) developed a forest map using Landsat 4, 5, 7, and 8, showing the canopy cover percentage. In their study all global land except for Antarctica and some Arctic islands was included. Furthermore trees were defined as all vegetation taller than 5m. To derive the canopy density percentage several deployed per-band metrics (reflectance values, mean reflectance values, and the slope of linear regression of band-reflectance values versus image date) were analysed per band (Hansen et al., 2013).

Another global mapping product showing the forest cover is the Global PALSAR-2/PALSAR/JERS-1 Forest/Non-Forest map based on SAR data, published by the Japan Aerospace Exploration Agency (JAXA). This binary map is based on the Global PALSAR-2/PALSAR/JERS-1 mosaic, which is composed of PALSAR/PALSAR-2-data. The per-pixel classification is based on the backscattering coefficients which are used to detect forests (Shimada et al., 2014).

On pan-European scale the Joint Research Center (JRC) published a forest cover map for the year 2006 with a spatial resolution of 25m. It is based on the data fusion and classification approach by Kempeneers et al. (2011), which works in two steps: first, the selected EO Data (in this case IRS-P6, Spot-4, and Spot-5 data with a spatial resolution of 25m) are classified into a generalized Land Cover (LC) map. For the classification, training data based on the Corine Land Cover (CLC) map (1990-2006) are used. In a second step, the generalized LC map is combined with a multi-temporal composite of coarse resolution MODIS data (250m) and thereby refined, so that the former classes now have several subclasses. By using this

method it is possible to keep the high spatial resolution of the EO Data although data with coarser resolution are used to refine the product (Kempeneers et al., 2011).

On a regional scale Potapov et al. (2015) focus on the former Eastern bloc countries and analysed the Landsat archive from 1985-2012. Forest loss is monitored annually whereas forest gain is estimated on a decadal scale, due to only marginal changes from year to year. Similar to the global product of Hansen et al. (2013), the methodology included a per-pixel quality assessment and the application of metrics derived from specific time-spans. The refined and extended methodology led to a significantly higher accuracy of the product than the global product (Potapov et al. 2015; Hansen et al. 2013).

In the tropical forests the research effort is also quite high and most of the studies aim at monitoring deforestation (e.g. Roy et al. 2002; Foody et al. 2003; Hansen et al. 2008; Miettinen et al. 2014; Achard et al., 2002; Fuller 2006). Achard et al. (2002) for example created a deforestation map of all tropical countries except Mexico by analysing the Landsat archive from 1990 to 2010. The forest cover is derived from the satellite data with the help of sampling units (10x10 km size). Afterwards, the land cover is classified into five categories regarding their tree cover by a supervised classification (Achard et al., 2002). Another project that deals with tropical forest mapping/monitoring and the development of the capabilities of EO based land monitoring is the EOMonDis project (https://eomondis.info/). Different multi-temporal techniques are used to estimate the forest cover in Cameroon, Malawi, Gabon and Peru: Among them, a multi-temporal classification is applied to create a Land Cover Map based on Landsat 8 and Sentinel-2 data. Besides, time-features are analyzed to monitor forest disturbances (Enßle et al., 2016).

Various approaches for forest mapping and monitoring are existing and used at the moment. A crucial factor which is influencing the forest mapping is not only to be found on technical side, but also in the definition of a forest itself. The ambiguity of classification systems with differences in the conceptualisation of forest around the world has been emphasized by Comber et al. (2005), just by taking the two physical characteristics tree height and crown canopy cover as a minimum requirement into account.

One of the most promising aspects for future improvements in the land cover mapping domain is in the increasing spatial and temporal resolution of EO data. Nearly all of the currently applied methods refer to high and medium resolution data. Currently available products based on Landsat for example have a spatial resolution of 25m maximum. The HRL Forest 2015 (based on Sentinel-2) have been produced with 20m resolution, but the upcoming HRL Forest 2018 is planned to be produced in 10m resolution. Therefore, the goal is to improve and further develop the existing approaches as well as the development of new ones for the higher resolution in order to fully use the additional information. Furthermore, a high temporal resolution is quite important to develop classification products with maximum accuracy (Mitchell et al., 2017; Hansen et al., 2013). On behalf of data availability it can be stated that the situation is quite good. Besides the Landsat archive the growing stack of Sentinel data in context of the Copernicus programme contributes to a broad range of medium to high resolution data which is constantly expanded.

2.3.3 HRL Grassland

The HRL Grassland chapter contains information about the HRL Grassland production, as well as the state of the art, concerning different technical approaches. Afterwards a desk study about the mapping of Mediterranean grassland, elucidates the challenges in this context and gives advice how to deal with it in future.

2.3.3.1 HRL Grassland production

The HRL Grassland 2015, comprising natural, semi-natural and managed grasslands of the EEA39 countries is one of five High Resolution Layers (HRL) on land cover characteristics within the context of Copernicus Land Cover Services (notably imperviousness surfaces, forest areas, natural and semi-natural grasslands, wetness and water, small woody features), commissioned by the European Environment Agencies EEA. It

is a binary product with 20m spatial resolution and a minimum mapping unit of 1ha that aims at providing a synoptic view on the distribution and expansion of the pan-European grasslands.

In answer to the technical constraints of the HRL Natural Grassland (NGR) of the reference year 2012, which has not met the common expectations nor the accuracy requirements, the methodology for the HRL Grassland product of 2015 has been fundamentally reconsidered and comes now with a revolutionized approach concerning definition, workflow and technical aspects, as well as an improved data base.

HRL GRASSLAND PRODUCT DEFINITION

The HRL Grassland 2015 is accompanied by both, a scientifically sound and solid definition about the diversity of grassland types and various typical grassland landscapes that have to be part of the grassland product, as well as a distinct declaration about what has to be excluded. Grassland within the context of this product represents herbaceous vegetation with at least 30% ground cover and with at least 30% graminoid species such as Poaceae, Cyperaceae and Juncaceae. Additional non woody plants such as lichens, mosses and ferns can be tolerated.

Table 2-2. Definition of Grassland according to the HRL Grassland 2015

ELEMENTS TO BE INCLUDED IN THE GRASSLAND PRODUCT	ELEMENTS TO BE EXCLUDED FROM THE GRASSLAND PRODUCT
▪ Natural, semi-natural, agricultural / managed grass-covered surfaces ▪ Grasslands with scattered trees and shrubs covering a maximum 10% ▪ Heathland with high grass cover, maximum of 10% non-grass cover ▪ Coastal grasslands, such as grey dunes and salt meadows located in intertidal flat areas with at least 30% graminoid species of vegetation cover ▪ Sparsely vegetated grasslands (>30% vegetation cover – cf. comment below) ▪ Grasslands in urban areas: parks, urban green spaces in residential and industrial areas ▪ Semi-arid steppes with scattered Artemisia scrub ▪ Meadows: grassland which is not regularly grazed by domestic livestock, but rather allowed to grow unchecked in order to produce hay ▪ Grasslands in urban areas: sport fields, golf courses ▪ Grasslands on land without use ▪ Natural grasslands on military sites	▪ Peat forming ecosystems dominated by sedges ▪ Reed beds and helophytes dominated systems ▪ Tall forbs, fern, shrub dominated vegetation ▪ Grasslands that have been observed as tilled (in the reference year or a certain period before, in that case they are considered as arable fields) ▪ Rice fields ▪ Vineyards, orchards, olive groves, (if more than 10% shrubs or trees) ▪ Tundra dominated by shrubs and lichens ▪ Grassland on fresh (and older) clear-cuts in the woods

The rate of 30% ground cover density shall be understood as a benchmark implicating that grasslands with ≥30% ground cover can usually be distinguished very clearly from bare ground on EO data with the resolution of 20m. According to this reference, the classification of grasslands focusses on “dense grasslands” that can be identified with high accuracy.

METHODOLOGY

The mapping of grasslands – and of vegetation in general - bases on the detection of canopy density, chlorophyll status and expansion of the vegetation cover during the growing season. It works best at those times of the year where plants show high photosynthetic activity. Due to this fact and in order to get a reliable data basis for the classification, the HRL Grassland 2015 uses a multi-seasonal, multi-temporal and multi-sensor approach.

- *Multi-seasonal* describes the use of EO data from different seasons concerning those periods of the year where grassland could be identified best and – taking account of agricultural management schemes as well as grassland mowing cycles - at the same time be well differentiated from croplands and bare grounds.
- *Multi-temporal* in this context means a time series of classifications using EO data from 1 up to 3 years for the reference period depending on the availability of suitable data (regarding cloud cover, covering of the area, etc.). Where necessary, EO data 2015+/-1, meaning data from the years 2014, 2015 and 2016, build the baseline for the reference year 2015. Data from preceding years cover the historic time period. However, the largest part of satellite data is from 2016 (~71%). The temporal series include images from 2015 (~18%), 2014 (~10%) and 2013 (~0,5%), respectively.
- *Multi-sensor* implies the use of several sensors to fill the gap in suitable data and to complement the advantages of optical data, namely Sentinel-2A (~59%), Landsat 8 OLI (~41 %), Landsat 7 ETM+, Landsat 5 and IRS-P6 with the benefits of SAR data from Sentinel-1.

The HRL Grassland 2015 is the result of an elaborate workflow, pursuing both, the accurate identification of grassland and at the same time the exclusion of distinct non-grassland areas.

All selected optical EO data (Sentinel-2A and Landsat 8 for the reference period, all sensors for the historic period) were used for a multi-scale and multi-sensor segmentation. These image segments, together with training samples of the main land cover classes, provided the basis for subsequent iterative supervised object-based classification of dense time series of both, optical and SAR data (Sentinel-1) with the support vector machine algorithm. Pursuing a strategy of exclusion, additional layers such as vegetation indices basing on Sentinel-2A and Landsat 8 enable the identification of tilled or harvested cropland and helped to exclude non-grassland areas. Potential overlaps were reduced by using thresholds from the HRL 2012/2015 concerning Imperviousness, Tree Cover Density and Permanent Water Bodies. The resulting intermediate scene-based grassland masks (each individually weighted reflecting their relevance within the classification) were then combined with a single SAR-based classification layer. The rule-based evaluation of the results of the optical classification in combination with those of the SAR classification allowed a further enhancement of the reliability and accuracy of the final grassland layer by ways of excluding critical non-grassland land cover that could not adequately be captured by optical classification, such as horticulture or vineyards.

Verified through a systematic stratified random sampling, the filtered and harmonized final HR GRA 2015 product proves an overall thematic accuracy of over 85%.

The GRA 2015 mask been derived by classifying nearly 4225 single satellite images (Sentinel-2, Landsat 8 OLI) from the 2015 reference year. In Figure 2-4, the total amount of images used for the production of the GRA 2015 mask, is displayed per year and satellite.

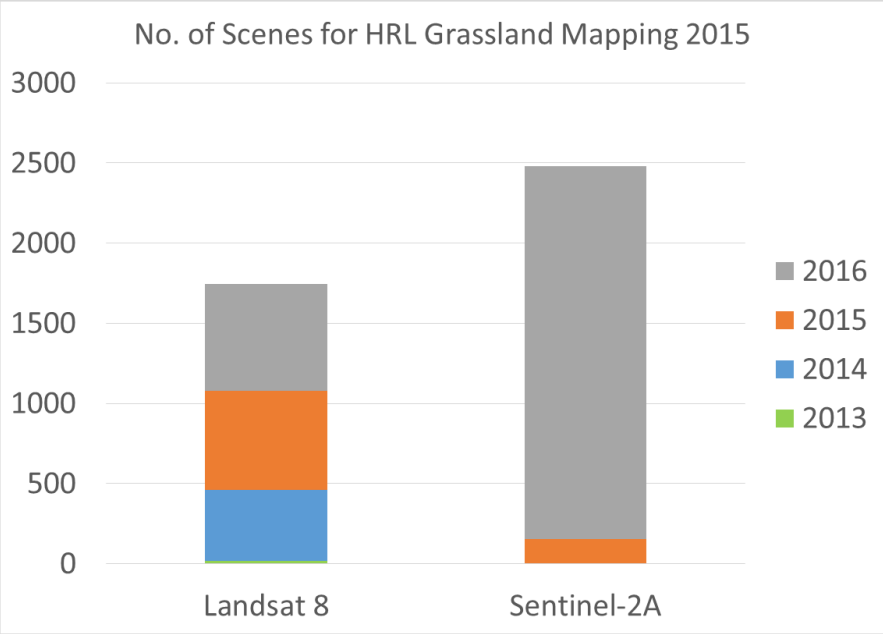


Figure 2-4. Summary of the total amount of images used for the production of the GRA 2015 mask, per year and satellite.

Additional products

The *Ploughing Indicator PLOUGH* indicates the time period (in years) since the last ploughing activity has taken place, respectively when grassland has been converted into cropland. For those countries with differing tilling regulations the PLOUGH then provides additional information on potential grassland areas.

Whereas the grassland layer derives from EO data of the reference year 2015+/-1, the ploughing indicator relates up to 6 preceding years, identifying those areas which have been tilled within this period of time. It highly depends on the availability of suitable historical data. The final HRL Grassland 2015 implies only the non-tilled areas.

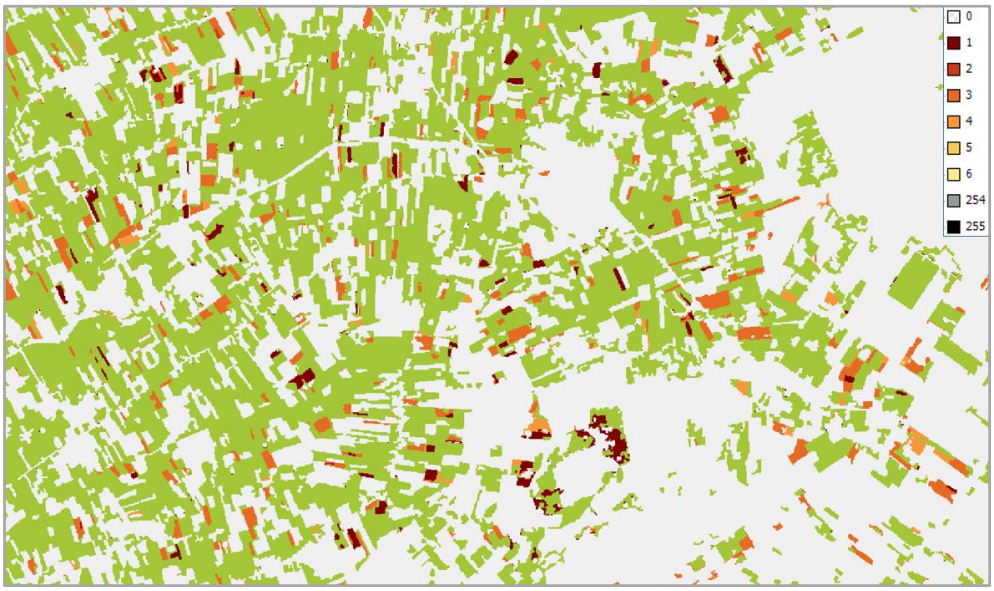


Figure 2-5. Final Grassland layer in Central Europe (green) and PLOUGH, indicating the number of years since the last ploughing activity in orange/red shades.

The *Grass Vegetation Probability Index GRAVPI* indicates the degree of reliability of the multi-seasonal optical grassland classification for the reference year of 2015 (EO data from plus/minus 1 year). It represents the number of scenes the optical classification bases on as percentage values. A high number of adequate imagery improves the accuracy and reliability of the final classification (indicated in bluish shades). Due to the variability of the data base (caused by limitation through atmospheric disturbances, cloud cover or technical constraints), GRAVPI values may differ in neighboring working units.

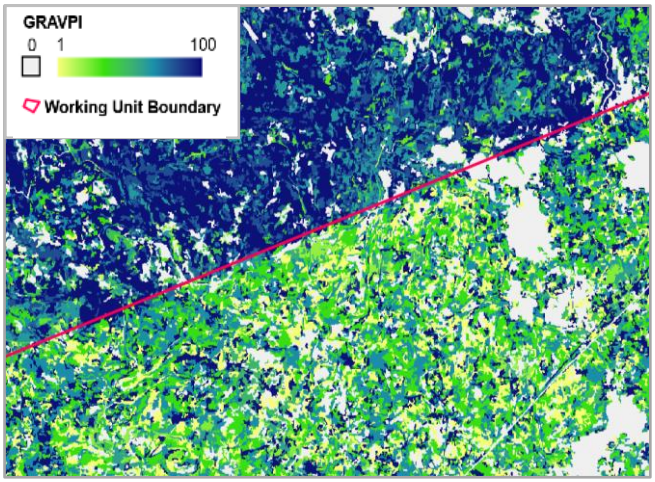


Figure 2-6. Example of GRAVPI from Turkey. The upper Working Unit (WU) provides a high number of adequate scenes for classification and thus a better data base than the WU below. The GRAVPI above consequently shows significantly higher percentages.

Like the main product HRL GRA 2015, PLOUGH and GRAVPI provide a spatial resolution of 20m and a MMU of 1ha.

2.3.3.2 Grassland state of the art

In the last decade, remote sensing technologies for the monitoring of diverse vegetation types and natural habitats as well as their biodiversity have been significantly advanced (Turner et al., 2003; Vanden Borre et al., 2011; Corbane et al., 2015). The increasing availability of high and very high spatial resolution multi-temporal data from multi-spectral and hyperspectral optical satellite sensors (e.g., RapidEye, Sentinel-2, the planned EnMAP), as well as from Radar (TerraSAR-X, Radarsat-2, Sentinel-1) and Light Detection And Ranging (LiDAR) sensors have boosted this development. Specifically, this progressively supports the transition from traditional statistical classification approaches to more effective machine learning algorithms (due to the increasing computational capabilities necessary for processing big amount of data) and is continuously fostering the development of newer more advanced methodologies (Waske and van der Linden, 2008). Still, present habitat mapping programs (e.g., CORINE land cover or the NATURA 2000 Annex II habitat maps) mainly account for visual image interpretation or field surveys, which are high time and cost demanding, but also strongly depend on knowledge and experience of the operator (Mander et al., 2005; Gross et al., 2009; Thoonen et al., 2010). Deriving key information for the assessment of biodiversity is highly supported through the mapping of grassland species and characterizing related parameters or indices, e.g., primary productivity, climate or habitat structure (Turner et al., 2003). In this framework, several methods for the monitoring of grasslands have been presented in the literature.

In general, grassland species occur mainly as plant societies with a great variety within each habitat (Corbane et al., 2015). However, it is still very challenging to distinguish homogenous habitats (Corbane et al., 2015; Hill et al., 2005). Accordingly, the most of current research activities aim at addressing the identification, delineation and change detection of habitats (e.g., in terms of areal coverage, field size,

spatial distribution and management practices) as well as the description of their status and quality. However, at present only few studies are tackling this issue by means of remote sensing techniques.

In general, so far only satellite data with low (> 300m, e.g., MODIS) or medium (30 - 300m, e.g. AWiFS) spatial resolution have been employed to derive national and continental land-cover maps. However, these are not suitable for a proper categorization of grassland habitats, which, to cope with their generally small extent, ideally requires high (3 to 30 m, e.g. Sentinel-1 & 2, Landsat or SPOT) to very high (< 3 m, e.g. TerraSAR-X, WorldView-2, Quickbird) spatial resolution imagery. Moreover, given the similarity of different grassland types in their physical appearance, high frequency acquisitions over the growing season are essential for characterizing their temporal behavior, which, instead, may consistently vary e.g., due to their use associated with different mowing practices (Schlager et al., 2013; Franke et al., 2012; Zillmann et al., 2014; Lucas et al., 2007).

2.3.3.2.1 Grassland monitoring by means of optical data

Pixel based approaches based on optical data

As optical remote sensing data with a high spatial as well as temporal resolution covering large geographical areas have only become available in the last 10 - 15 years (e.g., IRS-P6 (23.5m), RapdiEye (6.5m), and Sentinel-2 (10-60m)), various previous studies addressing grassland monitoring solely applied low spatial resolution satellite data for large area analyses, e.g. the NOAA-AVHRR or MODIS used to discriminate grasslands as a whole from other land-cover types (Hill et al., 1999; Wang et al., 2013), to estimate the aboveground biomass of grassland during the growing season (Zhao et al., 2014; Yu et al., 2010), to analyse grassland potential productivity dynamics and their carbon stocks (Li et al., 2013), to evaluate land degradation (Tasumi et al., 2014; Numata et al., 2007), or to determine grassland drought (Gu et al., 2007; Wan et al., 2004).

In contrast, medium resolution data such as the Landsat Thematic Mapper (TM) (Jensen et al. (2001), Wood et al. (2012), Price et al. (2002b)), the Landsat-7 Enhanced Thematic Mapper (ETM+) (Sanchez-Hernandez et al. (2007), Lucas et al. (2007)) or the Advanced Wide Field Sensor (AWiFS) on board the IRS-P6 satellite have been employed more recently in a variety of studies to classify different grassland types and habitats (Jensen et al., 2001; Sanchez-Hernandez et al., 2007; Lucas et al., 2007), to determine grassland changes (Rufin et al., 2015; Zha & Gao, 2011; Liu et al., 2004), to describe vegetation structure and management practises (Wood et al., 2012), or to evaluate grassland degradation (Price et al., 2002b).

High and very high resolution data sets, e.g., LISS-III on board the IRS-P6 satellite, RapidEye, IKONOS-2, or Quickbird, have been used not only to distinguish grasslands from crops (as for a test site in North-East Germany (Esch et al. (2014a, 2014b)) or in the context of a pan-European permanent grassland map (Zillmann et al. (2014)), but also to classify different grassland habitats. Buck et al. (2013) and Buck et al. (2015) integrated expert knowledge in form of raster information layer into the classification approach (where they tested the maximum likelihood and SVM classifiers) to map Natura2000 grasslands types, intensively used grassland and crops based on three RapidEye scenes. Stenzel et al. (2014) applied a Maximum-Entropy (MaxEnt) one-class classification approach (Phillips et al., 2004) on a time series of five RapidEye images over a test site in southern Bavaria (Germany), which generated a set of logistic probabilities maps that were finally combined creating one grassland map. Schmidt et al., 2014 used several RapidEye scenes and different combinations of vegetation indices as input for a SVM classification and presenting the best settings to discriminate semi-natural grassland classes. This study aimed to assess the most suitable phenological season to get optimized results and the best trade-off between the minimum number of individual scenes needed to achieve the best corresponding classification accuracy. They concluded that NDVI composites from early summer season are most important for such classification tasks. Furthermore, full spring season, late summer and midsummer seasons were also found to be important and contributed to a better grassland discrimination. Specifically, data from March, May and August were found to necessary to discriminate crops and grassland for Central Europe (Keil et al.

2013). However, these dates vary for different regions with changing climate conditions, different crop cultivations and land management practises (Zillmann et al., 2014).

Object based approaches based on optical data

While the above mentioned methods focused on pixel-based approaches, others focussed on the development of object-based approaches for discriminating grassland and diverse habitats. In particular, these are based on predefined objects (e.g., from existing geodata from national topographic maps, or segmentation of homogenous regions) and have the great advantage of including additional knowledge such as region based spectral and texture features, form features or context information (Bock and Lessing, 2000). Amongst others, Bock et al. (2005a) developed and assessed an object-oriented fuzzy-rule classification for habitat mapping at the regional scale (based on dual-date Landsat ETM+ scenes from 2001) and at the local scale (based on high resolution stereo camera (HRSC) scanner data from 2001) accounting for information derived from a soil and topographic map. Furthermore, Bock et al. (2005b) applied object-based classification for monitoring dry grasslands and wetlands by means of multi-temporal and multi-resolution EO data both at the regional (in a study site in Schleswig-Holstein, located in Northern Germany) and local level (in a study site in Wye Downs (UK)). While for the regional study a time series of Landsat TM/ETM+ scenes from the years 1990, 1995, and 2001 has been used, one pan-sharpened Quickbird image of 2002 has been employed for the local study to develop a hierarchical methodology based on fuzzy rules and nearest neighbour classification. Díaz Varela et al. (2008) studied the potential of the maximum likelihood classifier and the nearest neighbour decision rule for addressing both pixel- and object-based classifications of one Landsat TM image acquired over a test area in the Northern Mountains of Galicia (Spain), which is characterised by a heterogeneous landscape, also including habitats of the Natura2000 network. Franke et al. (2012) analysed the potential of multi-temporal RapidEye data for a large-scale assessment of grassland use intensity based on commercial decision tree software See5 (RuleQuest Research Pty Ltd, Australia) and using multi-temporal NDVI, Normalized Red-Edge Vegetation Index (NREVI), and Mean Absolute Spectral Dynamic (MASD) as input parameters. Secondly they tested a context-based classifier. Both approaches were implemented as object-based classification systems. Also, Corbane et al. (2013) successfully classified two habitat types (i.e., dry improved grasslands and riparian ash woods) using two RapidEye scenes and a DEM for a test site located in Foothills of Larzac in the Southern Massif Central (France). This was possible by applying an object-oriented sparse partial least square discriminant analysis. Schlager et al. (2013) introduced a classification approach specific for discriminating grassland habitats in the biosphere reserve Schwäbische Alb (Germany) based on a multi-sensor remote sensing data set consisting of an orthophoto composite, 6 RapidEye scenes, and LiDAR data set as well as vector data from the Authorative Topographic-Cartographic Information System (ATKIS®) and the Integrated Administration and Control System (IACS, German: InVeKoS). Petrou et al. (2014) applied an object- and rule-based classification methodology to map Natura 2000 habitats (i.e., two extended coastal lagoons, numerous channels, marshes and humid grasslands) in the Le Cesine test site located in the Apulia region in south-eastern Italy. The experiments were based on a pre-existing land cover map, two multispectral images from Quickbird and WorldView-2 as well as an Object Height Model (OHM) extracted from LiDAR data.

2.3.3.2.2 Grassland monitoring using SAR data

Similarly to optical imagery, also synthetic aperture radar (SAR) data have been successfully applied in several studies for discriminating different crop types (McNairn and Brisco, 2004; Ferrazzoli et al., 1997; Blaes and Defourny, 2003; Lopez-Sanchez et al., 2011; Wegmüller and Werner, 1997); however, they have been seldom employed for classifying grassland habitats. Furthermore, in such context studies accounting for multitemporal series of SAR images are extremely rare. Available data from current and past SAR satellite missions are mainly acquired in three frequency ranges: L-band (1-2 GHz; e.g., ALOS/ PALSAR, JERS-1), C-band (4-8 GHz; e.g., Radarsat-1 and Radarsat-2, ERS-2/SAR, Envisat/ASAR), and X-band (8-12 GHz; e.g., TerraSAR-X/Tandem-X, COSMO-SkyMed, PAZ). While C-band and L-band data have longer wavelength and can penetrate through vegetation (hence being more suitable for forest analyses), X-band

data are not penetrative and thus more suitable for short vegetation cover, such as grasslands. However, only in recent years the acquisition of high-temporal frequency SAR imagery has become possible, thus enabling a variety of new possibilities.

Hill et al. (2000) evaluated the applicability of Radarsat-1 C-band single polarisation (HH) data for monitoring grasslands in test sites located in Australia and Canada. In particular, they applied a clustering followed by a maximum likelihood classifier to different datasets obtained combining the backscattering information with texture features. The use of multiple images allowed a consistent improvement with respect to using a single one; moreover, the degree and regularity of surface roughness proved to be the most informative feature. Smith and Buckley (2011) assessed the suitability of multi-temporal Radarsat-2 quadpol imagery to classify native and improved grasslands as well as agricultural crops over a test site in southern Alberta (Canada). The classification on the Freeman-Durden decomposed data was performed by means of the See5 decision tree classifier (RuleQuest Research Pty Ltd, Australia). The results showed the potential to separate native grasslands from agricultural areas as well as native from improved grasslands and that the incidence angle of the acquisition has no influence on the classification accuracy. Schuster et al. (2011) showed that habitat-specific swath rules describing management practices are an important parameter in the conservation of semi-natural grasslands and can be used to indirectly map specific habitat types. They introduced a method to detect swath events based on a time series of eleven TerraSAR-X images (HH polarisation, Stripmap mode) over a nature conservation area west of Berlin (Germany) and analysed the temporal profiles of the backscattering coefficient σ0 by applying a rule-based approach to detect swath events. Results were compared to ground-truth data as well as to habitat-specific swath rules defined to conserve Natura 2000 habitats. Furthermore, Schuster et al. (2015) analysed the potential of grassland habitat mapping by means of inter-annual time series data (2009-2011) of RapidEye and TerraSAR-X data acquired over a 60km² test site in Northern Germany. Based on individual sets of five RapidEye and 15 TerraSAR-X scenes, after masking non-grassland areas they mapped seven grassland classes with a SVM and were able to achieve overall classification accuracies higher than 90%, with Kappa coefficient greater than 0.9. Betbeder et al. (2015) investigated the optimal number and key dates for the acquisition of dual-polarisation (HH/VV) TerraSAR-X images to classify wetland vegetation formations in a 6.7 km² test site located in the Bay of Mont-Saint-Michel (France). The available eight dualpol TerraSAR-X scenes were decomposed using the Shannon Entropy (SE) calculation and a SVM classifier with a Gaussian kernel was then used to categorise six classes (of which four are wet grassland types) based on training points collected in situ. Five images proved to be the best trade-off between the number of acquisitions and the final overall accuracy; moreover the best combination was obtained using scenes acquired in February, April, May, June, and July, i.e. when plants grow actively and hydrodynamic processes are vibrant.

A variety of approaches jointly apply multi-sensor imagery from SAR and optical satellites for the classification of vegetation classes, such as crop types (Brisco and Brown, 1995; Blaes et al., 2005; McNairn et al., 2009), and crops combined with more general land-cover classes (Waske and van der Linden, 2008, Waske and Benediktsson, 2007), or for the estimation of herbaceous biomass (Svoray and Shoshany, 2003). Smith et al. (1995) analysed ERS-1 SAR data together with Landsat TM, SPOT VIR, and airborne optical imagery to assess the combination of radar and optical data for monitoring rangeland in the Agriculture and Agri-Food Canada Research Substation at Onefour (Alberta) by means of discriminant function analysis (DFA). The combination allowed obtaining an improved categorisation of the vegetation classes with respect to considering each data type separately; moreover, while optical data proved to be more suitable to characterise the vegetation status, SAR imagery provided key information about the structure and surface topography. Also Price et al. (2002a) used a classification system based on the DFA to study the separability of three tallgrass land management practices in eastern Kansas (USA), where usually cool- and warm-season grass species occur, by means of three multi-seasonal Landsat TM and four multi-seasonal ERS-2 SAR images, as well as their combination. The results showed that by using Landsat TM data alone performances were better than those obtained with ERS-2 imagery and, when combined, the SAR data did not allow to increase the classification accuracy. Hill et al. (2005) showed the potential of improving the categorization of heterogeneous herbaceous cover in pastures and grasslands by combining

independent classifications obtained by means of mono-temporal Landsat-5 TM and Jet Propulsion Laboratory AirSAR data. Experiments were performed for a test site in the Cervantes area (Australia) using an unsupervised version of the Complex Wishart classifier for the C-, L-, and P-band polarimetric SAR data as well as a principal component analysis on the green, red and near-infrared Landsat bands followed by a centroid distance measure clustering. In particular, they were able to map vegetation types based on the different sensitivity of SAR and multispectral sensors to specific vegetation characteristics. Erasmi (2013) assessed the capability of combining optical (six RapidEye scenes) and SAR (four Radarsat-2 and six TerraSAR-X scenes) data for the classification of semi-natural habitats over the study site Schorfheide Chorin in eastern Germany and compared the results with single sensor classifications. The object-based classification was performed by means of a classification and regression tree (CART) algorithm. Results showed that single-sensor classifications based on multi-temporal RapidEye data outperformed the once carried out with TerraSAR-X and Radarsat-2 data and demonstrated that bi-sensor combinations of optical and SAR data resulted in classification accuracies between 60.83% and 84.53% (with Radarsat-2 polarimetric data providing higher classification accuracies than TerraSAR-X). Metz (2016) proposed a system which proved to be robust and confirmed the effectiveness of employing multi-temporal and multi-polarisation VHR SAR data for discriminating grassland types. Tamm et al. (2016) aimed to describe the relationship between Sentinel-1 A 12 day temporal interferometric coherence and mowing events on grassland. The study area includes 37 fields, six of which were in situ monitored on a weekly basis. In total 77 mowing events were observed on all test sites combined. Coherence is higher on bare soil than on fields with remaining vegetation and the increase in coherence after mowing events is highly dependent on the specific mowing method.

2.3.3.2.3 Time series approaches

Grasslands are highly dynamic throughout the time and its growing period with changing canopy density, chlorophyll status and ground cover and therefore do not have a unique spectral signature which allows a simple discrimination from other vegetated land cover classes (Zillmann et al., 2014). Especially grasslands and crops show significant variations throughout their growing cycle. Therefore, time series of data which mirror the phenological dynamics of grasslands are required. The usage of multi-temporal and multi-sensor data led to improved land cover classification especially of vegetated classes as it allows the observation of phenological effects. High temporal resolution of input data covering different seasons is also required to properly categorize grasslands (Metz 2016). Because of similarity of grassland types with other land cover classes as well as physical appearance the data need to cover growing seasons with higher temporal resolution to enable detailed characterization of temporal behavior differences and use the gained temporal information for better class discrimination and thus grassland classification with higher thematic classification accuracy. Analysis of spectral variability metrics allows discriminating between different land cover classes especially grass-dominated pastures from woody vegetation (Ruffin et al., 2015). Following, we present promising approaches which use dense time-series of data and derived metrics to classify different grassland related classes.

Zillmann et al. (2014) investigated an approach based on decision tree classifier C5.0 and optical multi-temporal imagery to generate a high-resolution pan-European grassland layer. They applied image segmentation and calculated seasonal statistics for various vegetation indices. They identified 7 indices to be useful for grassland classification especially regarding the discrimination of grassland and crops, namely: NDVI, ground cover (GC), Plant Senescence Reflectance Index (PSRI), Normalized Difference Infrared Index (NDII), Normalized Difference Senescent Vegetation Index (NDSVI), Wetness Index (WI), and Brightness. For each index seasonal statistics were calculated as they describe spatio-temporal phenological differences of vegetation and thus, enhance the discrimination between grassland and other vegetated land cover typed (especially crops). Yang et al. (2017) investigated a set of vegetation indices to detect changes of natural grassland to cultivated crops and the optimal timing of data acquisition, namely: Normalized Difference Vegetation Index (NDVI), Red-Green Ratio (RGR), Enhance Vegetation Index (EVI), Normalized Difference Infrared Index (NDII), Modified Triangular Vegetation Index II (MTV2), Shortwave Infrared Reflectance (SWIR32), and Plant Senescence Reflectance Index (PSRI). They verified that all

analysed indices were important for distinguishing native grassland and cropland. However, the optimal mix was changing with each month during the growing season (Yang et al. 2017).

Mueller et al. (2015) used Landsat time series to separate cropland, pasture and natural savanna vegetation using spectral-temporal variability metrics and random forest classifier. They concluded that deep temporal information derived from time series data is the key in a phenologically complex land cover system. Wang et al. (2017) used PALSAR mosaic data and Landsat 5/7 data to develop a pixel and phenology based mapping algorithm which helped to analyse the encroachment of red cedar into grasslands. The introduced approach can be also adopted for the classification of different grassland types. Also Cui et al. (2017) used a long time series of NDVI data to analyse the phenology response of grassland to draughts using the TIMESAT software (Eklundh and Jönsson, 2015). Lopes et al. (2017) discussed an approach using dense time series of satellite data such as Sentinel-2 to formulate the Spectro-Temporal Variation Hypothesis assuming that the spectral variability in time can be used as a proxy for grassland and different grassland species detection. Liu et al. (2017) utilized time series data of MODIS, VIIRS, Landsat sensors to monitor open grassland and oak/grass savanna and discussed the influence of spatial resolution. The following phenological metrics were identified to be essential to analyse the phenological cycle of open grassland and oak/grass savanna: the timing of the Onset of Greenup = the onset of the NDVI increase (OG); the full Maturity of the Green canopy = the onset of the maximum NDVI (MG); the commencement of senescence (or End of Greenness) = the onset of the NDVI decrease (EG); and full Dormancy of Green vegetation = the onset of the NDVI minimum (DG) (Liu et al. 2017). McInnes et al. (2015) found that native grasslands can be distinguished from spectrally similar tame pastures when using dense time series of NDVI data and generated seasonal profiles of the classes. The authors observed that the separation of the two classes was possible due to a different rate of spring green up at pixel level. The classification was performed based on simple linear discrimination function. Discriminant analysis builds a predictive model for group membership based on natural breaks in the data, using analysis of variance (ANOVA) techniques and multiple regressions (McInnes et al. 2015). The availability of vegetation index data in the early growing season was found most important for the discrimination of grassland and other spectrally similar land cover types.

The accuracy of grassland classification depends on the number of images in the time series, but more importantly on the optimal acquisition date and gap free data during the growing season. Many studies dealing with grassland detection based on remote sensing data have been using pre-existing land cover classifications information to avoid misclassification in areas where grassland can be excluded (Petrou et al. 2014). Depending on the assessment of tested approaches, this strategy can be implemented additionally to increase the detail and accuracy of the end result. Nevertheless, more research is needed on the spatio-temporal variation of the coverage of grass canopy and grass height (Rodríguez-Maturino et al., 2017).

2.3.3.3 Mapping Mediterranean Grassland with Multi-temporal Earth Observation Data

2.3.3.3.1 Intention of this desk study

With its pan-European component of High Resolution Layers (HRL) the Copernicus Land Monitoring Service aims at providing detailed information on land cover and land use, on change of land cover and land use and on land cover characteristics. The HRLs 2015 provide land cover information on five main themes, namely Imperviousness, Forest, Grasslands, Water and Wetness, and Small Woody Features. These layers derived from multi-temporal, multi-seasonal and multi-sensor EO data by application of elaborate methodological approaches, which all have been continuously refined during the last years, except for grassland. The HRL Grassland 2015 was completely novel, challenged by developing a methodology that would be suitable throughout Europe under highly variable conditions and at the same time ensured constantly high standard and reliability.

Whereas this specifically designed method for grassland detection proved to be most practicable and efficient for Northern and Central European areas and led to a highly accurate HR grassland product for the reference year of 2015, the outcomes regarding the Mediterranean region[1] showed potential for enhancement and fostered a second thought about a methodological adaption.

The Mediterranean region shows a considerable amount of natural and semi-natural grassland formation: roughly 50% of the Mediterranean basin are dominated by grasslands (Eurostat 2013) with exceptionally high biological diversity, representing ecosystems of High Nature Value[2] (Duarte 2011; Vrahnakis). However, mapping of these and other significant grassland areas by means of EO date is challenging due to differing vegetation seasons as well as differing management systems. Main limitations in the Mediterranean region are for example the identification of sparse and dry grasslands during arid summer months, the detection of grassland in wooded areas, the distinction of grassland and shrubs in abandoned regions or the differentiation of very detailed grassland and cropland plots in traditional small-scale farming in rural areas.

Methodological adaptions postulate an adequate knowledge and understanding of the climatic and geophysical conditions and the land use patterns in the Mediterranean region and of the possible consequences this has for the mapping of grassland with EO data. Thus, this study serves two purposes:
First, it aims at analyzing the characteristic features within the Mediterranean region of the EEA39 members which differ most from those experienced in the Northern and Central European countries and which may have influence on an effective and accurate detection of grasslands with remote sensing methods. These include the biogeographic conditions in the Mediterranean region, such as climate and soil and the resulting vegetation cover, photosynthetic activity and the growing peak of vegetation; and, as the differentiation between grasslands and non-grasslands poses one of the major challenges, the specific management systems concerning the cultivation of grassland and agricultural areas that could ease this differentiation through the identification of time slots when both types of vegetation differ most.

Second, it identifies changing parameters within the current methodological approach of mapping grassland and recommends adequate adjustments. An adaption of the time slots for satellite data oriented towards the specific vegetation peaks of Mediterranean grassland and a stronger involvement of the potential of SAR data can be seen particularly promising in view of an accuracy enhancement of a future HRL Grassland layer within the Copernicus Land Monitoring Service.

2.3.3.3.2 Bioclimatic conditions for grassy vegetation in the Mediterranean region

Climate - namely the provision with sunlight, suitable temperature and the availability of water - is the main factor that influences biological systems and affects the spatial distribution of plants, biomass production, growth cycles and vitality and thus sustains ecosystem functions and processes. The second factor is the potential of the soil in supplying vegetation adequately with nutrients and moisture.

In order to answer questions of where, when and what type of grasslands we could expect, further insights into the underlying geophysical conditions for vegetation growth are an important prerequisite.

[1] It has to be pointed out that the term “Mediterranean region” within this study refers to specific areas around the Mediterranean Sea which are characterized by Mediterranean climatic conditions as described in the next chapter. They are not synonymous with the national boundaries of the Mediterranean countries in a geographical sense. The interchangeable expression would be “Mediterranean basin”.

[2] Developed in the 1990s, the concept of High Nature Value displays those areas manifesting exceptional high biodiversity and representing typical landscapes which deserve protection. The concept aims at supporting these areas throughout the EU-territory by fostering the continuity of low intensity and sustainable farming systems across large areas of the countryside (EEA Report No 1/2004).

CLIMATE

Despite being Mediterranean countries regarding geography, most countries of the Mediterranean region are divided into several bioclimatic regions. Mediterranean climates (after Köppen and Geiger, see Figure 2-7) occur on the west side of the Mediterranean continental land masses between 28° and 45° latitude. They range from subtropical subhumid to dry climate with warm to hot summers, intensive sunshine and seasonal summer droughts[3] of variable length, and wet and mild winters with relatively high inter-annual variability (Spano et al. 2003; Peel et al. 2007; Zolotokrylin 2012), correlating to the climatic subclasses Csa (hot and dry summer Mediterranean climate), Csb (warm and dry summer Mediterranean climate) and Bsh (steppe-hot Mediterranean climate) regarding parts of the Iberian Peninsula. Mediterranean climates function as essential transition zones between temperate and dry tropical climates (Porqueddu et al. 2016) and are distinct and at the same time heterogenic as a result of the complicated morphology, orographic features, the large mass of water of the Mediterranean Sea and the influence of both, Atlantic and Continental macro weather conditions. That causes a high spatial variability of subregional and mesoscale climatic features depending on

- latitude
- altitude
- vicinity to the coast
- location on Eastern or Western coast
- location in mountainous coasts
- influence of the Atlantic Ocean
- location influenced by maritime or continental climate

The climatic classification after Köppen and Geiger bases on precipitation and temperature, allowing a general orientation on the geographical extension of the Mediterranean (Kottek et al. 2006; Peel et al. 2007; AGROMET/FAO 2006). Characteristic features of the Mediterranean climate type are:

- annual precipitation ranges from 250 to 900mm, mostly falling from November to April
- average temperatures in winter months go below 25°C
- the amount of time when temperatures fall below 0°C must not exceed 262 hours a year

(Aschmann 1973; Spano et al. 2003; Vrahnakis 2016; Rivas-Màrtinez et al. 2003, Rubel et al. 2011)

[3] Drought can be defined as an „extended period when evapotranspiration exceeds precipitation, causing the depletion of soil moisture and consequently reduction of ecosystem productivity" (Zolotokrylin 2012). Whereas dryness is a constant feature of arid areas, caused by climate, drought is a temporary phenomenon. In the Mediterranean region, seasonal droughts during the (arid) summer months are a common feature.

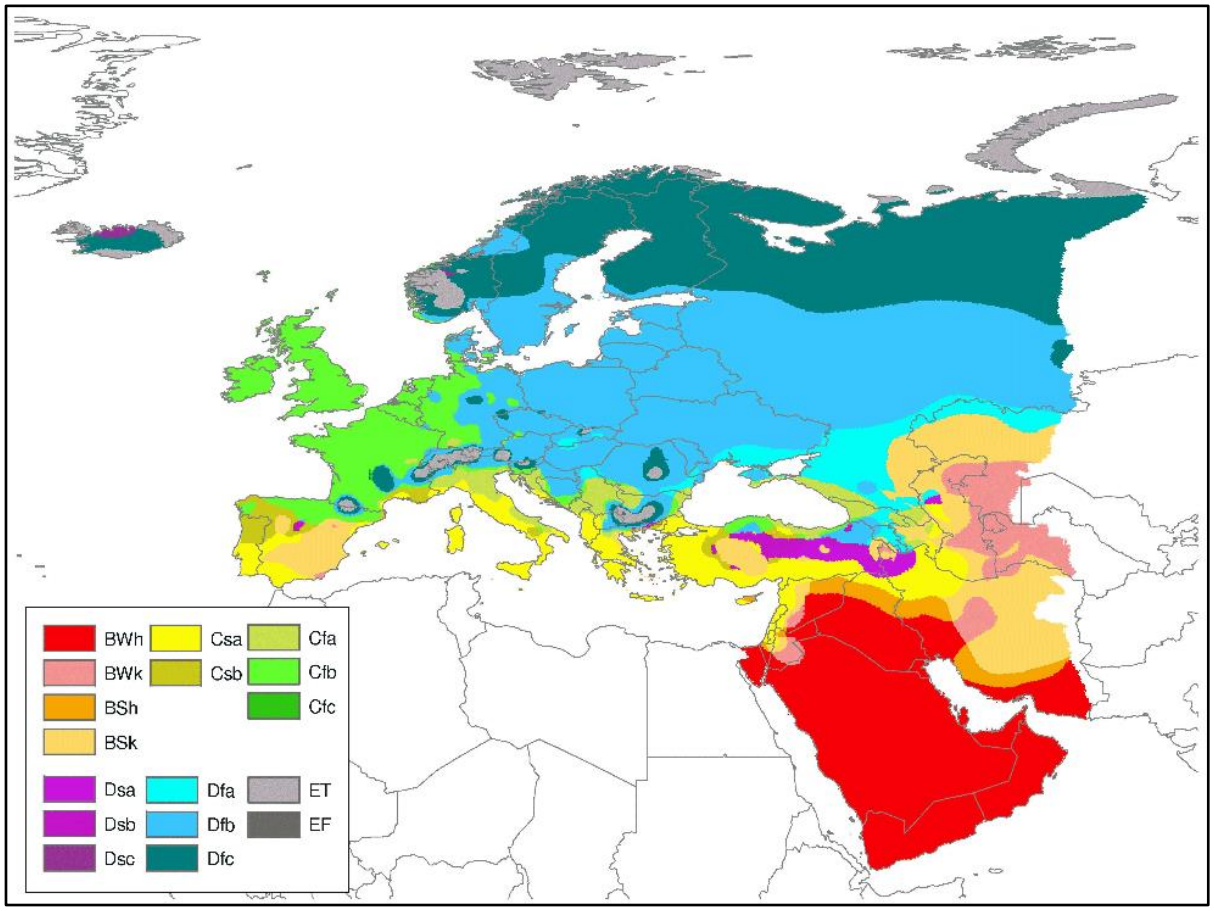


Figure 2-7. Mediterranean climatic map after Köppen & Geiger (Peel et al. 2007).

The tendency to and also the length of an arid summer period and consequently the risk for seasonal summer droughts increases from North to South and from East to West. Given the fact that continuing dry periods lead to degraded photosynthetic activity of plants, satellite data from summer months should be handled with care for grassland mapping, providing insufficient information on the status and the actual existence of vegetation.

SOIL

Due to its formation history, a variety of soils can be identified in the Mediterranean basin. Hence, grasslands provide a high diversity of plant species proving very flexible and adoptable to different soil conditions.

Soil plays an important role in detecting vegetation with remote sensing: soil characteristics do not only bear specific vegetation types as result of distinct nutrient and moisture content. Soil also influences the spectral response by mixing up its own spectral response with that of the respective vegetation cover. This effect is more pronounced in dry areas and arid months of the year, when vegetation gets sparse or withers as a result of drought. Particular attention has to be given to special types of soil as the detection of grassland in dry areas in the context of the HRL Grassland 2015 taught: saline soils for example show unusual deep purple shades. Due to the sparse vegetation and influenced by the dry conditions, the grassland vegetation is hard to identify because it mixes up with the spectral response of the saline soil. Hence, basic knowledge on distinct soil features like saline soils (which are frequent in the semi-arid regions) or the so-called "*ferra rossa*", ferruginous brown soils (e.g. in Spain) showing deviating reflectance in the optical spectrum, are indispensable for an accurate mapping of grassland vegetation.

Concerning land use, there is a clear distinction between fertile and marginal soils: Fertile soils with deep organic layer and abundant water supply are mostly used for (intensive) crop farming whereas soils which

are covered by grasslands show low fertility because of a low organic layer, lack of nutrients and often insufficient moisture which makes them marginal for planting cereals (Mesías et al. 2010). As the thickness of the fertile organic layer corresponds directly to the climatic conditions (more humid climate results in extensive soil and organic layer formation, more arid climate reduces these processes), it can be concluded, that grassland mapping in the Mediterranean regions should focus on less fertile, marginal soils, assuming that grassland vegetation would be the prior vegetation cover in these regions. As for fertile regions, there must be high awareness on the differentiation between grassland and the dominating cropland areas. In general, background knowledge on regional soil features is necessary for an accurate identification of grassland vegetation.

ADAPTATION OF VEGETATION TOWARDS BIOGEOGRAPHIC CONDITIONS

The biogeographic map of the EEA combines the previously stated climatic, geophysical and soil characteristics and provides basic information on real as well as potential climatically adapted vegetation cover (Canu et al. 2015; Smiraglia et al. 2013). Based on hydrologic cycles and the distribution of typical habitats according to the EU Habitats Directive, it answers the question of "*Where* can we expect typical Mediterranean vegetation?" and implies that vegetation in the given biogeographic regions follows very distinct annual life cycles.

Accurate mapping of vegetation requires data of those time slots when vegetation shows the highest level of photosynthetic activity. The question of "*When* can we expect vegetation?" is first and foremost determined by the local and seasonal climatic conditions, which means that vegetation flourishes if the provision with water and sunlight is adequate. The life cycle of all plants (i.e. growing season) starts as soon as temperature goes above 12°C and water availability is sufficient. It comprises germination and seedling emergence, stages of flowering and seed set and ends with dieback of parts or the whole plant or the entering of dormancy when temperature or humidity decreases (George and Rice 2012).

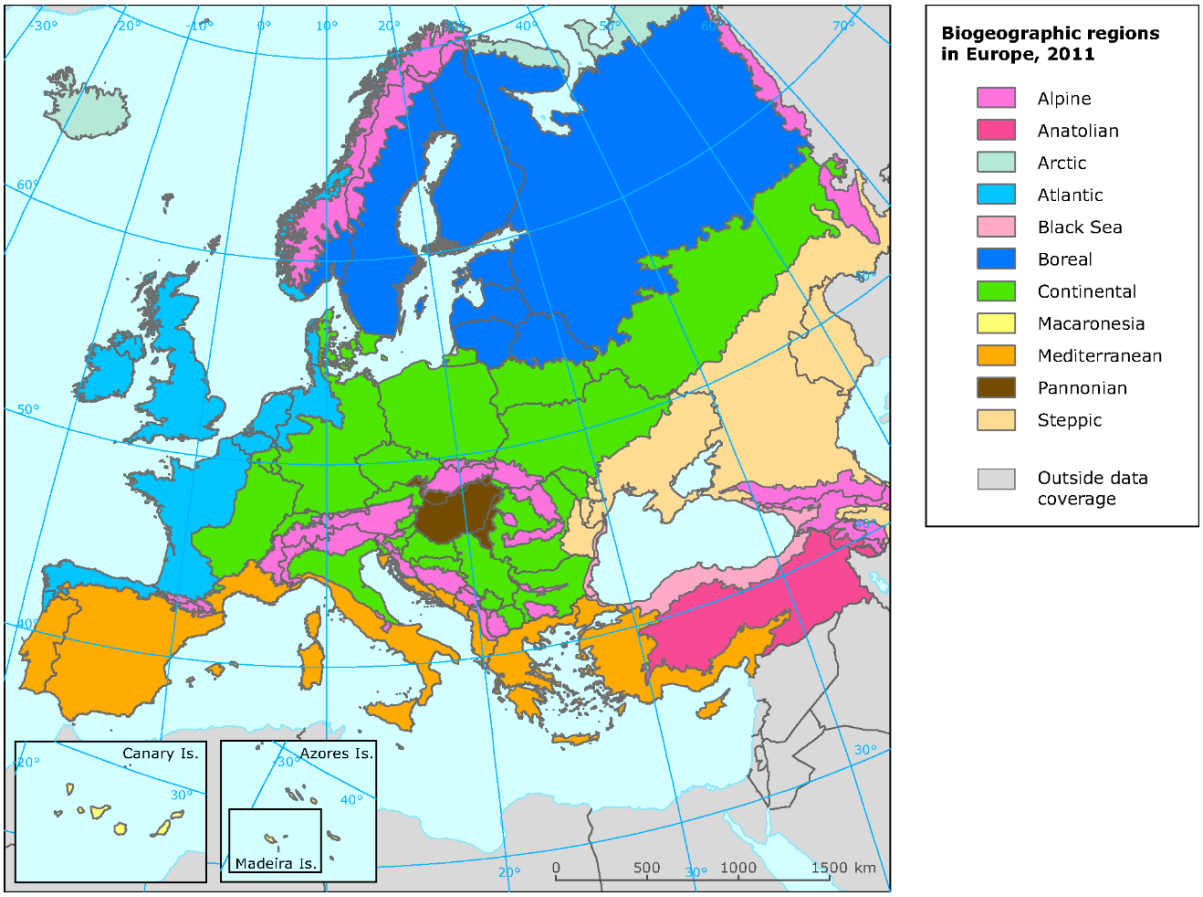


Figure 2-8. Biogeographic regions in Europe 2011 (EEA 2012).

In Mediterranean regions where there is a pronounced summer dry season with seasonal drought and mild winter temperatures with high water availability, the growing period of the whole vegetation cover experiences a shift: instead of growing from spring to autumn as usual in Central Europe, the growing season starts in early spring, shows rapid spring growth and reaches its peak in late spring, just before soil moisture is depleted at the beginning of the dry season. At this point, vegetation growth pauses, adopting to arid conditions and droughts with summer quiescence/dormancy (in the case of perennial plants) or withering and dieback (in the case of annual plants). With the onset of rain in autumn vegetation starts regrowth or sprouting and a new life cycle begins. Depending on temperature and the amount of precipitation in winter, some areas show vegetation growing even during the winter months (Ervin et al 2004): if temperatures remains above 12°C during winter months, high photosynthetic activity and water use efficiency can be observed, resulting in high growth rates in autumn and winter months. If temperature is moderate, vegetation growth slows down and restarts rapidly with increasing temperatures in spring.

Mediterranean grasslands have adopted similarly to the warm to hot and dry climate with a life cycle following the annual rainfall distribution, reacting particularly sensitive to variations in precipitation (Carmona et al. 2012). Consequently, ideal time slots for the detection of grasslands are:

- **Spring**: February-April for a first vegetation peak with high level of photosynthetic activity
- **Autumn**: September-November for a second vegetation peak after the beginning of winter precipitation; in Southern regions where winter temperatures stay above 12°C, vegetation can be detected throughout the winter months and could provide additional data for grassland mapping.

The distinct situation in the Mediterranean region places high demand on the selection and the assessment of satellite data. The objective of an optimal detection of grasslands has to consider several aspects: the shortened and shifted vegetation peaks of grasslands under dry conditions, a high inter-regional and inter-annual variability in the climatic patterns and an increasing aridity in summer going from North to South and from East to West, which results in quite restricted time slots for adequate satellite data.

2.3.3.3.3 Grasslands in the Mediterranean region

The term *grassland* involves several aspects: according to the working definition for grassland underlying the HRL Grassland 2015 (see chapter 2.3.3.1), it comprises a diverse range of plant species, of grassland types depending on the geophysical prerequisites and of Mediterranean landscapes shaped by grassland vegetation.

Due to its large-scale and synoptic approach of mapping grasslands, remote sensing data cannot aim at a detailed estimation of distinct plant species. However, being aware of the diversity of vegetation cover and of the different growing conditions that favour or discriminate a specific type of vegetation cover, is an important prerequisite for two reasons: for the accurate detection of grassland, due to its highly variable range of spectral characteristics and for the identification of adequate time slots for satellite data acquisitions, due to growing characteristics. The same thoughtfulness has to be paid for the different environments grassland is part of, because surrounding land cover features influence the spectral response of grassland and may complicate the clear differentiation between grasslands and non-grasslands.

GRASSLAND DOMINATED LANDSCAPES

In the Mediterranean region, grassland is traditionally part of characteristic landscapes such as

- wooded grasslands with oak trees, cork-oak trees or olive trees, providing the economical basis for sylvo-agro- or sylvo-pastoralism, p.e. *Dehesa* (Spain)
- grassland-shrubland mosaic used as pastures or basis for agro-pastoralism like *Mato* (Portugal), *Maquis* (France), or *Macchia* (Italy); the density of the shrub cover varies within these landscapes

- degenerated grassland-shrubland-mosaic with singular trees and taller shrubs due to intensive grazing, wild-fires or extreme droughts like *Garrigue* (France, especially Corse) or *Phrygana* (Greece, Turkey), or abandoned areas
- highland pastures

Whereas spacious rangeland or pastures are well detectable relating to large-scale and widely homogeneous spectral characteristics, these typical Mediterranean landscapes with their heterogeneous vegetation are challenging to map with remote sensing. Excluding scattered trees, as for the common sylvo-agro-pastoral areas, or shrubby areas in mixed grassland landscapes proved to be difficult with optical data only, as experiences within the HRL Grassland 2015 has shown. In many cases that meant also excluding a larger amount of grassland, due to the mixed spectral responses at a spatial resolution of 20m. Being able to detect texture and structure of the surface, SAR data could fill the gap. An accurate SAR classification could well enhance the optical classification by focusing on specific non-grassland area classes that can be better detected using SAR data, and can therefore be excluded from the grassland areas.

GRASSLAND TYPES

Wet and dry grasslands

Grassland types are strongly related to climatic conditions. In those areas showing temperate climate with sufficient precipitation in all seasons and adequate nutrient supply due to favourable soil conditions, the grassland types and their specific composition of grassy plants are similar to those of Central Europe. Additionally, there can be found regional grassland types such as wet grasslands, e.g in Bulgaria (the country belongs to the sub-Mediterranean climate type: Hájek et al. 2007). Wet grasslands are seldom in the Mediterranean region. More common is the type of dry grasslands due to predominant arid climate.

The origin of dry grasslands is often human intervention in the past, when Mediterranean forest landscape has been cleared in order to provide new arable land for an increasing population. Dry grasslands nowadays account for the majority of grassland biotopes on relatively dry and nutrient-poor soils overlaying acid rocks or deposits such as sands or gravels (Veress and Szigethy 2010). They have been used as common grazing pastures and are characterized by short plant cover and high biodiversity. Dry grasses are a typical part of the vegetation cover of grassland dominated landscapes such as steppe grasslands, Alpine grasslands, extra-zonal dry grasslands or Secondary grasslands[4]. An active grazing scheme is a precondition for preservation, otherwise those areas return to shrubland and later on to forest. Three functional types of Mediterranean dry grasslands can be identified: wintergreen perennial grasslands, wintergreen ephemeral grasslands, and, if moisture allows, summergreen perennial grasslands (Guarino 2007; Porqueddu et al. 2017).

Concerning the identification of grasslands with remote sensing, wet grasslands are well detectable with optical EO data, due to the high vitality of the grassland plants. Although grassland and cropland both show similar spectral responses during a similar annual growing period, they can be well differentiated by taking into account the differing management systems concerning different time slots for mowing in the case of grasslands and harvesting and tilling in the case of cropland. In contrast, the differentiation of wet grassland and flooded areas can be challenging. Experience so far shows that a well-considered selection of imagery, potentially complemented by SAR data (regarding permanent wet areas) shows convincing results.

[4] Secondary grasslands are grasslands following human intervention such as logging, forest clearing or fire events which is the case for a large area of the Mediterranean grasslands. Predominantly used as pastures, Secondary grasslands highly depend on permanent cultivation, be it mowing or grazing. Abandoned pastures are at risk of becoming overgrown by bushes or turning into forest (Porqueddu et al. 2014 and 2017).

Dry grasslands, however, are very difficult to map with remote sensing. Due to the reduced photosynthetic activity during the arid summer months, the then sparse and withered plant cover is hard to distinguish from harvested areas, dried crop cover or from bare soil. Hence, imagery from spring and autumn offer a more suitable base for grassland detection.

Annual and perennial grasslands

Annual grasslands plants are very common vegetation cover understory of woodlands and have different life cycles from perennial grassland plants. They are well adapted to the highly variable Mediterranean climate and to regular summer droughts. They produce a huge amount of seeds that survive for a long time in soil seed bank, waiting for early spring precipitation and warm temperatures to sprout. Therefore annual grasslands turn out to be reliably growing every year in the same areas (Cosentino et al. 2014). The life cycle of annual grassland plants usually starts in early spring, showing rapid growth during spring with germination development of seedlings and flower. As soil moisture is depleted, the plants wither and die.

Perennial grassland plants dominate most of rangelands and cease growing during summer drought (drought escape) until autumn, when rainfall allows growing anew. They show growing and increasing photosynthetic productivity in autumn reaching their peak in early spring and re-entering dormancy with the beginning of the dry season.

Wet and dry grassland types as well as annual and perennial grassland plants are both subject to the same regional climate conditions. Despite having evolved different strategies for conquering cold and dry stages of the year, both start their life cycles at the beginning of the rainy season reaching their highest level of photosynthetic activity at the end of spring. It is highly recommendable to adopt the classification method in focusing on this time of the year because grassland vegetation will then be well detectable (Cosentino et al. 2014). During summer dormancy, there is hardly any vegetation detectable. Satellite data for this time of the year provide only little additional information and should therefore be handled with care concerning the time series for image classification.

2.3.3.3.4 Land use and agricultural management schemes

The following map shows areas within the Mediterranean climatic region (area within red boundaries) and the predominant land cover type according to statistics of the European Union (Turkey: no data available). It illustrates the main difficulty of mapping grasslands: the statistic data distinguish several types of land cover, emanating from an approach of land use. Exempt from areas of *Artificial Dominance*, *Dispersed urban areas* and *Forest*, grasslands can be found in all other classes, even be partially included within the class of *Broad pattern intensive agriculture*. Consequently, a methodological approach for grassland detection has to take into account that areas of grassland are located in large areas which are mixed up with various types of land cover.

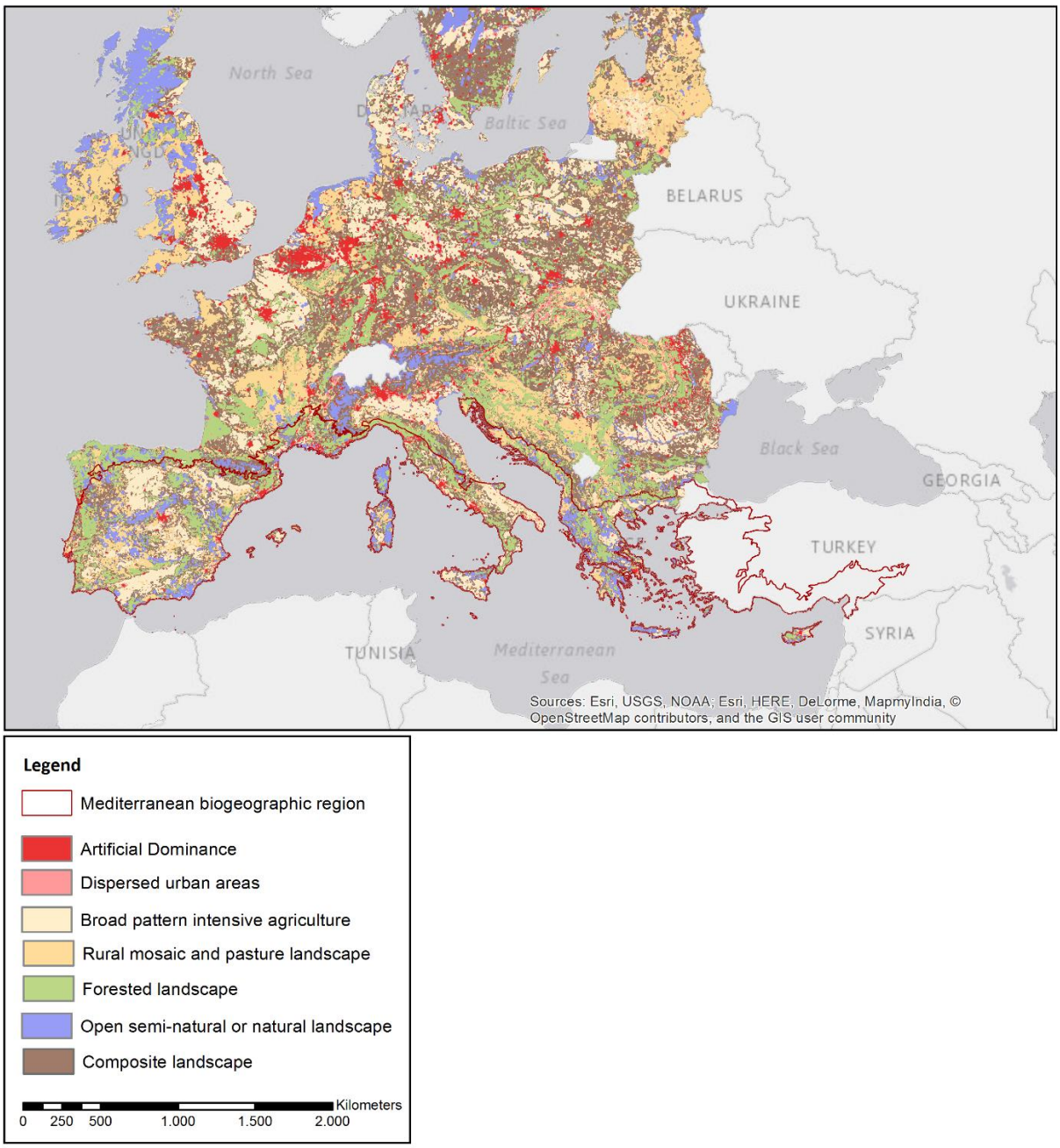


Figure 2-9. Dominant land cover in Europe and the Mediterranean region (red boundaries) (EU 2017 and EUROSTAT 2013)

Traditional Mediterranean agriculture is generally based on vegetation adapted to the local climatic conditions and to soil fertility and has developed unique agro-ecosystems especially for arid and semi-arid regions (Harlan 1992). Due to population and cultural development, land use patterns have been in constant change since the first settlements of man. However, following the climatic conditions, land use shows divergent patterns in the Northern and the Southern parts of the Mediterranean countries. In general, there is a tendency towards focusing on crop production in favourable areas which in turn leads to abandonment of vast grasslands and consequently widespread bush encroachment due to the reduced number of livestock (Landau et al. 2000; Bernuès et al. 2000; Ates et al. 2012).

CROPLAND MANAGEMENT

Besides crop farming, livestock farming and diverse types of pastoralism, land use patterns imply olive orchards, vineyards and horticulture. Hence, the predominant farming systems in the Mediterranean countries can be grouped into three major types of agricultural land use patterns:

- Irrigated systems
- Rainfed systems
- Agro-pastoral systems

Irrigated systems, implying intense type of agricultural land use patterns, occur independent from climatic conditions under both humid and arid regimes and occupy in most cases the more favourable areas concerning soil fertility. Diversity of crops and management schemes are varying, but there are seasonal cropping patterns:

- Winter crops (esp. wheat and barley) and
  Winter legumes (chick pea, lentil, faba bean): Planting or sowing starts in November/ December
  Harvest takes place in April/May
- Summer crops (esp. maize, rice): Planting starts in February/March
  Harvest takes place in June/July

Irrigated systems in their intense and economic form are characterized by larger parcels and are therefore well detectable by satellite data. During summer dry periods, irrigated fields stand out by their much higher vegetation activity compared to the surrounding areas.

Rainfall-based systems are highly dependent on precipitation pattern, their starting point and productivity as well as on the capacity of soil in storing humidity. The diversity of crop production rapidly drops as aridity increases. Generally, the productivity of those systems is low, mainly producing for small rural markets or for subsistence (ICARDA/Biradar). Rainfall-based systems show high variability with regard to crop types and annual management schemes. Remote rural areas show a high heterogeneity of agricultural patches which means that agricultural units tend to be smaller and also tend to cultivation of smaller patches of arable land. This makes it more difficult to differentiate the various vegetation cover with remote sensing data (EUROSTAT 2016: Agriculture and Environment).

Agro-pastoral systems mainly occur in the arid and marginal regions of the Mediterranean basin with soils of low fertility. The cropping pattern and its diversity and productivity within these systems is strongly associated with the occurrence, the yield and annual shifts of the rainfall season. In areas with less than 200 mm precipitation, barley-small ruminant production is the most common. 200-500 mm of annual precipitation allows the production of wheat and small ruminants, whereas precipitation above an amount of 500mm permits horticultural production and cash crop growing (Ates et al. 2012).

Both being strongly dependent on water supply and therefore sharing the same short vegetation period, the life cycle of grassland strongly resembles the life cycle of crops. For both, the onset of precipitation, the amount of water and the soil capacity in water storage are the prerequisites for developing a vital plant cover. In rural areas of the Mediterranean region, farmers rely on both, crop farming and pastoralism, adopting to the natural geophysical conditions. The map above reflects the heterogeneity of agricultural areas in the European Mediterranean region.

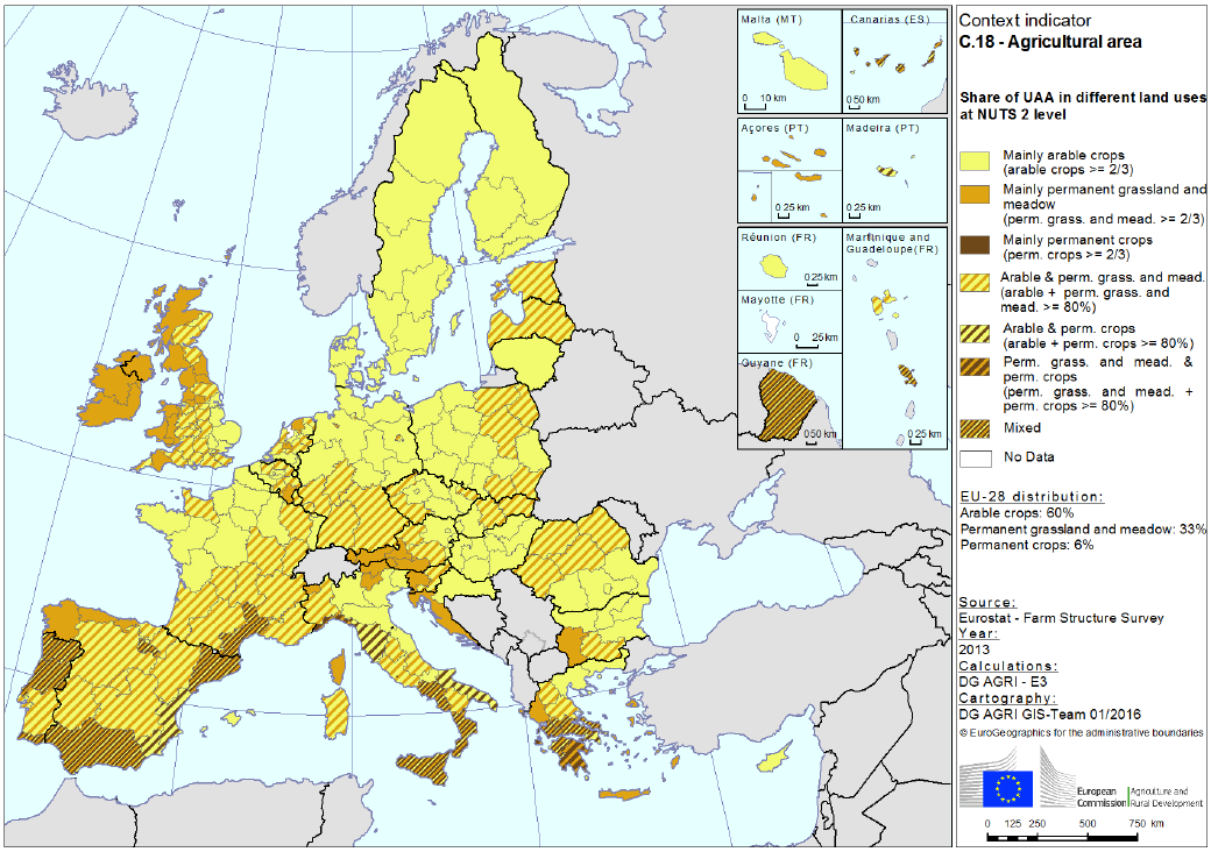


Figure 2-10. Share of utilized agricultural areas (UUA) in different land uses at NUTS 25 level, 2010 (EU2017).

The detection of grassland areas in agro-pastoral systems is highly demanding. Agricultural plots generally show small-scale structures which can barely be identified with optical data providing 20 m of spatial resolution. Moreover, small-scale farmers follow an individual cultivation scheme, being oriented towards annual variation of precipitation and temperature, changing soil quality and personal needs (Casas 2015; CIHEAM 2006; JOUVEN et al. 2010; Maranon 1988; Montserrat et al. 1990; Roggero et al. 2013; Todorovic 2016). Consequently, there is no general time scheme that could support a large-scale differentiation of cropland and grasslands within remote rural areas. However, it can be assumed that larger, coherent areas are detectable after tilling or harvesting events.

GRASSLAND MANAGEMENT

Corresponding the definitions of the HRL Grassland 2015, natural and managed grasslands both are part of the grassland product. Besides natural grassland areas in more elevated regions, there are two main categories concerning the land use pattern of grasslands:

- **Meadows**: grasslands that are harvested predominantly by mowing
  Meadows primarily occur within the humid-subhumid regions, providing enough plant cover and density to be used intensively for fodder production (Porqueddu et al. 2014); they are often part of agricultural areas.
- **Pastures and rangeland**: grasslands that are harvested predominantly by grazing
  Pastures and rangelands constitute the dominant management type within the Mediterranean landscape implying slightly different management schemes: Sylvo- or agro-sylvo-pastoralism

[5] NUTS, *Nomenclature des Unites Territoriales Statistiques*, is a classification system dividing the area of the European Union in three hierarchic levels (NUTS 1, 2, 3). This classification provides the basis for a pan-European cross-border comparison of statistical data (EC No 1059/2003 of the European Parliament and the Council 2003).

consisting of oak trees, shrubs, annual herbaceous species, fodder, winter cereal (Sitzia et al. 2011); pastures of the dairy cattle or sheep system (compared to beef cattle) shows high vegetation in spring caused by higher soil fertility.

The remote sensing-based differentiation of small-scale agricultural management from grassland management of pastures and rangeland is very challenging. Due to the extensive management of these grassland landscapes which is very common in the rural Mediterranean region, general characteristic management features such as distinct time slots for mowing, harvesting or indications of intensive grazing are hard to find (Catorci et al. 2012; Dusseux 2014; Jacques 2014; Louhaichi et al. 2012; Möckel et al. 2015; Salis et al. 2011 und 2015; EUROSTAT 2016: Agriculture and Environment). Generally spoken, early spring is the preferred grazing time for ruminants, when grasslands are vital due to warm temperatures and sufficient rainfall. The ruminants remain grazing until April or until all is grazed out. Depending on the length of summer drought and the general amount of precipitation in autumn and winter, the ruminants will graze again in the winter months or will be raised on a crop-residue, planted fodder or barley grain system (ICARDA/Biradar). It is the nature of extensive pastoralism[6] that it contributes to a sustainable management of grasslands, consequently no significant signs of grazing, respectively management are detectable with remote sensing during the course of the year.

There is only marginal human intervention concerning the management of pastures and rangelands, albeit in some areas, p.e. Sardinia, farmers do clearing cuts at the middle of February/March in order to stimulate plant growing in the upcoming grazing season (Porqueddu et al. 2016). Since livestock density in the rural areas is highly variable, too, even the grazing scheme and its intensity changes in timely and regional aspects.

2.3.3.3.5 Challenges for mapping grassland in the Mediterranean region by means of EO data

The detection of grassland by remote sensing is challenged by

- the heterogeneity of the physical landscape
- the heterogeneity of the Mediterranean climate plus high annual variability
- the heterogeneity of the farming systems
- the dry conditions in the Southern and arid areas
- abnormal conditions like droughts
- the problem of abandonment: due to the reduced number of livestock, rangeland risks to end up in widespread bush encroachment (Landau et al. 2000; Bernués et al. 2011) which could hardly be identified as grasslands

Optical data detect the photosynthetic active parts of the plant and thereby capture the vitality of vegetation. Thus, detection of grassland works best in its active growing period, but it shows high limitations in periods of degradation and drought. The selection of adequate time slots is the focal point in using optical data.

For temperate humid and sub-humid areas, the situation is similar to that in Central European countries: the time slot for detection will start in late spring/early summer, continuing until autumn. Due to sufficient supply of water, vegetation period is more influenced by temperature which means that data base ranges from April until September/October when temperature goes above 12°C. For those areas, the original methodological approach for the classification of grasslands has proven to be best practice.

In dry or arid areas however, the growth of vegetation depends essentially on the sufficient availability of water. Growth stops, plants die or wither and stop their photosynthetic activity. Thus, resting upon the

[6] According to the EEA, extensive grazing means that the stocking density of grazing livestock doesn't exceed 1 livestock unit per ha of forage area (EUROSTAT 2016).

detection of photosynthetic activity, optical satellite data of arid periods provide hardly any reliable information about the existence of grassland. Depending on the length of the arid period, possible time slots for recording grassland would be February to April and September to November or even the whole winter, when grasslands show high vitality due to high amount of precipitation at that time. At those very early and late times of the year however, the information provided by optical data might be severely limited. Coastal fog and clouds caused by seasonal mesoclimatic weather conditions during the winter months, atmospheric haze and shadow effects induced by the lowered solar zenith angle, significantly reduces the number and quality of suitable satellite data.

As **SAR** is able to act independent from sunlight and atmospheric interferences, SAR data are highly suitable for substituting missing information about vegetation cover. Regarding arid areas, SAR data from October/November and during winter months could give additional information about grassland vegetation cover disregarding atmospheric opacity, clouds or shadows and thus supplement an adequate database for the classification. Due to their ability of detecting texture and structure, SAR data are able to support the identification and classification of specific non-grassland classes which are better detectable using SAR imagery than optical imagery which eases their exclusion from the grassland area. Additionally, SAR coherence can aid in the detection of bare soil which can indicate mowing events of grassland or cultivation and ploughing of grassland areas and therefore their conversion into cropland. In this regard, SAR data and SAR classification provides high potential (see previous chapters).

2.3.3.3.6 Conclusion

Summarizing the main findings of this study, the following adoptions of the methodological approach for the mapping and detection of Mediterranean grasslands are recommended:

- Based on the bioclimatic conditions, the methodological approach has to be adapted for those areas showing a **prolonged arid period or summer drought**. That is the case for most Southern areas and Western coastal regions of the Mediterranean basin. There are Mediterranean subtypes of the *Temperate* climate classes which can be found in the hinterland of the Eastern part (Rivas-Màrtinez et al. 2004 and 2011; Peel et al. 2007) and may also show locally dry seasons and arid periods but not as large-scale as the *Mediterranean* one.
- Due to limitations in identifying dry or degraded vegetation with remote sensing methods, the methodological approach should focus on those time slots where grassland shows high photosynthetic activity and the most vital and dense plant cover, being **February to April** and **September to November** when precipitation and temperature allow the growing of vegetation. In Southern regions with mild winter temperature above 12°C, the whole winter months could be used for grassland detection.
- Dry vegetation has proved to be problematic because it can hardly be identified and differentiated from bare soil by optical data, consequently the **dry summer months** between May and August (in some areas even longer) have to be handled with care regarding the optical classification.
- In order to get a reliable and adequate data basis for the grassland classification, **SAR-data** could fill information gaps about grassland vegetation cover during autumn and winter caused by clouds, atmospheric constraints or shadowing.
- **SAR data** show high potential in identifying distinct texture and typical structures (as already used within the HR GRA 2015, see chapter 2.3.3.1 HRL Grassland production). SAR classification facilitates the exclusion of scattered trees within agro-pastoral landscapes, shrubland formations or the regularly structure of olive orchards and horticulture as well as of cropland areas which results in a significant enhanced grassland.

The advantages of optical classification concerning the detection of typical vegetation parameter of grassland and a recording time adjusted to the specific plant phenology in the Mediterranean region could well be complemented by an intensified involvement of SAR classification. These adoptive measures are

highly recommended for a suitable and accurate grassland detection regarding the specific conditions in the Mediterranean region.

2.3.4 Agriculture

Satellite remote sensing is an undisputed source of land information for a vast range of users at all geographical scales. The gap between remote sensing data producers and map users is increasing, enhanced by the fact that spatial data infrastructures are making a great volume of geographic information widely available; therefore, it is important to understand the various concepts and constraints underlying cropland mapping in the context of agricultural statistics. This is particularly critical in light of the fact that in agricultural surveys, land cover maps are often used to support stratification at the sampling design level. Indeed, simple cropland maps or more specific maps depicting cropping intensity can significantly reduce the sampling variance or the ground sampling effort and associated costs.

Land cover maps can highlight the non-agricultural strata that are not to be sampled, or the strata that could be sampled differently. As illustrated by Delincé (2015), if a non-agricultural stratum covers one third of the administrative area of interest, the reallocation of the entire sample to the remaining strata – including cropland areas – will provide a relative stratification efficiency of 1.51 at almost no cost. The efficiency of stratification clearly depends on the relevance of the land cover map selected for the stratification.

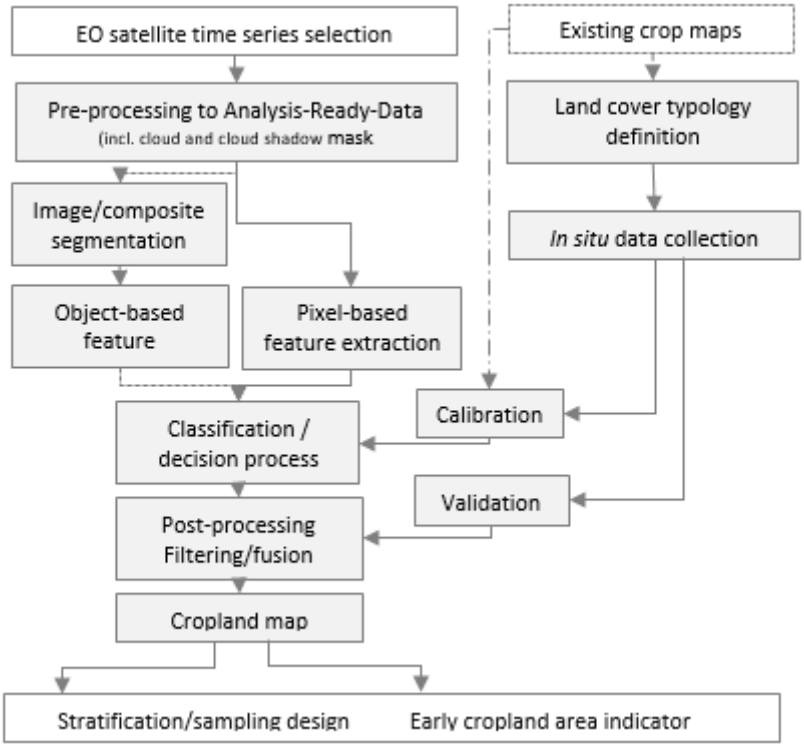


Figure 2-11. Workflow for cropland mapping from satellite observation time series. (Dashed lines correspond to alternative pathways).

This section reviews some key elements of the crop mapping process, as organized according to a standard workflow (Figure 2-11). The first steps of this process consist in the selection of the appropriate land cover typology, the collection of the *in situ* data and the acquisition of the remote sensing imagery. The digital exploitation of these satellite images requires a sequence of standard operations to be completed carefully, and thus derive an accurate cropland map. As land cover maps are readily available for some regions, the relevance of existing maps to agriculture will be discussed systematically on the basis of a set of well-defined criteria.

2.3.4.1 The concept of land cover for cropland mapping

Land Cover Classification Systems (LCCS and LCML)

To ensure full interoperability between typologies and provide common grounds for describing land cover, FAO developed the Land Cover Classification System (LCCS) as a conceptual framework for legend definition. Through a dichotomous modular-hierarchical system based on several sets of descriptors, namely the classifiers, the FAO-LCCS tool aims to explicitly clarify each land cover class, and therefore enables translating from one typology to another (Di Gregorio and Jansen, 2000; Di Gregorio, 2005). More recently, the LCCS framework has been modified into the Land Cover Meta-Language (LCML), to improve its flexibility with unbounded classifiers and a richer class description. The LCML is an object-oriented classification system in which each land cover feature is characterized by a series of elements that can be further detailed by a set of attributes.

For the sake of clarity, transparency and intercomparison, it is internationally recommended to use the LCML framework to define any given land cover typology prior to conducting any mapping effort. For instance, the recent land cover Globland30, which was delivered in 2014 thanks to highly intensive and comprehensive efforts, poorly defined the land cover classes related to agriculture; this seriously curtailed its use for many agriculture and livestock applications.

Agriculture in land cover typology

In the context of agricultural statistics, the stratification definition used for the sampling design relies primarily on the land cover classes related to agriculture. It is noteworthy that cultivated land is not, strictly speaking, a land cover class, but rather a land use class. For example, the land cover of a cereal field is more precisely a dense herbaceous vegetation, while only its land use should refer to agriculture or cropping activity. However, all existing land cover typologies integrate agriculture-related classes because of their importance for the landscape structure and for map users.

While agriculture may at first seem to be the easiest 'land cover' class to map for, this is a major source of misunderstanding and discrepancies between existing land cover maps, even when simply considering cropland and no cropland. This situation is exacerbated when considering the vast diversity of agricultural lands throughout the world, from double-cropping rice fields in Asia to the Mesoamerican traditional *milpa* intercropping system, from the European fallow lands to African perennial plantations such as cacao under the forest canopy.

The World Program for the Census of Agriculture 2020 (vol. 1, p. 82) proposes the following definitions, obtained by aggregating LCML classes:

(i) **Arable land** is land that is used in most years for growing temporary crops. It includes land used for growing temporary crops during a twelve-month reference period, as well as land that would normally be so used but is lying fallow or has not been sown due to unforeseen circumstances. Arable land does not include land under permanent crops or land that is potentially cultivable but is not normally cultivated. Such land should be classified as "permanent meadows and pastures" if used for grazing or haying, "forest and other wooded land" if overgrown with trees and not used for grazing or haying, or "other area not elsewhere classified" if it becomes wasteland.

(ii) **Cropland** is the total of arable land and land under permanent crops.

(iii) **Agricultural land** is the total of cropland and permanent meadows and pastures.

(iv) **Land used for agriculture** is the total of "agricultural land" and "land under farm buildings and farmyards".

Based on the LCML framework, Di Gregorio (2013) established a precise and comprehensive cropland nomenclature to define cropland. However, in the context of agricultural statistics, the definition may raise additional questions, such as the fact that the cultivated area of interest is neither the sowed surface nor

the harvestable one, but rather the area actually harvested. This is not only a semantic discussion for researchers, as the differences can be large in case of drought or floods.

Other than this important discussion, the land cover typology must be workable and compatible with the source of data. For satellite remote sensing, the Joint Experiment for Crop Assessment and Monitoring network (JECAM) endorsed a definition for annual cropland due to the annual nature of the Earth Observation time series: "the annual cropland from a remote sensing perspective is a piece of land of minimum 0.25 ha (min. width of 30 m) that is sowed/planted and harvestable at least once within the 12 months after the sowing/planting date. The annual cropland produces an herbaceous cover and is sometimes combined with some tree or woody vegetation."

The focus on annual cropland is more precise from a mapping point of view, and enables dealing with inter-annual changes of land cover, due for example to cropland extension or the abandonment of cultivated lands.

It is important to note that the definition adopted by JECAM also includes the concept of the Minimum Mapping Unit (MMU), which defines the smallest unit to be considered in the mapping process. For example, the mapping process of the EU's CORINE Land Cover Database was initially set at 25 ha, thus considering only landscape features larger than 25 ha. Such a specification may lead to the discarding of small fields scattered in an urban or forest landscape, which may induce a significant bias in the resulting agricultural land map.

Alternative approaches for land characterization

Other initiatives, driven by well-targeted objectives, focus on the delivery of single land cover class products or binary masks. For instance, the global croplands extent was derived from multi-year 250-m MODIS time series using a set of 39 metrics to depict cropland phenology and to derive a global per-pixel cropland probability layer using a global classification decision tree algorithm (Pittman *et al*., 2010). Hansen *et al*. (2013) obtained a bare soil/no bare soil map at global scale by processing the full archive of Landsat data since 2000 for its tree cover product. All of these initiatives offer the advantage of providing a map product that is focused on the land cover class of interest. Conversely, a major drawback is the absence of any concern for complementarities between products, which may lead to significant spatial inconsistencies or semantic incompatibilities.

The retrieval of biophysical variables from satellite time series results in a quantitative description of the land surface thanks to empirical regression or to physically-based model inversion. Indeed, remote sensing products corresponding Leaf Area Index (LAI), fraction of Absorbed Photosynthetically Active Radiation (fAPAR), albedo, etc. provide direct estimates of undisputable variables that can also be measured on the ground. The seasonal evolution of these biophysical variables can characterize the land surface, and could sometimes be interpreted in agricultural land cover classes of interest or directly used for stratification. However, the capability to identify these biophysical variables from high-resolution, free and open-access satellite imagery, such as that provided by Sentinel-1 and -2, has developed only very recently. The time series available since years at coarse spatial resolutions (250 m to 1 km) are only useful for stratification purposes in certain agricultural landscapes, which either have very large field sizes (as typically occurs in Argentina, Ukraine, the United States of America, Russia, etc.), or with uniform and non-fragmented landscapes comprising many small but similar fields cultivated according to a same crop calendar (e.g. in the North China plain or in case of irrigated rice plains).

2.3.4.2 Image processing and cropland map production

Any land cover map production consists of a sequence of main processing steps. For each of these steps, several conceptual and algorithmic choices are possible. Waldner *et al*. (2016) have shown that crop mask

accuracy varies more from one agricultural region to another rather than from one state-of-the-art method to another. Clearly, certain methodological choices may be more appropriate than others; however, ultimately, the quality and quantity of the remote sensing input and of the calibration data set play an even more important role, in most cases. The key to success is probably the adequacy of the methodological choices adopted for a given quantity and quality of input Earth Observation and in situ calibration data, and with regard to the landscape characteristics to be mapped.

As introduced in figure 1, four main steps in the land cover production chain may be clearly identified: (1) image segmentation; (2) feature extraction; (3) classification; and (4) postprocessing, including filtering and/or fusion.

Image segmentation

The land is discretized into pixels by satellite imagery, while on-screen visual interpretation delineates homogeneous patterns. An image raster made of pixels and a vector made of objects are the two main conceptual models designed to describe the spatial dimension of the world. When the spatial resolution is close or larger than the size of the land cover elements to be mapped, land cover information is generally extracted at the pixel level and the segmentation step is not necessary. For VHR or high-spatial-resolution imagery providing pixels much smaller than the land cover elements, the vector model is usually preferred and the image should be segmented into objects by means of image segmentation algorithms.

Image segmentation groups adjacent pixels into spatially continuous objects according to their spectral characteristics and their spatial context, aiming to capture meaningful spatially discrete land objects. The object-based approach is well-adapted to image texture extraction, has intrinsic contextual information avoiding a salt-and-pepper effect in the classification output, and supports multiscale interpretation thanks to hierarchical or multilevel segmentation (Radoux and Defourny, 2008). On the other hand, this step is also an additional source of error compared to the pixel-based approach. As explained above, it is mostly recommended to proceed with object-based classification when the pixel size is much smaller than the landscape elements. Typically, metric and decametric images are often segmented into objects, while hectometric-resolution images are not. In exceptional cases, pixel- and object-based production chains have been designed; consider the interactive production of the GlobeLand30 land cover map (Jun Chen *et al*. 2015).

Feature extraction

The feature extraction step consists in computing, from the remote sensing images or time series, the most discriminant variables to be used as input for the classification algorithm. These features may be of various natures: (1) spectral, as the multispectral reflectance or the derived indices, such as the NDVI or any other vegetation, chlorophyll or soil index; (2) temporal, as the minimum, maximum or amplitude of a variable over a given time period; (3) textural, as the local contrast, entropy or any other variable derived from the co-occurrence matrix; and (4) a spatial or contextual variable that is particularly suited to the object-based approach.

Currently, three main strategies may be observed in the field of land cover mapping. First, classical strategies rely mainly on spectral features and, possibly, some simple temporal features based on NDVI time series, considering that these are the sources of all other features in any case. In light of increasingly powerful computing performances and the dissemination of machine-learning algorithms, many remote sensing specialists now consider that “more is better” (in terms of features) and rely on classification algorithms to select the most discriminant ones. Third, knowledge-based strategies aim to integrate external expert knowledge by designing ad hoc features according to the classification target and by retaining only those features deemed meaningful according to experts’ rationale (Lambert *et al*., 2016).

Classification

The classification step consists in one or many numerical processes to finally allocate every pixel or object to one of the classes of the land cover typology. The vast diversity of classification algorithms can be split into two main types: the supervised type, which uses a training data set to calibrate the algorithm a priori; and the unsupervised type, which produces clusters of pixels to be labelled a posteriori as land cover class in light of in situ or ancillary information. More recently, forerunning steps of supervised classification are found very useful and consist in automatic cleaning of in situ training data sets or active learning to build a more efficient training data set, by iteratively improving the performance of the classifier model.

The set of methods used to classify images in land cover classes is constantly expanding and is summarized in Table 2-3 in terms of strengths and disadvantages. A review of these methods was recently completed by Davidson (2016).

Table 2-3. Strengths and weaknesses of algorithms used for large-area classification of satellite image data (based on Gómez *et al*., 2016).

Algorithm	Strengths/characteristics	Weaknesses
Maximum Likelihood (Parametric)	• Simple application • Easy to understand and interpret • Predicts class membership probability	• Parametric • Assumes normal distribution of data • Large training sample necessary
Artificial Neural Networks (Non-parametric)	• Manages large feature space well • Indicates strength of class membership • Generally high classification accuracy • Resistant to training data deficiencies – requires less training data than Decision Trees (DTs)	• Needs parameters for network design • Tends to overfit data • Black box (rules are unknown) • Computationally intense • Slow training
Support Vector Machines (Non-parametric)	• Manages large feature space well • Insensitive to Hughes effect • Works well with small training data set • Does not overfit	• Needs parameters: regularization and kernel • Poor performance with small feature space • Computationally intense • Designed as binary, although variations exist
Decision Trees (Non-parametric)	• No need for any kind of parameter • Easy to apply and interpret • Handles missing data • Handles data of different types (e.g. continuous, categorical) and scales • Handles non-linear relationships • Insensitive to noise	• Sensitive to noise • Tends to overfit • Does not perform as well as others in large feature spaces • Large training sample required
Random Forests (Non-parametric)	• Capacity to determine variable importance • Robust to data reduction • Does not overfit • Produces unbiased accuracy estimate • Higher accuracy than DTs	• Decision rules unknown (black box) • Computationally intense • Requires input parameters (#trees and #variables per node)

Post-processing

Postprocessing operations can improve the classification output thanks to the possibility to apply various filtering techniques or to fuse various classification outputs. First, macroscopic errors can be corrected interactively, as they are clearly identified by systematic visual inspection. Basic filtering operators over sliding window of 3 pixels x 3 pixels or 5 pixels x 5 pixels, such as a majority filter removes the salt-and-pepper effect induced by pixel-based classification. More interestingly, such a majority filter could also be applied to pixel-based classification output using objects obtained by multispectral reflectance image segmentation, thus providing a much smoother land cover map.

Fusion techniques are required to merge outputs from the ensemble classifier. A single output map can be obtained by majority voting either where the ensemble chooses the class on which all classifiers agree (unanimous voting); at least one more than half of the classifiers agree (simple majority); or several classifiers agree (plurality voting). Weighted majority voting can be used when some classifiers are expected to perform better than others, or are weighted by the associated probability or membership of the classification output.

2.3.5 New land cover products

The current known limitations among the available land cover products can be summarized as too low spatial and temporal resolutions, as well as some inconsistencies between the different datasets. The low to medium spatial resolution, ranges from 100m (for Corine Land Cover, CLC) to 1km (for Global Land Cover, GLC) – which is useful for cartographic purposes mainly, as an insight for business intelligence, but not for new thematic reporting concerning urban planning or biodiversity strategy, for example. The lack of guaranteed consistency for all available ancillary data retrieved from national datasets can affect the various classes and nomenclature chosen or the temporal range covered. More importantly the datasets themselves – status layers as well as change layers – can sometimes be an aggregation of national data, which have been produced using various methodologies, from full manual process, semi-automated one to customized mixed of both.

All those summarized issues call for the emergence of new land cover (LC) products, which should exhibit new properties to increase their spatial and temporal consistency. The most obvious improvement should be an important increase in the spatial (expected to be at least down to 30m, even 10m) as well as temporal resolutions (every year) of the status layers and their updates, synonym of quicker deliveries – this should be enable by the design of all Sentinels, if fused data is made obtainable in order to decrease the impact of cloudy skies on the optical image production. Users need tends towards update being made every three years, and possibly every year in the long term.

The increase of temporal and spatial resolutions offered by the Sentinel constellation will also result in a higher thematic accuracy, through the enrichment of the existing classes used in the various LC nomenclatures, available at the moment. There is a real need for a better characterization of the cultivated summer and winter crops, their turn-over from one year to the next, as well as the tree species present in the forest cover. Those quicker deliveries will also de facto lead to a better monitoring of the different kinds of change or transitions from one LC to another. This aims at creating a sixth HRL, focused on the agricultural LC.

The creation of a pan-European HR LC layer will be obtained by merging together all the currently available layers, in addition to this new agricultural layer. This merge constitutes an opportunity to enforce a logical consistency between the current and upcoming thematic products, which are being produced independently, without requiring post-processing to ensure the spatial and temporal coherence.

Previous attempts at mapping land cover at a global or continental scale all suffered from the scarce amount of good quality data available for such a task, as well as the scarcity of reference data or ground-truth data, still valid at the moment of the production.

Three Finer Resolution Observation and Monitoring of Global Land Cover (FROM-GLC) versions have been produced (Gong, et al., 2013) (Yu, Wang, & Gong, Improving 30m global land-over map FROM-GLC with time series MODIS and auxiliary datasets: a segmentation based approach, 2013) (Yu, et al., 2014). However, the first map, FROM-GLC, exhibits an accuracy of 63.69%, at a 30m resolution, with 9 classes – while using a dataset of images dating from 20 years ago at the moment of the creation of the map. The supervised classification was trained on 90000 samples; and those two facts combined constitute the main reasons to explain the low accuracy and the huge amount of manual enhancement needed.

The second map, called FROM-GLC-seg, employed MODIS data, which resulted in a slight improvement of the overall accuracy, at 64.42%, but the same methodology was applied. Finally, the third version, FROM-GLC-agg, was an aggregation of the two previous maps at a coarser resolution, for an accuracy of 65.51%.

A wall-to-wall land cover map at country scale – in this study, France – was produced by the CESBIO based solely on Landsat-8 datasets (Inglada, et al., 2017a) for the reference period 2016. The production is fully automated and uses existing datasets as reference data for training and validation in supervised classification, without further manual enhancement. This processing chain uses the full time series, regardless of the cloud amount, and produces maps with 17 land cover classes while providing a complementary confidence map at pixel level.

The envisioned methodology for this first phase to produce such new products characterizing land cover can be decomposed into two steps:

- The generation of temporally stable objects, which can be defined as a geometric border information, based on exogeneous data, such as Open Street Map data;
- The creation of a layer of temporally variable classes, which uses image segmentation, based solely on the Sentinel-2 time series for the moment.

The most important constraint for those new LC products lays in the consistency and continuity of those with the previous LC products.

3 Testing and Benchmarking

This chapter addresses the testing and benchmarking of the candidate methods identified in chapter 2. The benchmarks concerns first the inputs of classification (section 3.1), i.e. automated reference sampling, compositing methods, indices and time features, and second the time series classification methods by thematic field (section 3.2). For each benchmark, the candidate methods and the benchmarking criteria are described in detail. Then, the implementation and results of benchmarking are presented. Finally, main outcomes and recommendations of the analysis are summarised.

3.1 Input data for classification

During the last decade, supervised classification techniques have replaced unsupervised classification techniques as the prevalent technique for large-area LC/LU mapping with time series data (Gómez et al. 2016). In order to train an accurate supervised classifier the two most important components are a suitable reference data set and a powerful set of discriminative features.

Commonly used supervised classification algorithms cannot cope with the irregular time series of remote sensing data over large areas. This occurs particularly between neighboring satellite sensor footprints due to different acquisition dates and, in case of optical imagery, within one scene due to clouds and cloud shadows (3.1.2). In order to transform the data to input features that can be used directly in the classification, the original time series data is transformed to temporal-spectral metrics, so called time features (3.1.3). Time features do not suffer from missing values and can capture the temporal-spectral characteristics of a given pixel for the separation of land cover classes (Egerov et al. 2015).

The other important component for training an accurate supervised classification model is the labeled training data, a set of data points with known location and land cover class in the area of interest. Nowadays, a lot of ancillary data is available that facilitates sample collection for training data (Gómez et al. 2016), e.g. field crop type data that is provided by European farmers in order to receive subsidies. Also, forest and leave type sample data can be derived from existing land cover maps. Although most land cover classes are relatively persistent over time, the sample quality can still be improved by suitable reference sampling techniques (section 3.1.1).

3.1.1 Automated reference sampling

For persistent land cover classes, such as forest, grassland, arable land or impervious surfaces, it is a common approach to automatically sample training locations and labels from outdated maps. This information can be combined with the predictors or features extracted from up-to-date remote sensing image data, to derive a new training dataset which can be used to produce a new up-to-date LC/LU map. Obviously, such automatically generated training samples contain as well wrong labels due to (i) LC change that occurred between the outdated map and the up-to-date imagery, or due to (ii) samples drawn from stable but in the outdated map incorrectly classified regions. Such erroneously labeled samples can be considered outliers in the training dataset, due to the unusual feature patterns.

So far, most approaches try to minimize the amount of outliers by applying a negative buffer before performing the spatial sampling and therefore, to avoid the selection of samples at LC class borders (according to the outdated map) and by excluding very small polygons (Radoux et al. 2014, Inglada et al. 2017). The assumption is that state-of-the-art machine learning classification algorithms can cope with the remaining amount of outliers. However, it is still desirable to reduce the number of outliers as much as possible in order to obtain the best possible model quality. That is particularly relevant when a larger number of wrong samples remain in the sampled dataset with the above methods.

Since outliers are a common problem in many real world datasets, several machine learning algorithms exist to solve the problem. The selection of potential methods and analysis of their performance for additional data cleaning has been evaluated and is shown in the following subsections.

3.1.1.1 Description of candidate methods

For the problem of cleaning automatically generated training datasets for large area remote sensing classification problems, the algorithms should be efficient for large sample sizes, should work well for high-dimensional datasets and should deal with complex unknown distributions. The Isolation Forest (iForest) is a promising state of the art approach that fulfils all these properties (Liu et al. 2008). Additionally, it does not require the features to be scaled and is not very sensitive to parameters leading to overfitting or underfitting. It can be assumed that, as in the case of the frequently used Random Forest classifier (Breiman 2001), good results can be achieved with default parameters. The latter aspect is particularly important for the outlier detection because, in contrast to the case of a supervised classification task with reliable labels, tuning of parameters would be a non-trivial task.

The performance of the iForest was compared to the One-Class Support Vector Machine (OCSVM) (Schölkopf et al 1999), a Support Vector approach that is suitable for outlier detection with high dimensional datasets and complex non-linear class distributions. It is worth mentioning that the Support Vector Data Description (SVDD), another frequently used method for outlier detection, is similar to the OCSVM an when used with a Radial Basis Function Kernel gives the same solution than the OCSVM (Tax & Duin 2004).

3.1.1.2 Benchmarking criteria

The most important benchmarking criteria is the error rate of the outlier detection approach, i.e. the fraction of false positives (outliers that are not identified as such) and false negatives (inliers that are identified as outliers). Apart from the **threshold-specific** performance, it is worth to investigate threshold-independent performance of an outlier detector. Most outlier performance models are able to return a continuous valued decision function instead of a binary decision or prediction (inlier/outlier). The binary decision is simply the result of a (default) threshold applied on the continuous decision function. Thus, given a threshold, all samples with decision function values larger than the threshold are considered inliers and all samples with decision function values smaller than the threshold are considered outliers. Here the kappa coefficient is used as threshold-specific performance measure. A common **threshold-independent** performance measure is the area under the ROC (Receiver Operation Characteristic) Curve (AUC). It can be considered a relative measure for the **potential outlier detectability**. In other words, the higher the AUC the better the achievable outlier detection given that the suitable threshold can be found. Taking into account both threshold-specific and threshold-independent performance measures is important to get more comprehensive picture of the strength and weaknesses of an approach. For example, let us consider the threshold-specific results of an iForest result with a non-optimal-threshold and OCSVM result with a non-optimal-threshold. It is possible that the OCSVM is better than the iForest. At the same time it is possible that the iForest is better than the OCSVM given the most suitable threshold is used for both. In such a case, it can be eventually be concluded, that the threshold selection algorithm has to be improved but not the algorithm used to derive the continuous decision function values. Thus, considering threshold-independent and threshold-specific results allows a more comprehensive assessment of the approaches and strengthens the conclusions and potential improvement measures to be taken eventually.

Most outlier detection algorithms require a user-defined parameter that defines the assumed fraction of outliers in the data set (Tax 2001). Of course, in many applications it is not only unknown *which are the outliers in the dataset* but also *how many outliers are in the dataset*. Estimating the fraction of outliers from the data is a difficult problem and needs to be addressed in the future. By now, an important benchmarking criteria to be investigated is the sensitivity of an algorithm with respect to the assumed fraction of outliers, i.e. how much does the detectability performance degrade in case the user-defined outlier fraction assumption deviates from the true fraction of outliers.

It is worth mentioning that in case of the iForest only one model needs to be trained for different assumed outlier fractions. The assumed outlier fraction only influences the value of the threshold, which is used to convert the decision function to binary decisions. In case of the OCSVM the assumed fraction of outliers also influences the decision function itself. Thus, a new model needs to be trained whenever another fraction of assumed outliers is to be considered.

Other relevant criteria for the selection of a suitable approach are (i) the ease of use of an algorithm, i.e. the number of influential parameters and its sensitivity to parameters and if the input data needs to be scaled, and (ii) the suitability of the algorithm for large datasets, i.e. its computational complexity.

3.1.1.3 Implementation of benchmarking

As mentioned above, the fraction of outliers in the dataset is required to be set as a user-defined parameter. In order to investigate the sensitivity of this parameter with respect to the true amount of outliers several datasets with different outlier fractions have been created from a real dataset. This dataset contains the classes non-forest (260 polygons à 9 pixels), broadleaf forest (100 polygons à 9 pixels) and coniferous forest (104 polygons à 9 pixels). The samples of each class have been contaminated by a growing fraction of outliers – defined in 10 steps by increasing the fraction by 0.05 for each step (0.05, 0.1, 0.15,…, 0.5) – from the other classes. For example, in order to create a dataset with a fraction of 0.1 contaminated samples, the features of randomly chosen 10 % of the coniferous polygons have been replaced by the features of randomly chosen 10 % polygons of the broadleaf forest and non-forest polygons. This results in 300 training data sets for the three different classes and the 10 outlier fraction steps. In order to reduce the statistical uncertainty of the results five replications of different polygons are switched. As a consequence 1500 datasets were generated with known outlier fractions and outlier samples on which the outlier detection approaches have been tested.

In the current analysis, only the fraction of assumed outlier parameters (called the contamination parameter) was changed when setting up the iForest and OCSVM (called the nu parameter) models. The other parameters have been set to sensible default values. Particularly, the OCSVM is trained with an RBF kernel and gamma parameter corresponding to 1/#Features, where #Features is the number of features. As mentioned above, the nu parameter is not investigated and therefore not varied. For the iForest, the number of samples and features to draw from, for constructing a base estimator of the forest, is set to 256 and #Features.

3.1.1.4 Results of benchmarking

As mentioned above, the outlier detection approaches are evaluated based on the AUC, a threshold-independent performance measure, and the kappa coefficient, a threshold-specific performance measure.

Comparing the threshold-independent accuracies (AUCs) grouped by class (non-Forest, broadleaf, coniferous) and the methods (iForest, OCSVM) reveal the following interesting insights (Figure 3-1). First, in case of the Non-Forest class both methods are hardly better than a random predictor since an AUC value of 0.5 corresponds to a random prediction. Instead, the AUCs for the other two classes are much higher, thus the outliers can be distinguished from inliers. Distinguishing outliers from inliers is more accurate in case of the coniferous forest type compared to the broadleaf forest type. For both forest classes the performance of the iForest is significantly better than the one of the OCSVM. Particularly, the mean and median AUCs are higher and the variation is lower. The high variation of the OCSVM AUCs in case of the coniferous forest is of particular interest and might be related to a higher parameter sensitivity of the OCSVM.

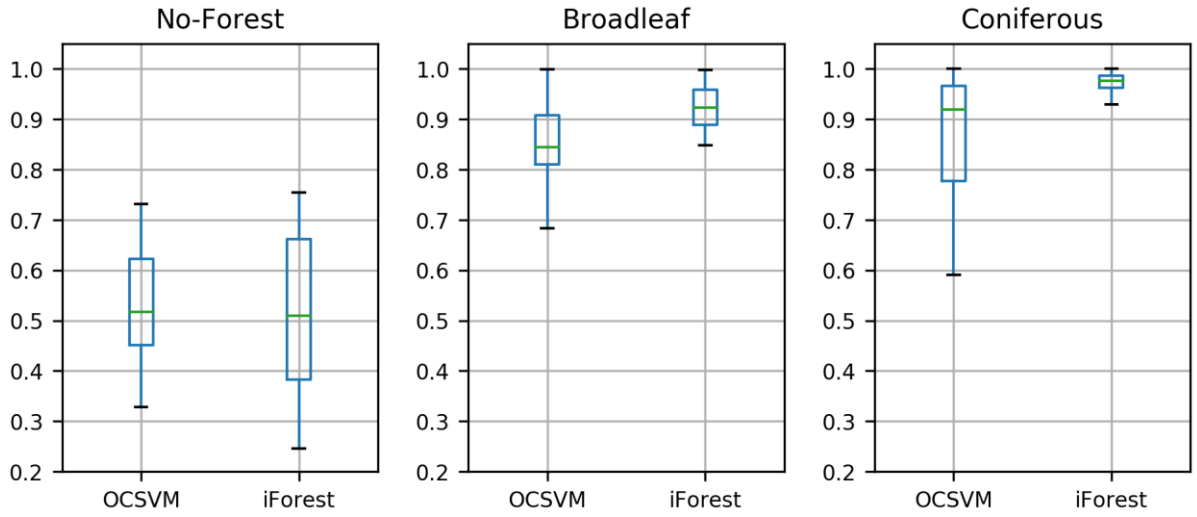


Figure 3-1. Boxplots of AUC values given the class and outlier detection approach achieved over all respective experiments, i.e. varying random replications (5), outlier fractions (10) and assumed outlier fractions (10). Thus, one boxplot is constructed from 500 values.

In case of both methods and all classes the AUC values decrease with increasing outlier percentages (i.e. the outlier fraction multiplied with 100 %) (Figure 3-2). The figure also shows that in case of the iForest, the AUC is constant over the percentage of assumed outliers. This is the case because the AUC is a threshold independent measure that is calculated based on (i) the decision function values and (ii) the above described property of the iForest, stating that the decision function is not influenced by the percentage of assumed outliers (but only the binary decision). An interesting pattern of the OCSVM is that the AUC increases with an increasing percentage of assumed outliers.

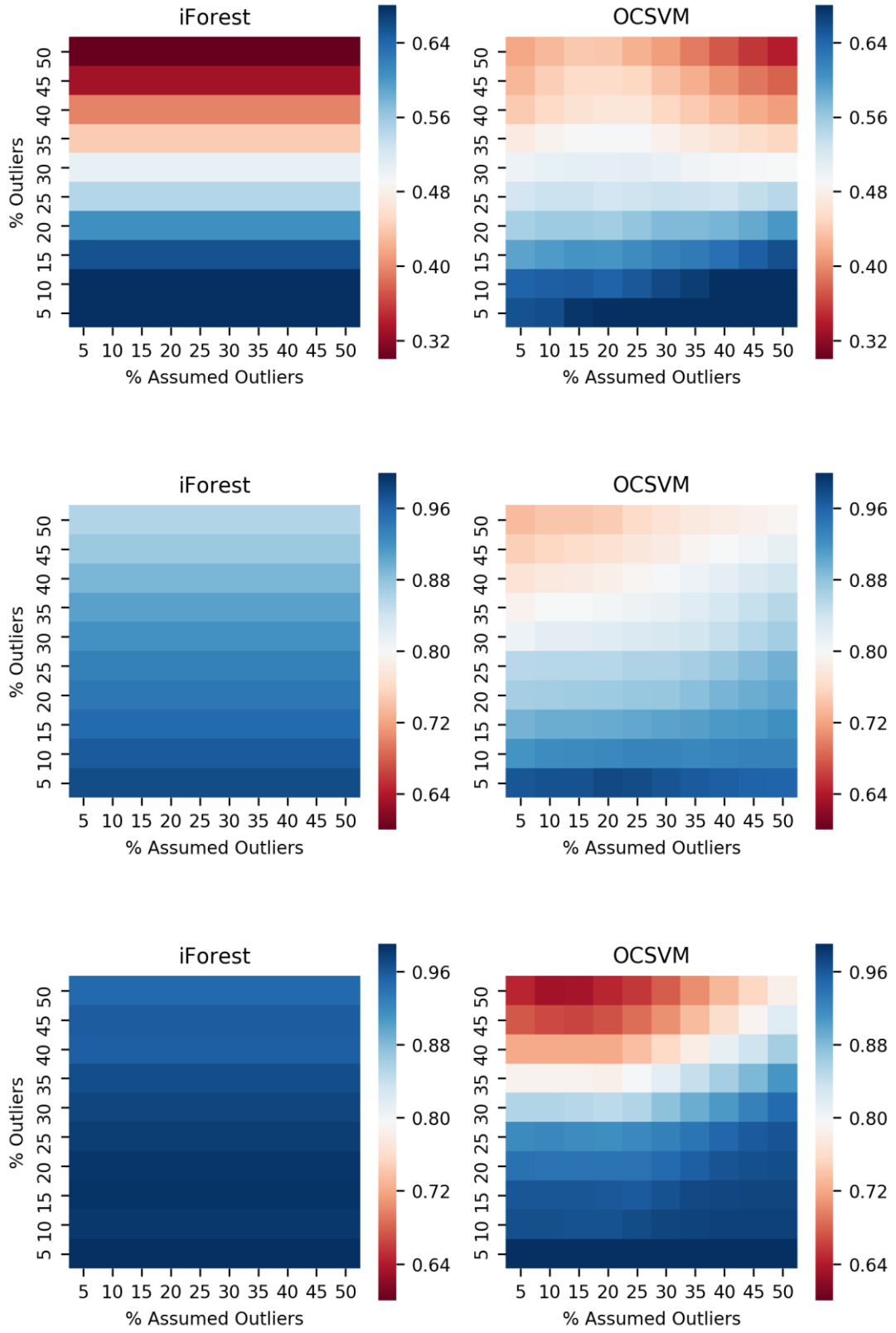


Figure 3-2. Mean AUC for the three classes non-forest, coniferous and broadleaf forest (from top to bottom), the outlier detection approaches iForest (left) and OCSVM (right) dependent on the percentage of assumed outliers (x-axis) and percentage of outliers (y-axis). Each value is the mean AUC of the five random replicates.

The AUC revealed interesting insights in the potential outlier detectability for the different methods and datasets. However, for the actual outlier detection the decision function needs to be converted in binary decision. In case of the OCSVM, where the fraction of assumed outliers is used to train the decision function model, the threshold of 0 is used standardly for the conversion. In case of the iForest, where the fraction of assumed outliers does not influence the decision function, the threshold is selected such that the fraction of assumed outliers is below the threshold. Thus, the threshold is the quantile of the decision function values corresponding to the fraction of assumed outliers.

With the decision function values converted to binary predictions (inlier and outlier) and the true class membership it is possible to derive a confusion matrix containing the classification performance metrics. Cohen's kappa coefficient as threshold-specific performance measure, shows some similar patterns as the threshold-independent AUC (Figure 3-3). The outliers in the coniferous forest class can be better identified than in the broadleaf forest type. In the non-forest class the outliers cannot be identified. It is more important for an accurate outlier prediction that the fraction of assumed outliers does not deviate strongly from the fraction of outliers. This is particularly true for the two forest type classes and the iForest. In case of both forest type classes and both outlier methods, it seems to be favorable – with respect to the to the kappa coefficient – to assume a higher fraction of outliers as it is present in the dataset.

It has been argued before that it cannot always be assumed that the percentage of outliers is known in all applications. For example, when reclassifying up-to-date remote sensing data with reference samples derived from an outdated map there the following two sources of information can help to estimate outliers of the dataset: first, the accuracy assessment of the outdated map and second, the expected land cover change between the target and non-target classes. However, it can also be shown that the histogram of the decision function values can give insights in the percentage of outliers. Figure 3-4 shows the decision value function histograms with different outlier percentages. It is remarkable that with an increasing number of outliers the histograms develop from unimodal and right skewed histograms (with the outliers at the left side) to a bimodal histograms. Thus, as long as the target class is well separable from the rest of the classes (i.e. the outlier samples) the outliers will cluster in a distinguishable mode at the left of the histogram and are separated by a gap between the outliers on the left and the inliers on the right of the histogram. In practice this observation can be helpful when automatically or semi-automatically generating reference samples.

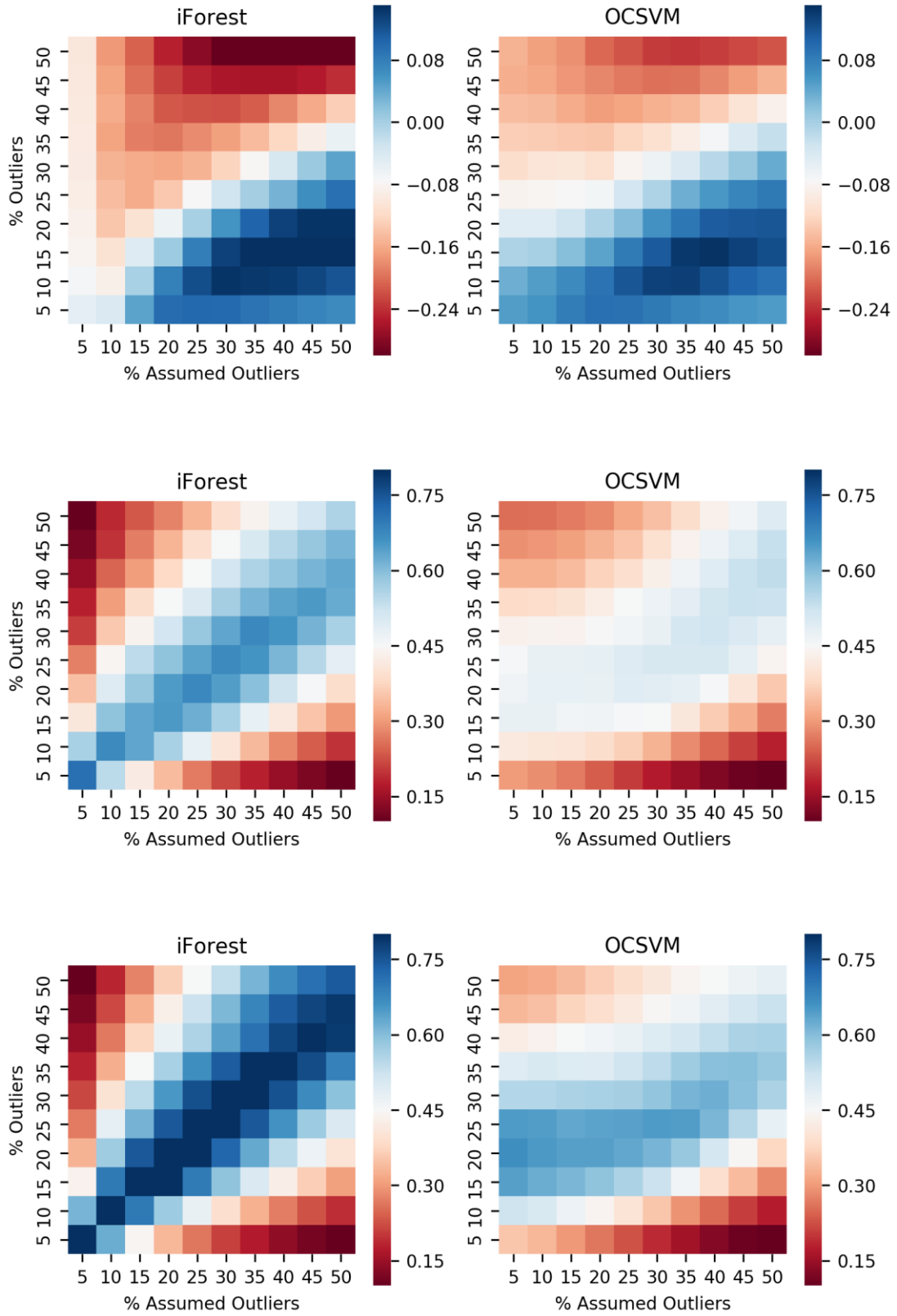


Figure 3-3. Mean kappa coefficient for the three classes non-forest, coniferous and broadleaf forest (from top to bottom), the outlier detection approaches iForest (left) and OCSVM (right) dependent on the percentage of assumed outliers (x-axis) and percentage of outliers (y-axis). Each value is the mean kappa coefficient of the five random replicates.

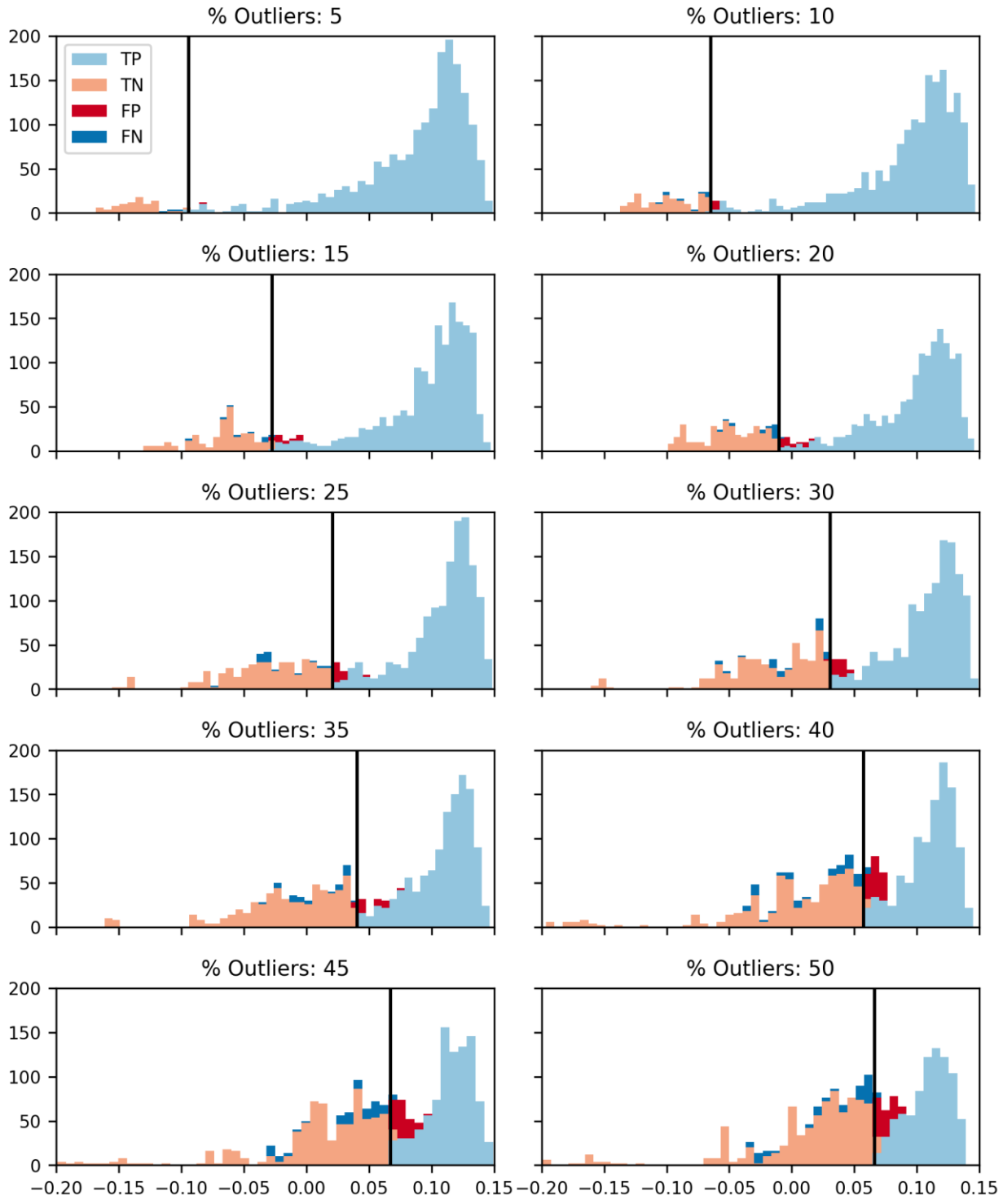


Figure 3-4. Histogram of the iForest decision function values for the coniferous forest class containing different percentages of outliers (see subplot title). The black vertical line shows the location of the threshold when the assumed outlier percentage corresponds to the actual outlier percentage. Given this threshold the colours reveal the true positives (TP), i.e. inliers predicted as inliers, true negatives (TN), i.e. outliers predicted as outliers, false positives (FP), i.e. outliers predicted as inliers and false negatives (FN), i.e. inliers predicted as outliers.

As a consequence of the analysis the iForest turns out to be a powerful and suitable method for the detection of outliers in an automatically sampled reference datasets. Its advantages compared to the OCSVM are:

- High potential separability (AUC) when using default parameters. This is very important since an algorithm that is sensitive to parameters and needs carful tuning is not suitable for outlier detection because there are no known inlier/outlier labels to tune the model.
- The fraction of assumed outliers does not influence the decision function values.
- Scaling the input features is not necessary (according to the literature).

Suitable for high-dimensional and large sample datasets (according to the literature). Particularly, the algorithm can easily be parallelized since each base learner can independently be processed.

3.1.1.5 Summary and conclusions

It has been shown that the iForest, which is able to separate outliers from inliers, exhibits additional important properties valuable for an outlier detection method. It is therefore suitable to be used for such purposes in future applications, e.g. where training samples are sampled from outdated LC maps and used to produce up-to-date maps.

The problem that the fraction of outliers in the dataset needs to be known can be approached by analyzing the histogram of decision function values. If the class of interest is relatively well separable and due to the fact that the assumed outlier fraction does not change the decision function values, a reasonable strategy, to define the threshold, is by analyzing the decision function values with a suitable thresholding approach. A review of potential thresholding approaches and a starting point for further research in this direction is the comprehensive review by Sezgin and Sankur (2004). Another approach, which would not require the binary inlier/outlier decision would be to use the decision function values as instance weights when using the automatically sampled reference data for training a new machine learning classifier. Doing so, the samples that are more likely outliers are assigned to having less weight and become less influential during the model building. In other words, instances (i.e. reference samples) that are more likely inliers (high decision function value) are more influential in the model training than instances that are more likely outliers (low decision function values).

Further research is also required in order to better understand why the outlier detection of the non-forest class failed. It has to be noted that compared to the other two considered classes, this class is an extremely heterogeneous composition of a wide variety of different classes. It is possible that the relatively small amount of reference samples used in this study is not able to well represent such a complex distribution and that the outlier detection can be assumed to improve with a much larger amount of reference samples. Further research in that direction is required in order to increase the knowledge about the potential as well as limitations of outlier detection for different types of classes or distribution characteristics.

3.1.2 Compositing methods on S-2 time series

Spatial continuity and consistency in large scale mapping are important criteria in global and regional vegetation monitoring, land cover change analysis, and land cover mapping activities. The following sections explore methods to reduce heterogeneity in the imagery (different orbits, acquisition dates, cloud/shadow contamination) through temporal synthesis of daily optical satellite observation, i.e. compositing. Various algorithms have been developed to produce a cloud-free synthesis from optical time series, each correcting for angular effects and atmospheric variations differently. In this benchmarking, two main categories of compositing are selected: time interval algorithms and feature-based algorithms.

First, this section describes the candidate methods to be compared (3.1.2.1). Second, the benchmarking criteria are detailed (3.1.2.2). Then, the implementation and benchmarking results are presented and discussed (3.1.2.3). Finally, the main outcomes of the analysis are summarized in section 3.1.2.4.

3.1.2.1 Description of candidate methods

This benchmarking assesses the performance of various compositing approaches applied on land surface reflectance of Sentinel-2 images. Three methods considered are time interval algorithms (Maximum Value Compositing on NDVI, Mean Compositing and Weighted Average Compositing) and two are feature-based algorithms (Knowledge-based Compositing and Quantile Compositing).

Maximum Value Composite on NDVI (MVC NDVI)

This best pixel method selects, for a given compositing period and on a pixel-by-pixel basis, the date of the valid pixel which has the highest NDVI (Holben, 1986). Reflectance values of each spectral bands are retained for each pixel location according to the date selected.

- *pixel value = reflectance value at the date where the NDVI of the pixel is the maximum for the compositing period, for each spectral bands*

Mean Compositing (MC)

This method treats all cloud-free reflectance values as estimates of the signal, and any remaining variability after cloud removal as an unpredictable noise. It consists of averaging all valid reflectance values for each pixel and each spectral band acquired during the chosen compositing period (Vancustem et al., 2007a). The MC algorithm need to fulfill three conditions to be relevant from a statistical point of view: (i) an efficient quality control procedure able to discard any odd value, (ii) an accurate geometric correction, and (iii) a compositing period which is a multiple of the view zenith angle (VZA) cycle of the instrument.

- *pixel value = mean of reflectance values of all valid L2A in the compositing period, for the corresponding pixel, for each spectral band*

Weighted Average Compositing (WAC)

This method averages all cloud-free reflectance values acquired during the compositing period giving more weight to the images closer to the middle of the compositing period in order to enhance the fidelity to the central date (Hagolle et al., 2015). The weighting must be light enough so that it does not finally select only one date, and finally looks like a best pixel method. The weight is computed for each L2A image based on the time difference between the L2A date and the central date of the time series.

- *pixel value = weighted average of reflectance values for each L2A in the compositing period, for each spectral band. The weighting strategy gives a weight of 1 to the central date, and of 0.5 to the first and last date of the compositing period. Weights of L2A images between the beginning/end and the middle of the composite are interpolated.*

Knowledge-Based Compositing (KC)

This feature-based method extracts relevant spectral and temporal features at specific events of the growing season (Matton et al., 2015; Waldner et al., 2015; Lambert et al., 2016). These features are defined according to generic characteristics of crop growth: (i) the growing of crops on bare soil after tillage and sowing; (ii) a higher growing rate than natural vegetation types; (iii) a well-marked peak of green vegetation; and (iv) a fast reduction of green vegetation due to harvest and/or senescence. Five distinct remote sensing stages in the crop cycle are defined at the pixel scale: (i) the maximum value of red; (ii) the maximum positive slope of the NDVI time series; (iii) the maximum value of NDVI; (iv) the maximum negative slope of the NDVI time series; and (v) the minimum value of NDVI. The final spectral-temporal

features corresponded to the reflectance values observed at these stages. A Whittaker smoothing is first performed on the L2A time series and NDVI time series prior to the feature extraction.

- *pixel value - Max. Red = reflectance value at the date of the time series with higher value in red band, for each spectral band*
- *pixel value - Max. NDVI = reflectance value at the date of the time series where NDVI is the highest, for each spectral band*
- *pixel value - Min. NDVI = reflectance value at the date of the time series where NDVI is the lowest, for each spectral band*
- *pixel value - Max. positive slope NDVI = reflectance value at the date of the time series where the gradient of NDVI is the highest, for each spectral band*
- *pixel value - Max. negative slope NDVI = reflectance value at the date of the time series where the gradient of NDVI is the lowest, for each spectral band*

Quantile Compositing (QC)

This feature-based method proposes statistical measures from a multi-temporal stack of good quality satellite observations. Metrics consist of measures derived from all L2A observations. A 0-10 and a 90-100 interval quantile means (mean of all valid observations between the defined thresholds of the quantile) of reflectance values are computed for all spectral bands, based on the distribution of valid NDVI along the time series.

- *pixel value - Quantile 10 = mean of the reflectance values for the dates of the time series with the minimum NDVI values (for each pixel the 10 % of lower NDVI values from the time series are used), for all spectral bands*

- *pixel value - Quantile 90 = mean of the reflectance values for the dates of the time series with the maximum NDVI values (for each pixel the 10 % of higher NDVI values from the time series are used), for all spectral bands*

3.1.2.2 Benchmarking criteria

Five performance criteria are used to assess and compare the compositing outputs. The first criteria is a qualitative analysis, consisting in a visual examination of the composites, and the others are quantitative analysis (temporal consistency, fidelity to medium date image, data gaps and artefacts analysis).

Visual analysis

A systematic visual examination and comparison of the colour compositions (R:NIR-b8, G:Red-b3, B:Green-b2) of the composited products were realized. Qualitative criteria such as the presence of haze, speckle effect and spatial consistency are analysed for each composites of the five methods.

The MVC, MC and WAC are compared using the same compositing period and frequency, namely monthly composites, while KC and QC are compared on the entire time series.

Temporal consistency

This first quantitative analysis evaluates the spectral consistency over time by studying the temporal profiles of the individual reflectance bands coming from stable surfaces for which reflectance is not supposed to vary in the time series.

The samples were carefully selected in order to consider only "pure" land cover pixels. They were selected as much as possible in valid and cloud-free area, i.e. not covered by any cloud/cloud shadows/ambiguous cloud. Three land cover types were selected: water, roof top and bare soil. These three land cover types are represented for the Belgium site, while sufficient areas of roof top for South Africa and of roof top and water for Mali couldn't be find. The samples were manually delineated based on very high spatial resolution images (ESRI World Imagery), the 2012 Corine land cover map for the Belgium site, and the 2014 NLC South Africa map for the South Africa site. One region of interest (ROI) was sampled per land cover type with the following rules: (i) ROIs have to be homogeneous on the orthophotos, and (ii) ROIs are selected at the center of land cover features in order to avoid boundaries effects.

For each date, mean and standard deviations of reflectance values are computed based on all the pixels contained in the ROI, for all spectral bands. Then, in order to better visualize the stability of the over time, standard deviation of the ROI mean values are computed comprising all the composites of the time series, for all spectral bands.

This analysis is only realized for the three time interval algorithms (MVC NDVI, MC and WAC) as their outputs are monthly composites along the time series, allowing a temporal examination, while the features-based algorithms outputs are computed on the entire time series.

Fidelity to medium date image

This second quantitative analysis assesses the fidelity of cloud-free areas of the composites with the medium date image (L2A level) of the composite. The statement behind this analysis is that in a perfect world, the Level 3A synthesis of the middle of the composite period should be identical to a cloud-free Level 2A acquired at that date, if it existed (Hagolle et al., 2015).

As a results, the fidelity criterion is to best mimic the information content of a single cloud-free image considered as reference image. It measures the difference between the composite surface reflectance value and the L2A surface reflectance value for all the cloud-free pixels, when a relatively cloud-free L2A image is available for a date close to the central date of the composite (+/- 8 days). Composites having a high fidelity to the central date allow to have a temporally consistent time series.

The following statistics are computed:

- 70 % percentile: Maximum absolute value of the difference between level 3A (composite) and level 2A (central date), for the 70% of pixels which have the lowest absolute value of difference.
- 95 % percentile: Maximum absolute value of the difference between level 3A and level 2A, for the 95% of pixels which have the lowest absolute value of difference.

The comparison of this fidelity criterion is realized for the three time interval algorithms (MVC NDVI, MC and WAC). The feature-based algorithms (KC and QC) are computed on the entire time series and a fidelity to the middle of the time series would not make sense.

Remaining proportion of data gaps

This third quantitative analysis assesses the remaining proportion of data gaps after the synthesis. It provides the average value, for all the composites of the time series, of the pixels with no value within the

image footprint, and divide by the number of pixels which should have been observed if at least an image had been completely cloud-free.

- Residual gaps = $\frac{\text{number of pixels in data gaps within image footprint}}{\text{number of pixels within image footprint}}$

This analysis is achieved on the five compositing algorithms.

Artefacts

This last quantitative analysis assesses the amplitude of the artefacts observable at the limits of zones obtained with the same set of dates. This is assessed by the standard deviation of the average difference of reflectance values between pixels at the external borders and pixels at the internal border of the contiguous zone. This analysis is achieved on the five compositing algorithms.

3.1.2.3 Implementation and results of benchmarking

The benchmarking is achieve on Sentinel-2 cloud-free images. The implementation has been done on the test sites in Belgium (tiles 31UFR and 31 UFS), in South Africa (tiles 35JMJ and 35JNJ) and in Mali (tiles 29PRP and 29PTU). These three sites were chosen because (i) they all cover various land cover types needed for the spectral consistency analysis and interesting for classification purpose, and (ii) the effects/artefacts of their different cloud coverage can be compared in the compositing outputs.

Composites are generated on a monthly regular basis for the MVC NDVI, MC and WAC along the time series. Seasonal composites are generated for the entire period for the KC and the QC (Table 3-1). Table 3-2 summarizes the compositing periods and tests realized for each method.

Table 3-1. Length of time series per site.

Site	Time series
Belgium	01-01-2017 to 30-11-2017
Mali	01-07-2016 to 30-04-2017
South Africa	01-07-2016 to 30-04-2017

Table 3-2. Tests and compositing periods for the composite benchmarking achieved on the five compositing methods.

Test	**MVC NDVI**	**MC**	**WAC**	**KC**	**QC**
Compositing period					
Monthly regular basis	V	V	V		
Seasonal basis				V	V
Tests					
Visual analysis	V	V	V	V	V
Temporal consistency	V	V	V		
Fidelity to medium date image	V	V	V		
Data gaps	V	V	V	V	V
Artefacts	V	V	V	V	V

3.1.2.3.1 Visual analysis

In this section, the five algorithms are visually examined. First, MVC, MC and WAC are compared together as they are monthly composites, and then KC and QC outputs that represents features computed for the entire time series. Finally, drawbacks and advantages of time interval algorithms and feature-based algorithms are pointed out.

Figure 3-5 Figure 3-5 shows false colour compositions of the composited products of the MVC NDVI, MC and WAC for the three sites in Belgium, Mali and South Africa. At this scale, no large differences are visible between these monthly composites, except that MVC NDVI outputs provide more contrasted outputs compared to MC and WAC.

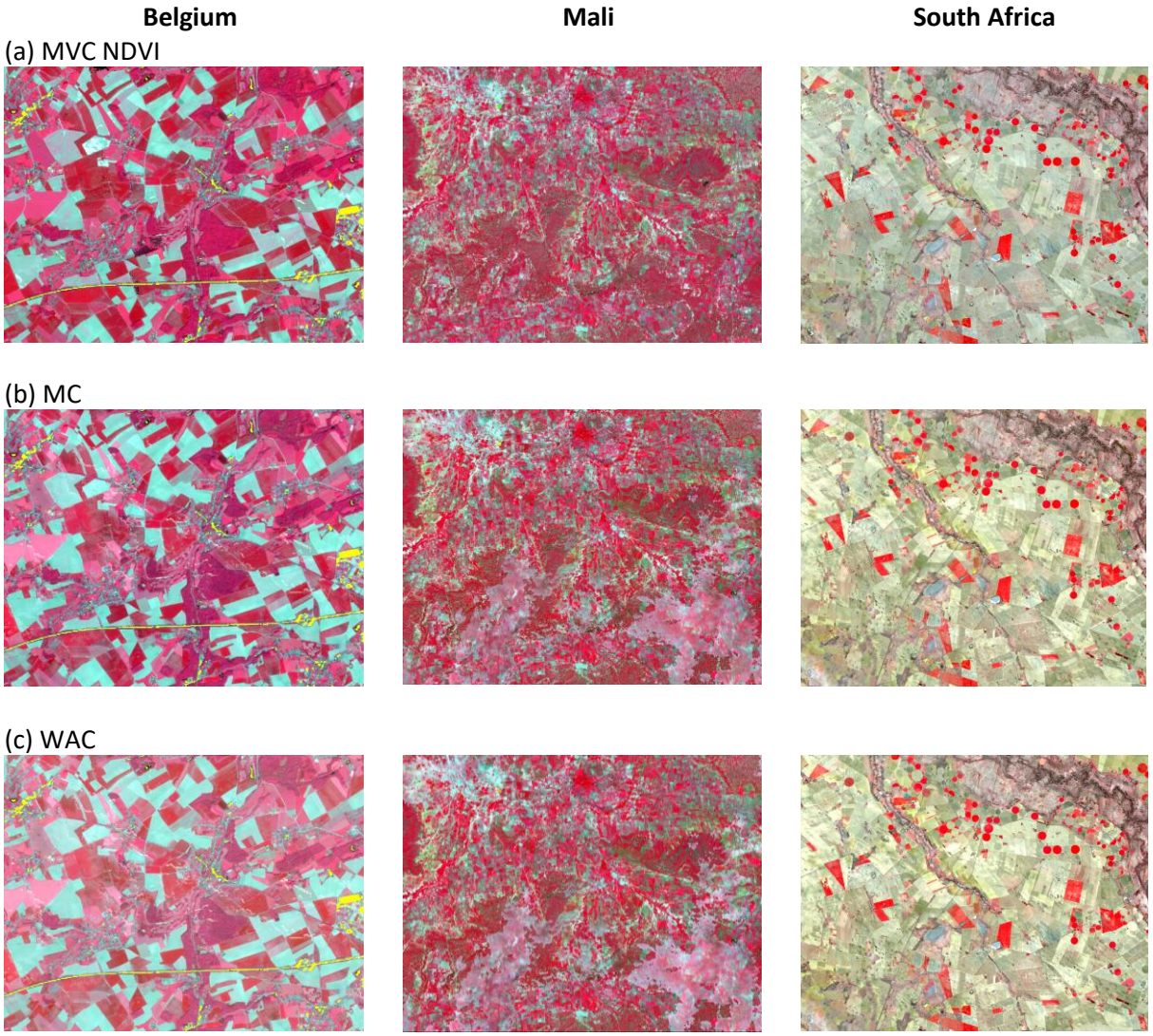


Figure 3-5. False colour (b8, b3, b2) monthly composites over the Belgium site (2017-05), Mali site (2016-08) and South Africa site (2016-09) of the (a) MVC NDVI, (b) MC and (c) WAC algorithms.

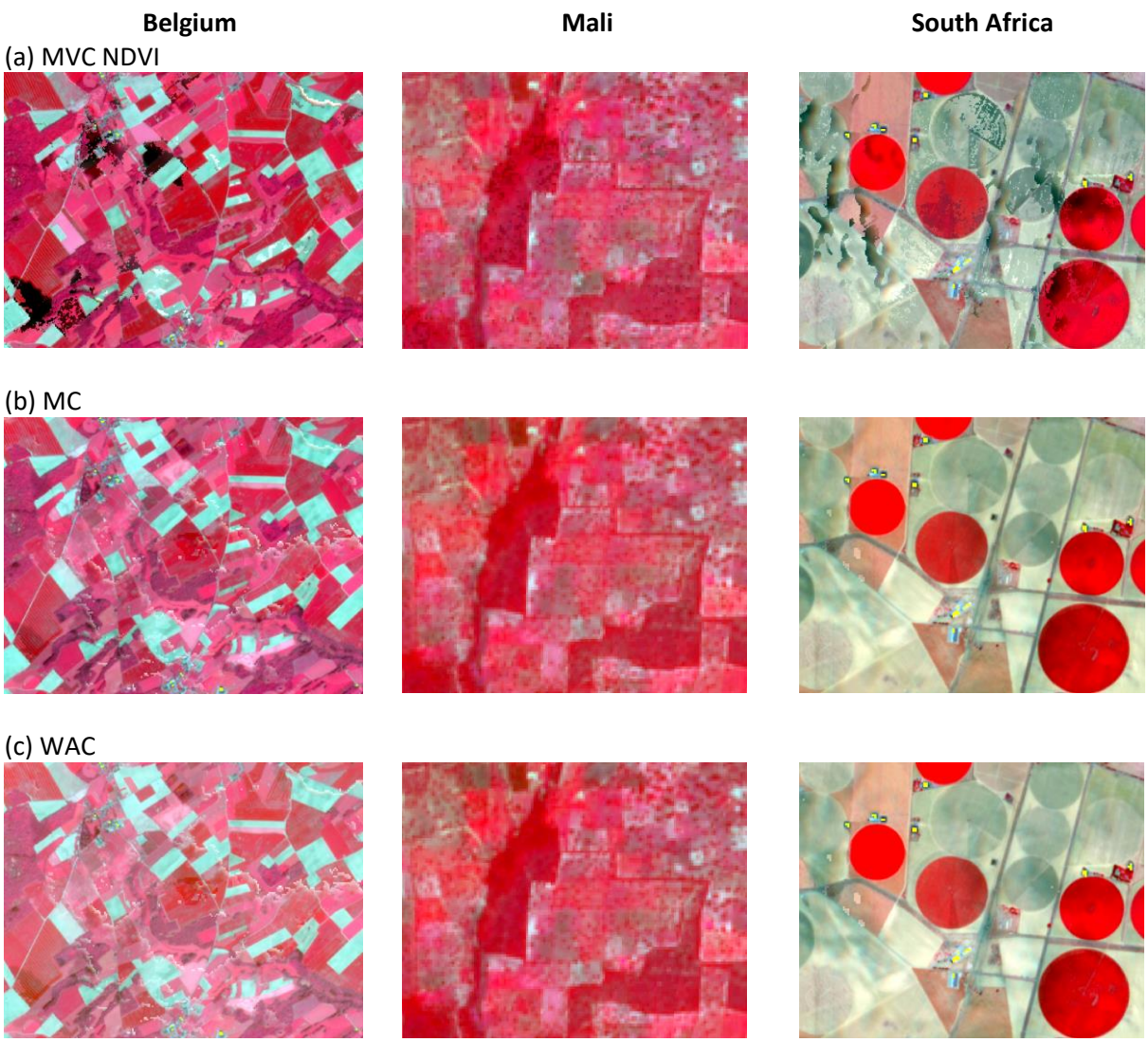


Figure 3-6. False colour (b8, b3, b2) monthly composites over the Belgium site (2017-05), Mali site (2016-10) and South Africa site (2016-10) of the (a) MVC NDVI, (b) MC and (c) WAC algorithms.

However, when zooming at the field scale, such as in **Figure 3-5. False colour (b8, b3, b2) monthly composites over the Belgium site (2017-05), Mali site (2016-08) and South Africa site (2016-09) of the (a) MVC NDVI, (b) MC and (c) WAC algorithms.**

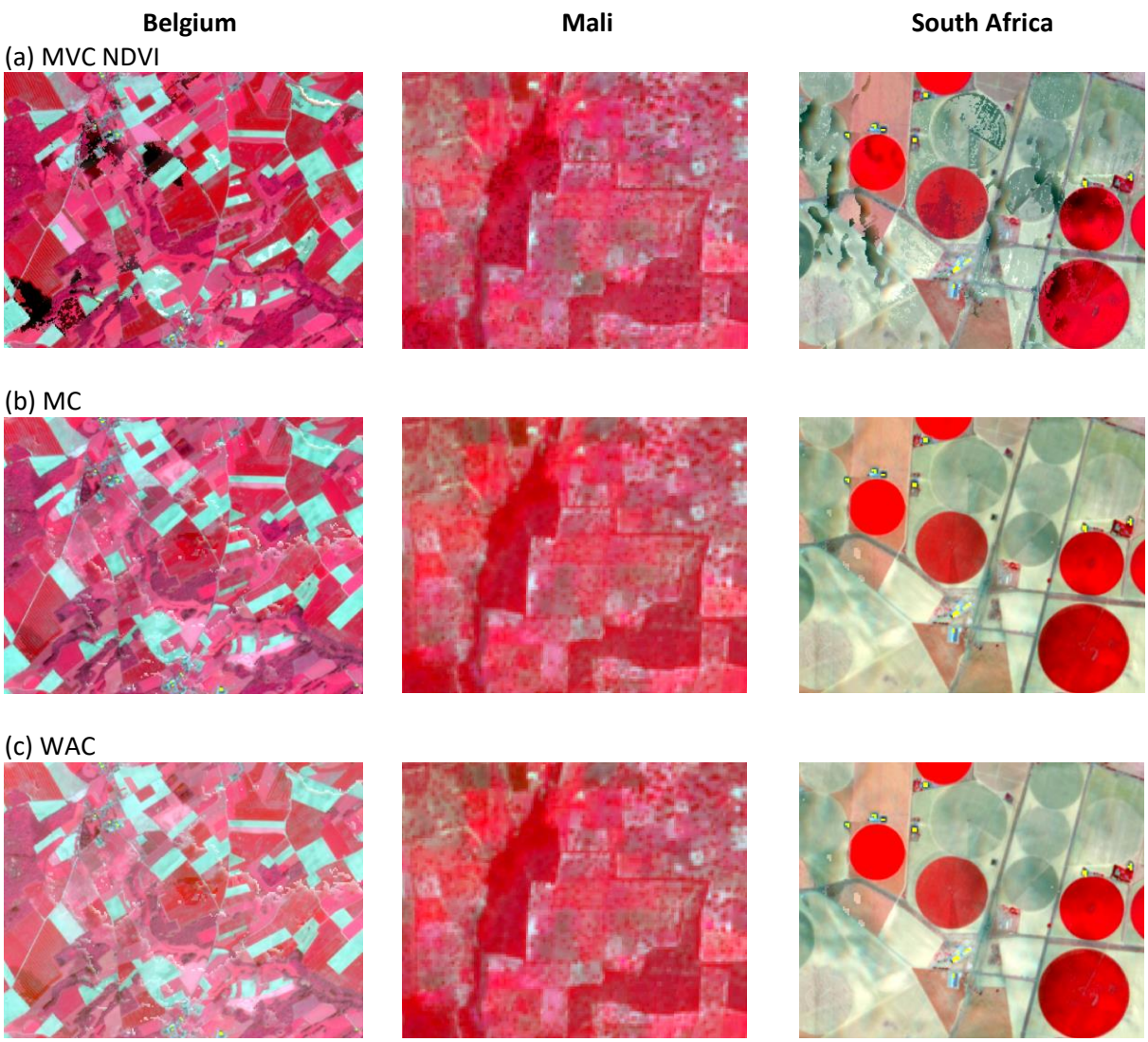


Figure 3-6, strong differences appears. The MVC syntheses exhibit a large noise or speckle like effects. This effect is particularly visible for center-pivot irrigated crops of the South Africa site. This type of compositing, also known as “best pixel method”, only selects one date for each pixel and discards the others. It results in very noisy composites because the selected date for adjacent pixels may have been acquired under different acquisition geometries, or may be affected differently by a cloud shadow or undetected cloud. In addition, surface reflectance may have changed within the compositing period for different dates from one pixel to the other, leading to a noisy image. This noise is not observed in MC and WAC composites. They are designed to minimize artefacts by selecting the largest number of valid points within the available set of dates. As a result, the possibility to observe artefacts when the set of dates changes is reduced. The visual comparison between these two methods shows indeed that MC and WAC strategies produce cleaner images than MVC.

MC and WAC show large similarities, it is not possible to discriminate them visually. The main drawback of these two methods is the sensibility to artefacts of the cloud mask. This is particularly visible in the Belgium site, prone to high cloud cover, in **Figure 3-5. False colour (b8, b3, b2) monthly composites over the Belgium site (2017-05), Mali site (2016-08) and South Africa site (2016-09) of the (a) MVC NDVI, (b) MC and (c) WAC algorithms.**

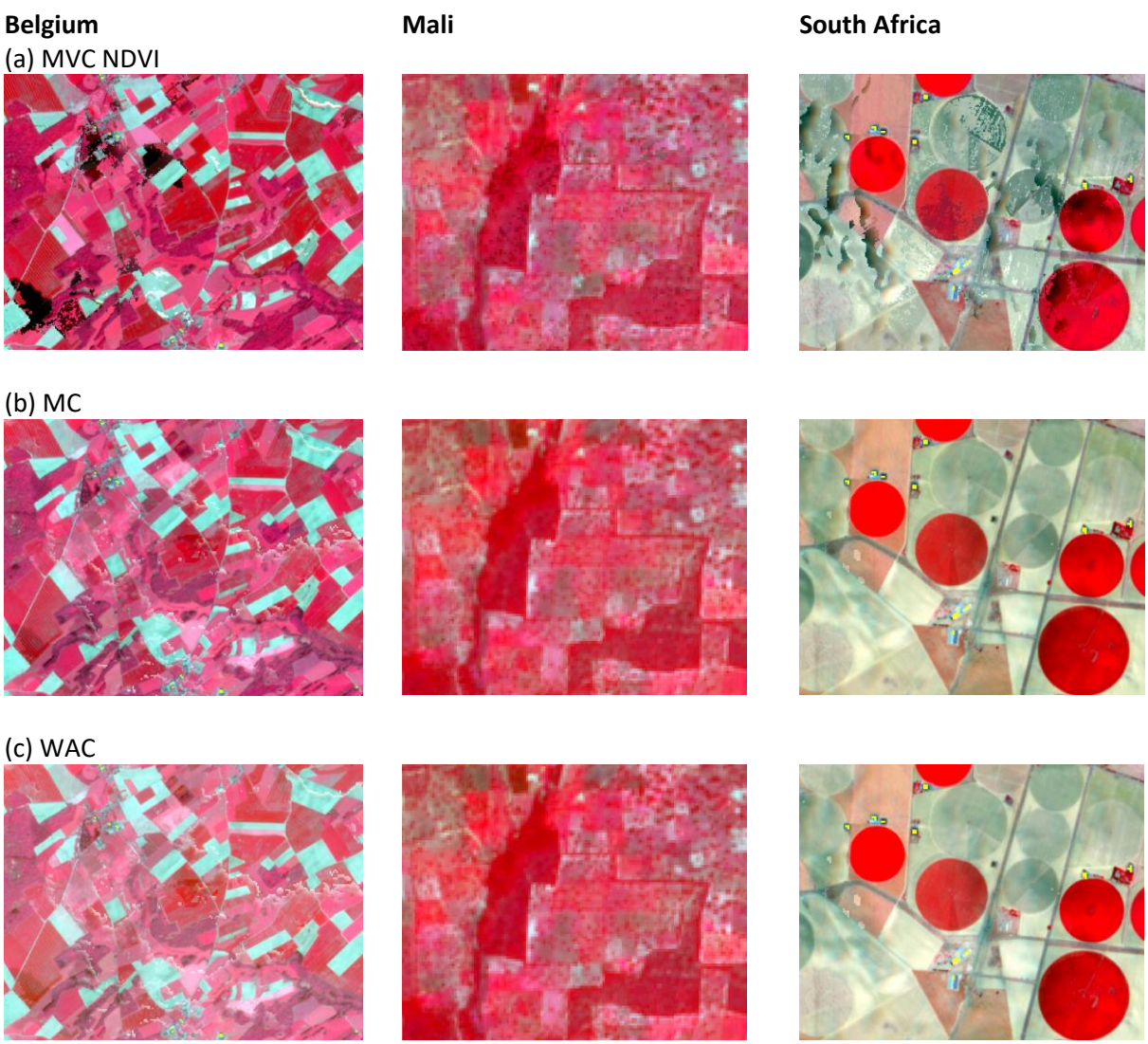


Figure 3-6. If too few images are available (only one or two), which is the case with only Sentinel-2A available until July 2017, and if the cloud mask is not performant enough, artefacts will be visible in the average compositing methods (Figure 3-7). In this case, borders of large clouds are poorly detected, and the remaining haze effects affects the reflectances. Undetected cloud shadows lead to more artefacts in the MVC NDVI products, while it is smoothed in MC and WAC composites (**Figure 3-5. False colour (b8, b3, b2) monthly composites over the Belgium site (2017-05), Mali site (2016-08) and South Africa site (2016-09) of the (a) MVC NDVI, (b) MC and (c) WAC algorithms.**

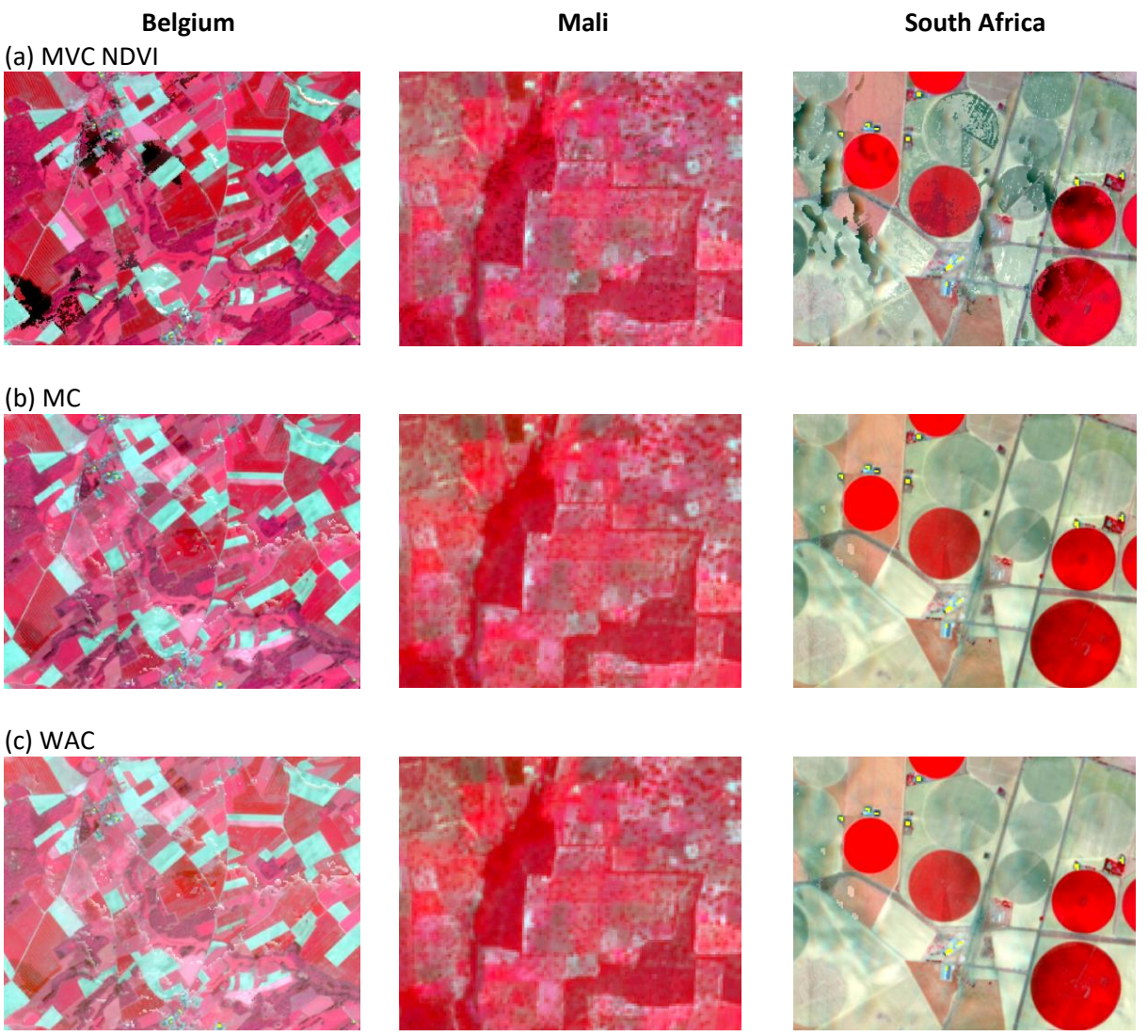


Figure 3-6, Belgium site). Although cloud detection is supposed to be much more accurate when performed at high resolution and with a large diversity of spectral bands including the 1.38 µm spectral band able to detect thin cirrus cloud, the Sentinel-2 cloud mask presents too many artefacts concerning the delineation of clouds borders, the haze and cirrus detection and removal, and the detection of cloud shadows. Improvements are necessary to produce composites with sufficient quality for land cover mapping.

Figure 3-7. False colour (b8, b3, b2) monthly composites over the Belgium site (2017-06) of the MC algorithm, showing strong artefacts due to undetected haze or cloud borders.

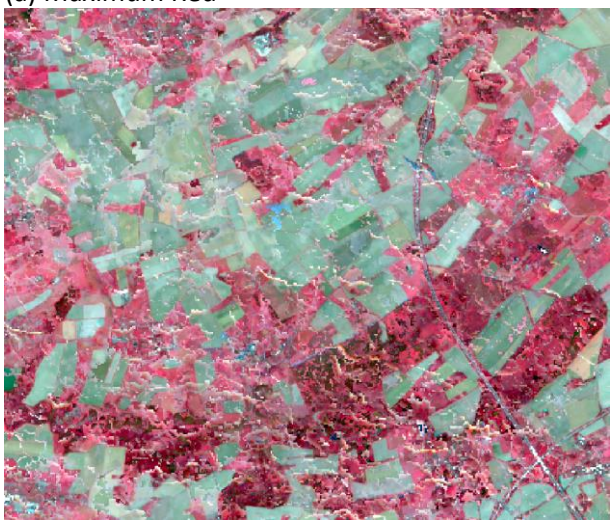
Figure 3-8 and Figure 3-10 show knowledge-based features extracted over the Belgium site and Mali site respectively. They target key phases of the crop cycle such as (a) the bare soil after harvest or before sowing (Maximum Red), (b) the growth rate (Maximum positive NDVI slope), (c) the peak of photosynthetically activity (Maximum NDVI), (d) the green vegetation reduction due to harvest or senescence (Maximum negative NDVI slope) and (e) the minimum vegetation cover (Minimum NDVI). Cropland appears clearly distinct from other classes. Depending on how the time-series cover the crop cycle, specific features tends to give a homogeneous response over the cropland regardless of the crop types (Waldner et al., 2015). The features based on slopes are more sensitive to noise and produce patchy results, which is especially in the Belgium site (Figure 3-8). These phenomena can be a source of additional noise for the classification. Part of the noise is related to the spectral temporal features themselves (Lambert et al., 2016). Spectral-temporal features are based on extreme values and are thus more sensitive to noise, as noise itself is characterized by extreme values.

Compared to KC products, features of QC produce cleaner images, as observed in Figure 3-9 and Figure 3-11. They are mean of all valid reflectance values between the defined thresholds. Thus, the effects due to extreme values is smoothed. No particular artefact is visible on these two quantiles. However, other quantiles could be computed to get more inputs for classification algorithms.

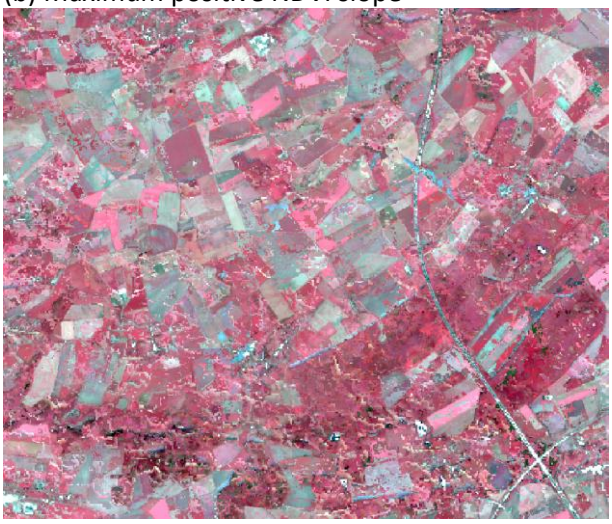
Finally, Figure 3-12 compares the outputs of the five algorithms, considering the beginning of crop season and the middle of crop season for the Belgium site. If the time interval algorithms provide regular composites, in this case each month, it is clearly observed that it can result in partly or totally unusable product as input for a classification due to cloud cover. On the contrary, being computed on the entire time series, feature-based algorithm provide fewer inputs but of better quality. Also, due to their smaller compositing period, time interval algorithm are much more sensible to cloud mask artefacts, as visible in the composites of middle crop season in Figure 3-12. Concerning the beginning of crop season, when bare soils are still present, quantile 10 of QC performs better than Maximum Red of KC.

KC - Belgium

(a) Maximum Red



(b) Maximum positive NDVI slope



(c) Maximum NDVI

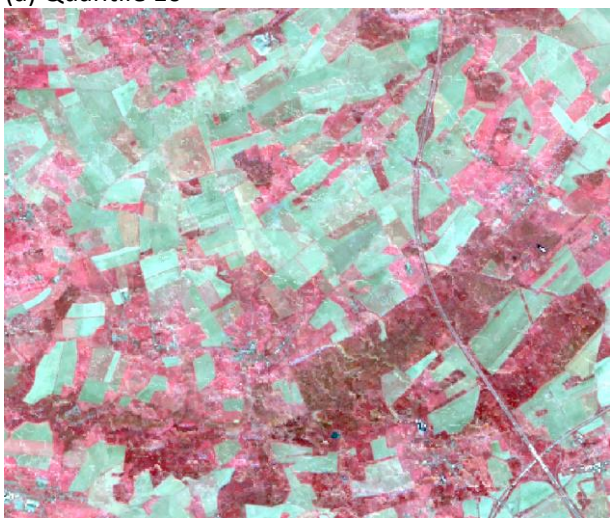
(d) Maximum negative NDVI slope

(e) Minimum NDVI

Figure 3-8. False colour (b8, b3, b2) knowledge-based features over the Mali site: (a) Maximum Red, (b) Maximum positive NDVI slope, (c) Maximum NDVI, (d) Maximum negative NDVI slope and € Minimum NDVI.

QC - Belgium

(a) Quantile 10



(b) Quantile 90

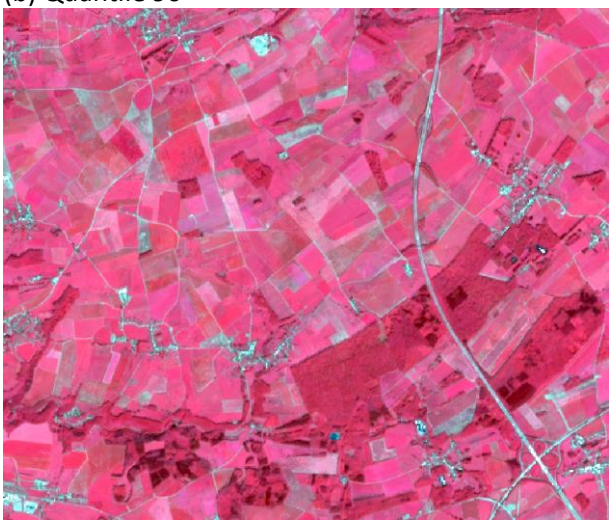
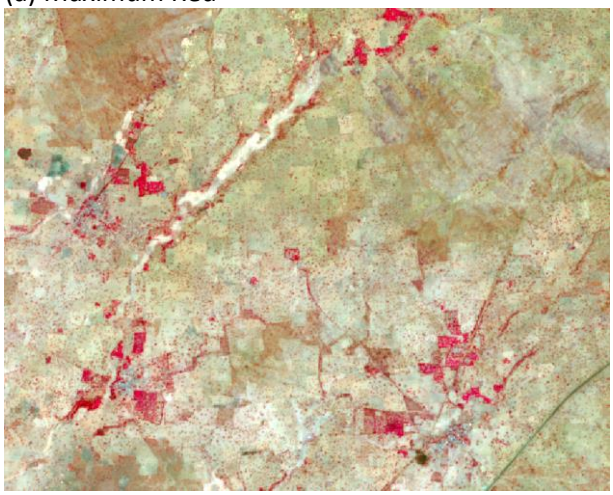


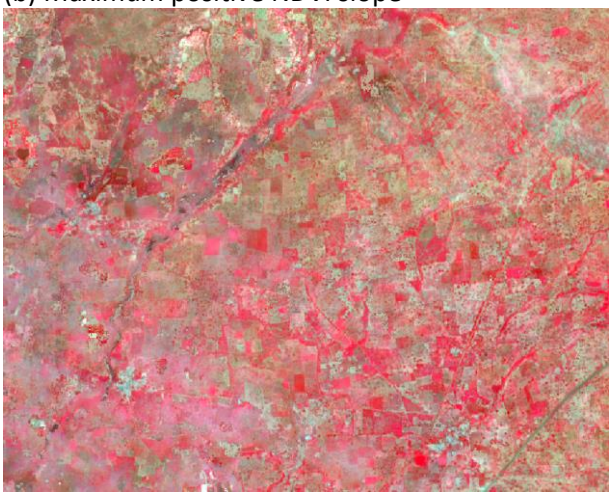
Figure 3-9. False colour (b8, b3, b2) quantile compositing features over the Belgium site: (a) Quantile 10 and (b) Quantile 90.

KC - Mali

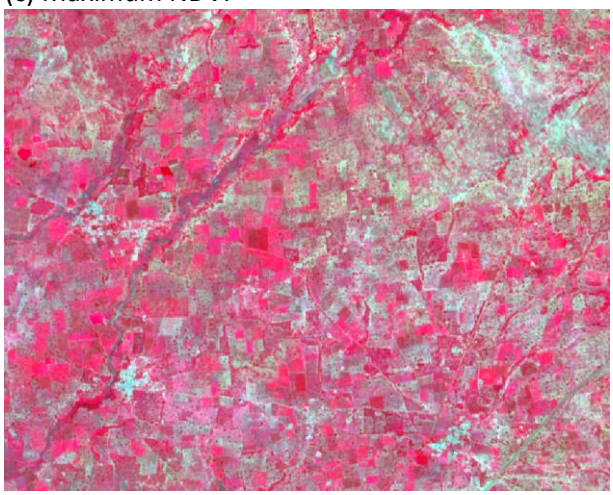
(a) Maximum Red



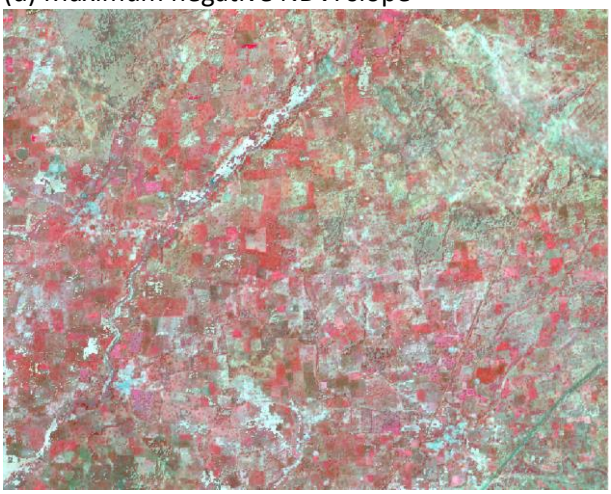
(b) Maximum positive NDVI slope



(c) Maximum NDVI



(d) Maximum negative NDVI slope



(e) Minimum NDVI

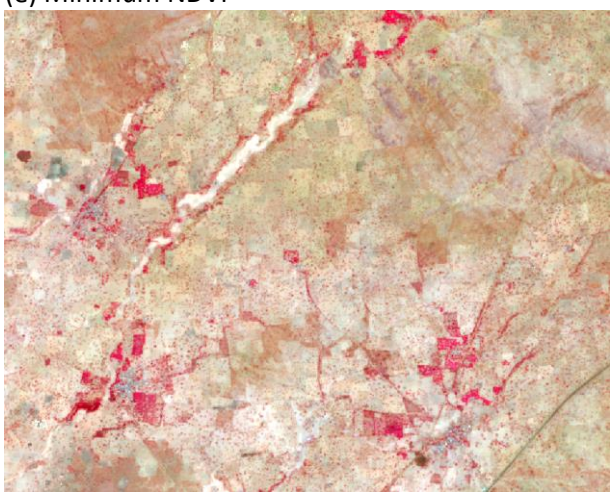
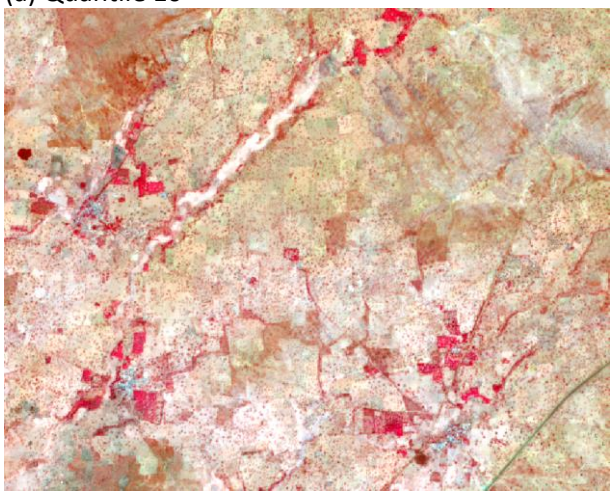


Figure 3-10. False colour (b8, b3, b2) knowledge-based features over the Mali site: (a) Maximum Red, (b) Maximum positive NDVI slope, (c) Maximum NDVI, (d) Maximum negative NDVI slope and € Minimum NDVI.

QC - Mali

(a) Quantile 10



(b) Quantile 90

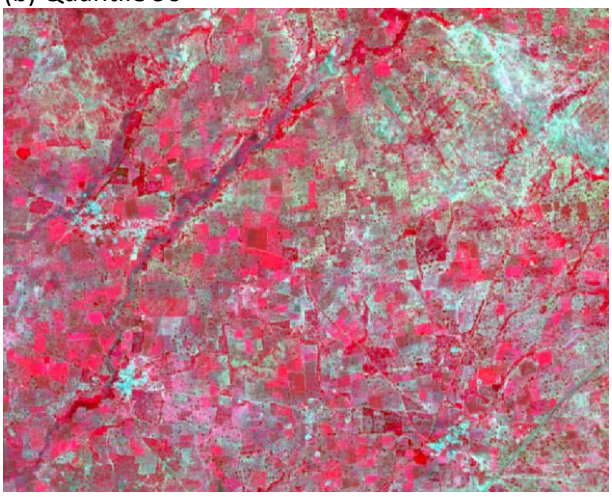
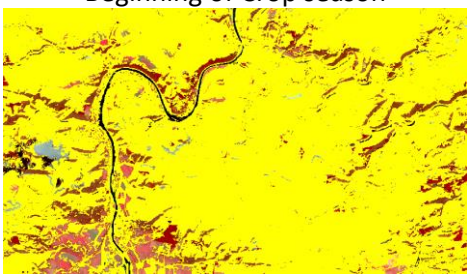


Figure 3-11. False colour (b8, b3, b2) quantile compositing features over the Mali site: (a) Quantile 10 and (b) Quantile 90.

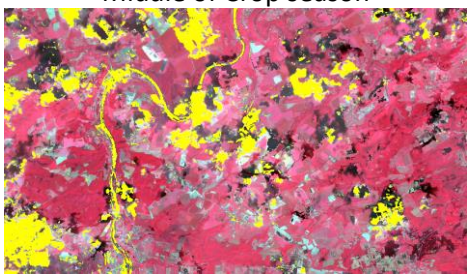
Belgium

(a) MVC NDVI

Beginning of Crop season

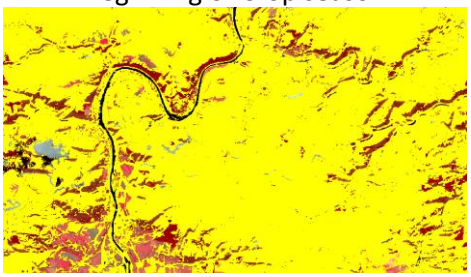


Middle of Crop season

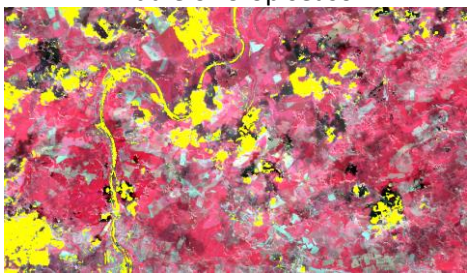


(b) MC

Beginning of Crop season

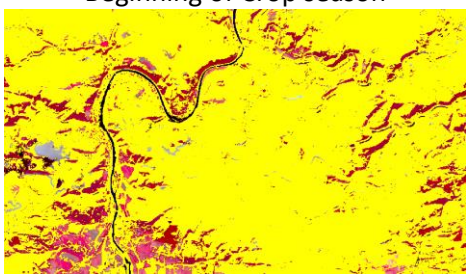


Middle of Crop season

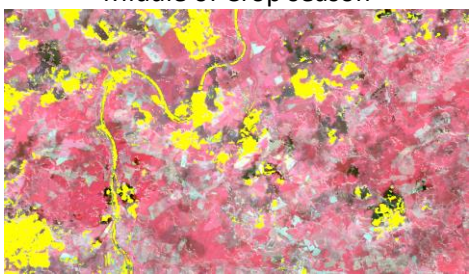


(c) WAC

Beginning of Crop season

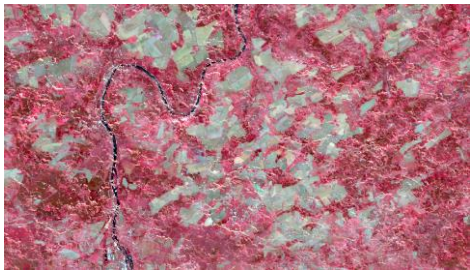


Middle of Crop season

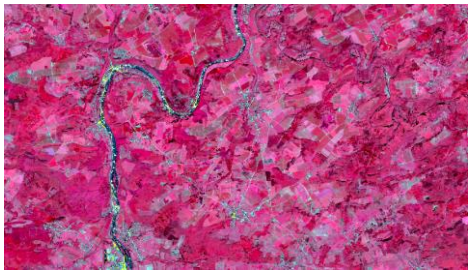


(d) KC

Maximum Red

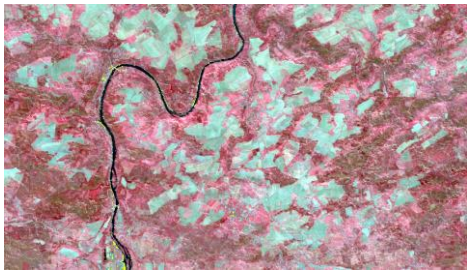


Maximum NDVI



(e) QC

Quantile 10



Quantile 90

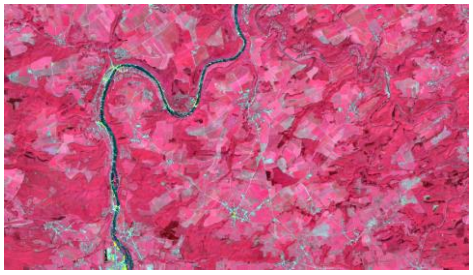


Figure 3-12. False colour (b8, b3, b2) of monthly ((a) MVC, (b) MC and (c) WAC) and features ((d) KC and (e) QC) composites comparing beginning of crop season (left) and middle of crop season (right). Yellow pixels are invalid pixels (cloud mask).

3.1.2.3.2 Quantitative analysis

Temporal consistency

This analysis focuses on the effects of compositing on reflectance values over various invariant land cover types (i.e. not vegetation) over time. This analysis is of major interest for land cover classification as it indicates temporal consistency of the compositing methods. This analysis concerns the time interval algorithms, i.e. MVC NDVI, MC and WAC, as their outputs are time series of monthly composites.

The ROI mean values for three spectral bands (b1: blue, b3: red and b8: NIR) are presented along the time for the three land cover types in Figure 3-13 a (roof top), b (bare soil) and c (water), coming from the Belgium and South Africa sites. In order to better visualize the temporal stability of reflectance values over time, standard deviation of the ROI mean values are computed over the entire time series. They are displayed for NIR, red and blue spectral bands in Figure 3-14 for roof (a and b), bare soil (c and d) and water (e and f).

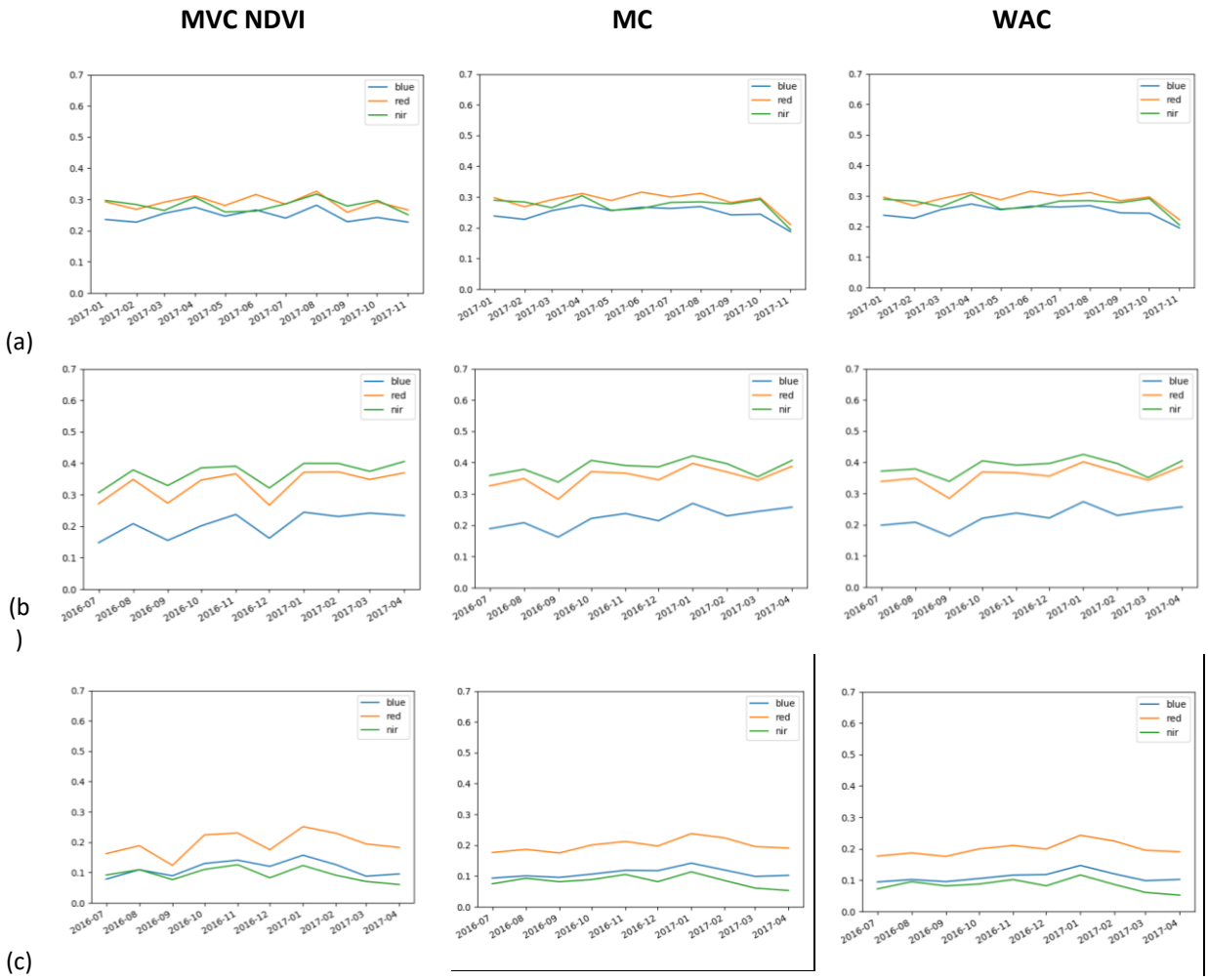


Figure 3-13. Temporal profiles of average surface reflectance for (a) roof top in Belgium, and (b) bare soil and (c) water in South Africa for MVC NDVI, MC and WAC composite time series.

For the three land cover types, MVC NDVI standard deviations are systematically higher than those of MC and WAC (green bars in Figure 3-14). It is also visible in the temporal profiles in Figure 3-13. It indicates that MVC NDVI composite time series are noisier, as concluded by the visual analysis of spatial consistency.

This difference is less present for roof tops, which is the more invariant surface compared to bare soil, which can contain small vegetation variations or water, which can vary according to e.g. sediments. Being a “best pixel method”, MVC NDVI could be sensitive to these small variations if they present extreme values.

In a general manner, standard deviations are not higher than 0.06 for most of spectral bands and land cover types, which indicates an acceptable temporal consistency. MC and WAC composite time series show very similar temporal profiles and standard deviations of reflectance values over the entire time series. They present less variations than the MVC NDVI.

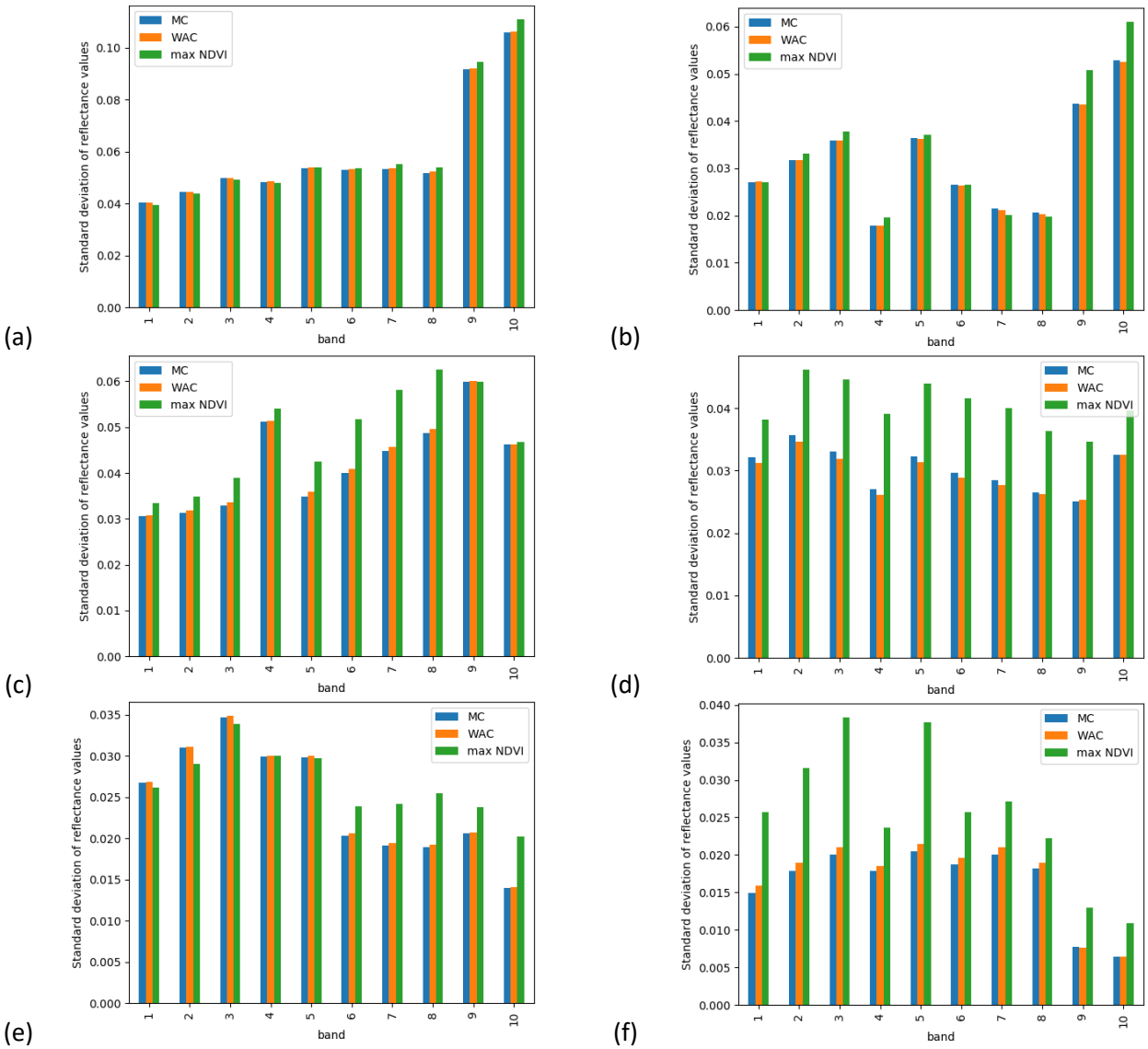


Figure 3-14. Standard deviation of average surface reflectance over roof top in (a) Belgium and (b) Mali, bare soil in (c) Belgium and (d) South Africa, and water in (e) Belgium and (f) South Africa, derived from the three time interval algorithms.

Fidelity to medium date image

This analysis confirms the visual examination in the previous section, with very large amount of artefacts for the MVC NDVI. The MVC NDVI has the worse fidelity to central date, especially in the NIR band. In the spring season, the vegetation is growing and the MVC NDVI tends to select the latest date with the greatest NDVI for vegetated pixels, which are therefore different from the images at the center of the compositing period. Regarding the MC and WAC, the observed performances are similar, with small advantage for the WAC.

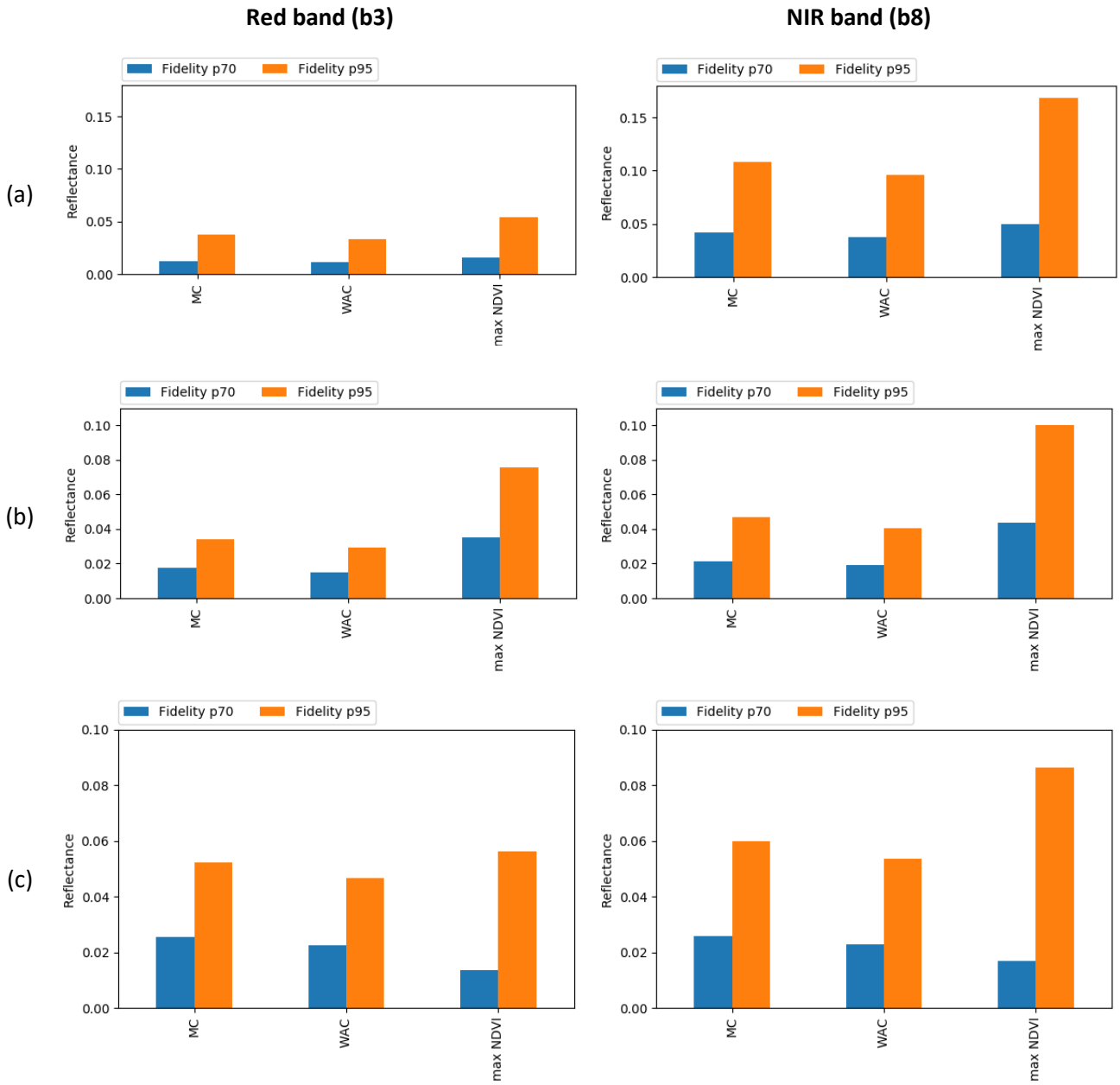


Figure 3-15. Fidelity to central date in the Red and NIR bands for MVC NDVI, MC and WAC for (a) the Belgium site, (b) Mali site and (c) South Africa site.

Remaining proportion of data gaps

Figure 3-16 shows the average percentage of data gaps remaining in the composites for the Belgium site. Given that the same compositing period was used for the time interval algorithms, i.e. MC, WAC and MVC NDVI, the three methods have exactly the same amount of remaining data gaps. Differences between the Maximum Red, Maximum NDVI, Minimum NDVI and the two Maximum slope NDVI features are due to the fact that the computation of a slope is not always possible if not enough data are available.

This analysis clearly shows the advantage of working with feature-based algorithms for cloudy sites like Belgium, as already observed in the visual analysis.

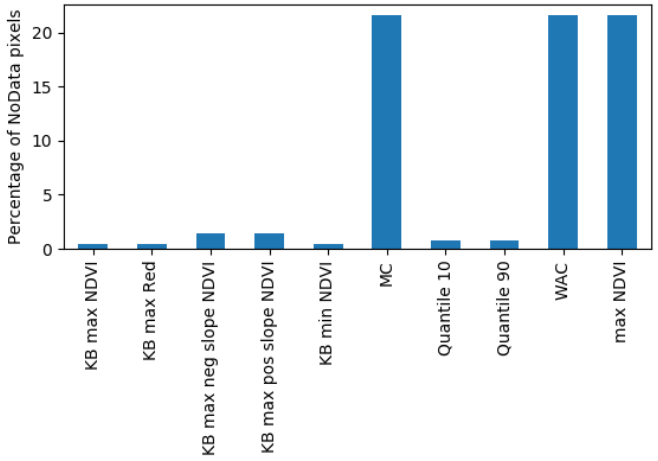


Figure 3-16. Average percentage of data gaps remaining in the composites for the Belgium site.

Artefacts

This analysis assesses the amplitude of the artefacts observable at the limits of zones obtained with the same set of dates. Figure 3-17 shows the standard deviation of the average difference of reflectance values between pixels at the external borders and pixels at the internal border of contiguous zones, for (a) Belgium and (b) Mali (Red and NIR bands).

In a general manner, more artefacts are presents in the Belgium site. This is probably due to the higher cloud cover, leading to more patches coming from different set of dates. Time interval algorithms show higher values of artefacts than features-based algorithms. This confirms the visual analysis showing more noise and artefacts in monthly composites.

Unexpected high artefacts in the WAC may come from the weights higher for the central date. Then, the reflectance values of the different set of dates results in more different values. Indeed, for all connected groups of pixels with the same set of dates, the average difference between the external border and the internal border of the contiguous zone will be higher is a lower weight is applied on the extreme images of the compositing period.

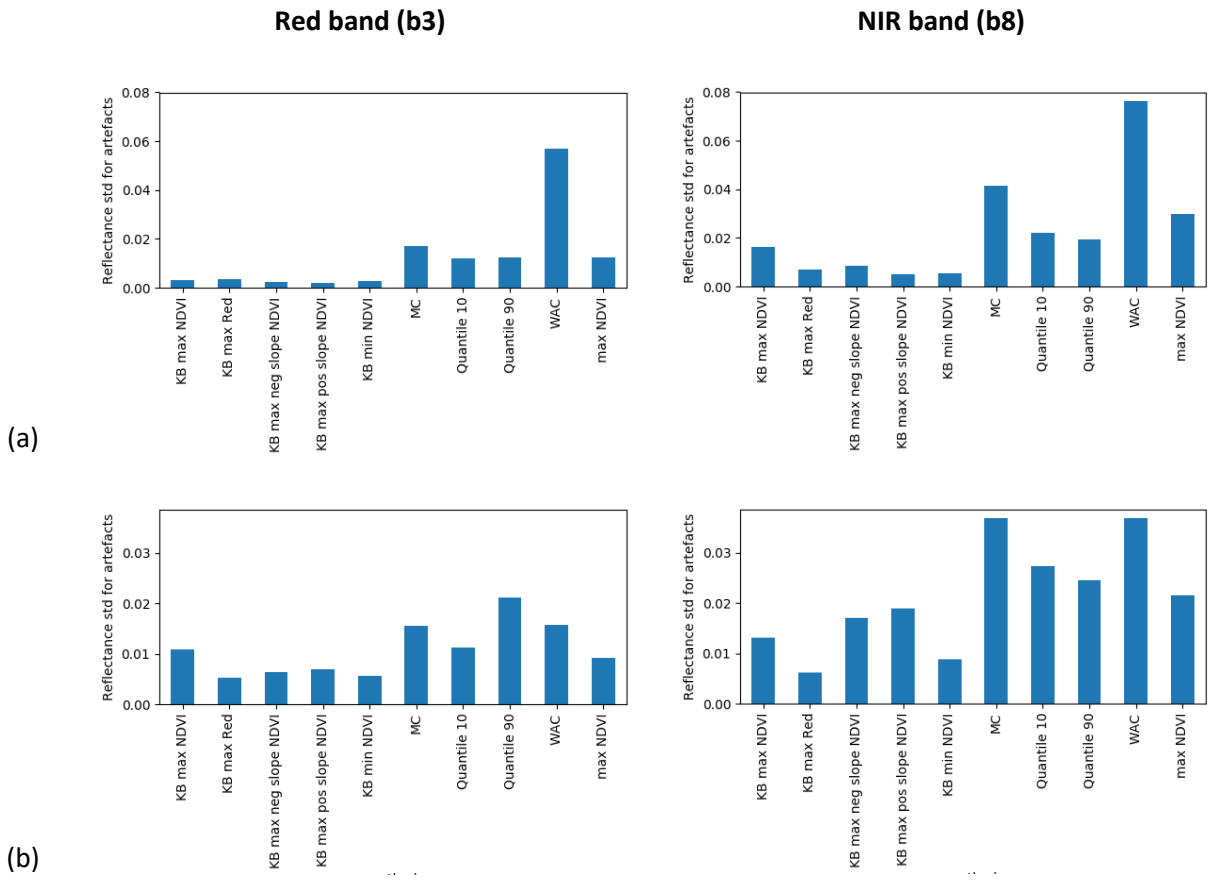


Figure 3-17. Artefacts in the Red and NIR bands for the five selected algorithms for (a) the Belgium site and (b) the Mali site.

3.1.2.4 Summary and conclusions

This analysis assesses the performance, advantages and drawbacks of five compositing approaches applied on land surface reflectance of Sentinel-2 images. The five methods considered are Maximum Value Compositing NDVI (MVC NDVI), Mean Compositing (MC), Weighted Average Compositing (WAC), Knowledge-based Compositing (KC) and Quantile Compositing (QC). One visual and four quantitative analyses examine the consistency as well as the noise introduced into composite images of the reflectance data time series.

Visual comparisons and quantitative analysis of the composites consistency provides complementary and coherent conclusions. The main advantages of feature-based algorithms (KC and QC) are a better spatial consistency achieved, thanks to the use of the entire time series as input, as well as very few data gaps compared to time interval algorithms. The time interval algorithms present the advantage of providing more composites for the same length of time series. Indeed, more outputs are available with monthly composites than only several features for the entire year. However, due to the short compositing period, some monthly composites could be partly or totally unusable because of the cloud cover. In addition, also due to their smaller compositing period, products of time interval algorithms are much more sensible to cloud mask artefacts.

More specifically, the features of KC based on slopes are more sensitive to noise and produce patchy results, especially for cloudy sites. The other features are very homogeneous with a high spatial consistency. Compared to KC products, features of QC produce cleaner images. However, other quantiles could be computed to get more inputs for classification algorithms.

The MVC NDVI outputs presents lower temporal and spatial consistencies than MC and WAC, which produce more homogeneous and very similar composites. The larger noise is due to the fact that this method only selects one date for each pixel and discards the others, compared to MC and WAC that are designed to reduce this effect by averaging all valid observations. However, MC and WAC are more sensitive to cloud masks artefacts because of the average of all valid pixels including those that are not supposed to be valid (undetected haze or cloud borders). These artefacts lead to patches and spatial inconsistencies visible in the products. In addition, undetected cloud shadows are strongly visible in MVC NDVI outputs, compared to MC and WAC.

The Sentinel-2 cloud mask presents too many artefacts concerning the delineation of clouds borders, the haze and cirrus detection and removal, the detection of cloud shadows and cloud commission for bright surfaces. Improvements are necessary to produce composites with sufficient quality for classifications, and for a benchmark interpretation based on compositing methods rather than on mask artefact.

3.1.3 Indices

A thorough list of envisioned indices has been reported in the document D31.1a. For this first phase, the focus will be set on the following indices, among the most used, the NDVI and the NDWI (also named NDMI). The BI will eventually be integrated to complement those exploratory studies.

During the MULTIPLY workshop that took place on the 5th-8th February 2018, it has been stated that the following phenological variables will be retrieved using different physical radiative transfer models (RTM) and made available on the platform, after the processing of Sentinel-2 and Sentinel-1 images, on demand:

- LAI, in optical and in microwave domain;
- faPAR;
- soil moisture and soil roughness
- canopy chlorophyll content
- canopy optical depth or thickness
- canopy height
- canopy water content
- leaf color

Those phenological indices and their contribution to the project (for example in the characterization of the type of crops and the species of trees) may be explored in more details once the platform become operational, if it is possible in the second phase.

3.1.4 Time Features

In the ECoLaSS Deliverable D6.1 – D31.1: Methods Compendium: Sentinel-1/2/3 Integration Strategies (AD06) several spectral, textural and also temporal indices are described which are of potential relevance as input for image or time series classification. The following sections describe the time features methodology (Valero et al., 2016) which was applied for the testing and benchmarking of methods for forest (section 4.1.2) and agriculture (section 4.1.4). The preliminary set of implemented features will be explained (section 3.1.4.1), followed by feature selection and a consecutive classification workflow implementing the time features (section 3.1.4.2).

3.1.4.1 (Preliminary) Set of Implemented Features

From the data described in 3.1.3 and the ECoLaSS WP 31 Deliverables (AD06), a set of different temporal-spectral features (time features) for varying time periods was calculated. Time features are able to capture

statistical properties and information about significant changes (due to seasonal patterns, extreme events or human activity) contained in the time series (Valero et al., 2016). They can be flexibly computed from reflectance or index data and can act as powerful input features for various classification or regression tasks. When dealing with different periods for vegetation phases in different geographic areas, the use of remote sensing time series data can be limited (Valero et al. 2016). This effect is mitigated by the time features, as their information is not directly related to the acquisition dates, they do not require prior knowledge of the change event dates or in general manual selection of scenes.

In case of Sentinel-2, the time features were calculated for the indices Brightness Index (BRIGHTNESS), Inverted Red Edge Chlorophyll Index (IRECI), Normalized Difference Vegetation Index (NDVI) and Normalized Difference Water Index (NDWI) (Table 3-3). In case of Sentinel-1, the time features were calculated for the VV and VH polarizations, the ratio of VV and VH and the normalized difference of VV and VH. The specific calculation and characteristics of the time features are described in more detail below.

Table 3-3. Time features calculated for various bands and indices.

Sensor	Bands / Indices	Time features
Sentinel-2	• Brightness (derived through summation of the values of the bands Green, Red, NIR and SWIR1) • IRECI (Inverted Red Edge Chlorophyll Index) • NDVI (Normalized Difference Vegetation Index) • NDWI (Normalized Difference Water Index, based on SWIR and NIR)	min, max, mean, std, p10, p25, p50, p75, p90, pdiff75/25, pdiff90/10, maxmean, activity, difmin3, difmax3, difdif3mean Postrend(NDVI only), negtrend (NDVI only)
Sentinel-1	• VV (Gamma0) • VH (Gamma0) • Norm. Difference VV/VH • Ratio VV/VH	min, max, mean, std, p10, p25, p50, p75, p90, pdiff75/25, pdiff90/10, maxmean, activity, difmin3, difmax3, difdif3mean

The features are considered as separated in two classes of different complexity, referred to as "simple" and "complex" time features. Simple time features are commonly used statistical metrics which are calculated over time using all valid (particularly cloud and cloud shadow free) observations. This includes the minimum (min), maximum (max), mean, standard deviation (std), different percentiles: 10^{th} (p10), 25^{th} (p25), 50^{th} (p50), 75^{th} (p75) and 90^{th} (p90), and the differences between the 90^{th} and 10^{th} (pdiff90/10) and 75^{th} and 25^{th} percentiles (pdiff75/27) of the time series.

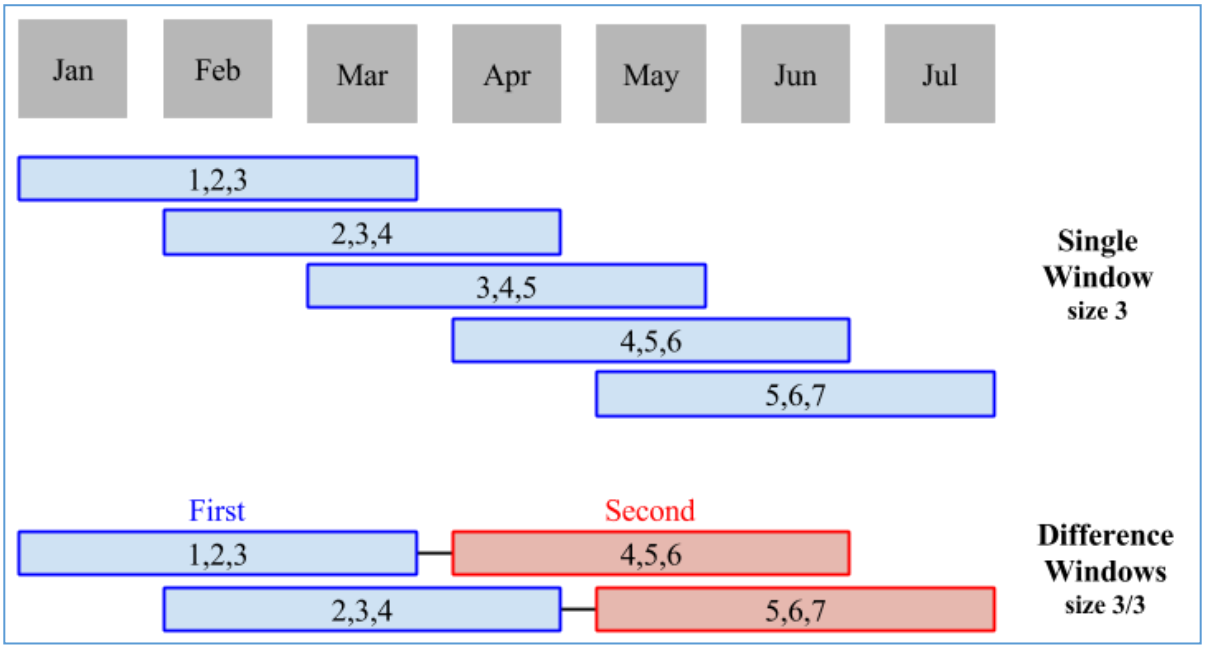


Figure 3-18. Temporal window concept: Single sliding temporal window (e.g. for calculation of mean_max) (top) and difference sliding temporal window configuration (e.g. for calculation of dif_max (bottom)); both examples have a window size of 3 consecutive observations.

The "complex" time features are calculated by the application of a temporal sliding window from the time series stack (Figure 3-18). At each window step, the information of the respective scenes inside the window range is integrated and used to iteratively update the desired time feature. E.g. the mean_max is the "stabilized" maximum value of the time series, iteratively updating the maximum feature by the mean of the scenes at each window step. The dif_max, dif_min, and dif_dif features use two offset temporal sliding windows (“difference windows”) to iteratively update the feature by the respective difference of the window scene complexes. These features represent the maximum positive (dif_max) and negative difference (dif_min) within the time series. The dif_dif feature is the difference of dif_max and dif_mean. The calculation of the dif_max is detailed in Figure 3-19. At each window step, for each pixel, the feature is only updated when at least one scene in both scene complexes is valid and cloud free for a specific pixel. Pixels, for which no update from the initial feature value of 0 was triggered keep this state. If at each iteration step no update was possible due to at least one of both scene complexes being completely cloud masked, the pixel is flagged as cloud masked in the final time feature. Instead of using a temporal window, the pos_tr and neg_tr loop through the time series and iteratively integrate information from the previous and recent scene to find pixels with significant positive or negative value transitions ((e.g in the case of a change from vegetation to bare soil) between consecutive scenes.

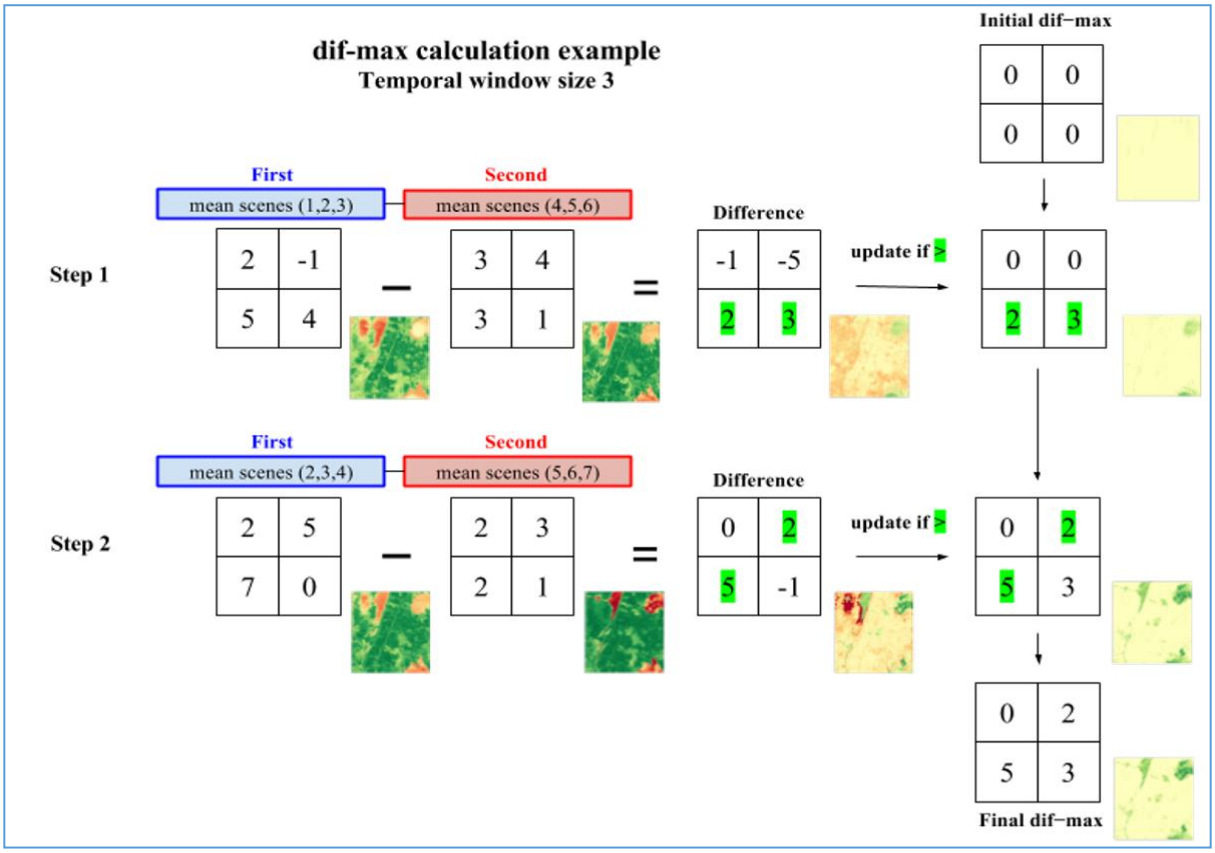


Figure 3-19. Concept of the calculation of a complex time feature shown for the dif_max time feature.

3.1.4.2 Feature Selection

One of the main challenges when deriving LC/LU maps over large areas is the generation of suitable spatially coherent layers or time series features for the analysis (usually supervised classification). The irregular nature of ordinary remote sensing time series data (e.g. due to clouds within a scene, different acquisition times between orbits) can be resolved via a best-available pixel composite approach (from the time series of each pixel, only the least cloudy one is combined in a composite image) – as mentioned in section 3.1.2 – or by calculating spectral-temporal time series metrics (e.g. mean, standard deviation, percentiles, etc.), see the previous section.

Building a large set of features is computationally expensive and it is desirable to reduce this cost by only building the features that turn out to be useful for the respective classification task. However, the optimal set of useful features is usually not known in advance. In order to tackle this problem the classification workflow of this work (Figure 3-20. Classification workflow.) explicitly addresses feature selection before the feature calculation for the full dataset is carried out.

The workflow comprises the following steps:

1. Extraction of raster values at reference data locations (where class labels are known) for all the available acquisitions, bands and indices. This results in a small data subset to work on before building the final features for the whole image footprint.
2. Calculation of the potential time features from the extracted data. Together with the known labels at the extracted sites, this yields the combination of labels and predictors/features required for training a classification model.
3. Splitting the full reference data in a training and test set.

4. Training the classifier based on the training set. Here, the first training step is a recursive feature elimination. This algorithm finds a small subset of all the potential input features with which a comparable (and sometimes higher) accuracy can be achieved compared to a full-feature model. After the suitable subset of features is known, the final model is trained with the selected feature.
5. Generation of an accuracy report based on the independent test data.
6. Calculation of the selected features for the whole raster data.
7. Prediction / mapping with the calculated raster features via the final model. This step yields the predictions (classes), class-probabilities (one layer per class), and three reliability layers (max. probability, breaking ties, entropy).

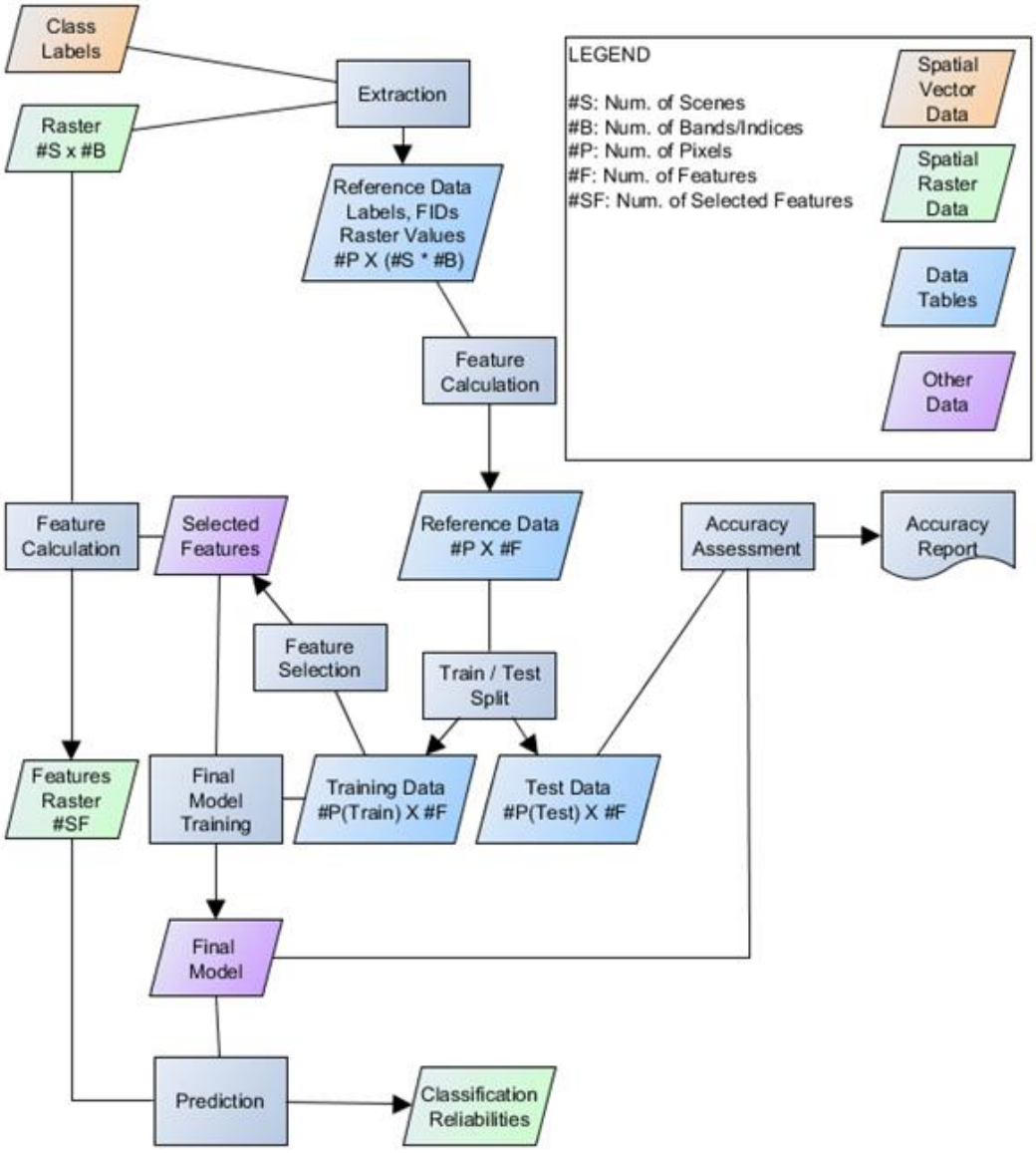


Figure 3-20. Classification workflow.

3.2 Time series classification methods

This subchapter addresses the testing and benchmarking the time series classification methods. This benchmark is addressed separately for different thematic fields: Imperviousness (section 3.2.1), Forest (section 3.2.2), Grassland (section 3.2.3) , Agriculture (section 3.2.4), and new land cover products (3.2.5). For each of the thematic classifications, different inputs, classification methods and parameters are assessed.

3.2.1 Imperviousness

The following subchapters comprise the testing and benchmarking of the time series classification methods for HRL Imperviousness.

3.2.1.1 Description of candidate methods

The objective of this Work Package is to develop a framework for times series analysis for thematic classification based on Sentinel multi-sensors constellation. In this section, the Imperviousness High Resolution Layer is addressed.

The material and input data

Following the WP32 and the time series preparation, the pre-processed Sentinel data (both Sentinel-1 and Sentinel-2) and time series are used for the tests. Pre-processing has been performed by the JR, as detailed in WP32.

Table 3-4 - Band order for ECoLaSS S2 data, as provided by JR (cf. report D7.1a on WP32) – the 1st, 11th and 13th bands, at a 60m-resolution, are only useful for the TOA to BOA processing, and are not therefore not included in the BOA final product.

ECoLaSS Stack number	Original band number	Spectral band
1	2	blue
2	3	green
3	4	red
4	8	NIR
5	5	NIR red
6	6	NIR red
7	7	NIR red
8	8a	NIR
9	11	SWIR
10	12	SWIR

Following the results of the WP31 (separability of the information for thematic classifications), the input data selected are constituted by:

- All the pre-processed Sentinel-1/2 images with all the spectral bands listed in Table 3-4;
- A subset of the full dataset based on the cloud cover and the useful images;
- A spectral subset of the full or partial dataset based on specific bands – ECoLaSS bands number 2, 3, 4, 7 and 9 – that avoid most band overlaps, thus making the most significant spectral extract;
- And a combination of spectral indices – here, the NDV and the NDBI.

Therefore, the current outcomes of the tests conducted for the WP31 solely rely on multispectral information. This kind of information is in fact essential to discriminate landscape elements but is not sufficient. A more effective detection could require advanced feature computation, that would be able to

discriminate objects. A large set of computable variables can be regrouped according to their properties as follow:

- **Texture and Structure:** Texture and structure analysis consists in extracting information on the spatial arrangement of pixels. Amongst numerous existing techniques, the following are particularly interesting, regarding the discrimination of impervious surfaces:
    - **Grey Level Co-occurrence Matrix (GLCM):** it is a widely used texture analysis technique in remote sensing. It consists in the distribution analysis of co-occurring pixel values at a given offset. Numerous indexes are derived from this matrix to extract texture properties (Haralick, Shanmugam, & Dinstein, 1973).
    - **Signal decomposition:** Signal decomposition techniques are used to provide a multi-resolution representation of the original image in a series of components related to a specific direction. Wavelet and Gabor analysis applied to VHSR images showed their efficiency for detecting textured objects (Lefebvre, Corpetti, & Hubert-Moy, Estimation of the orientation of textured patterns via wavelet analysis, 2011a), (Lefebvre, Corpetti, & Hubert-Moy, Wavelet and evidence theory for object-oriented classification: Application to change detection in Rennes metropolitan area, 2011b).
    - **Structural Features Set (SFS):** This method is based on a direction lines analysis. It implies computing the spectral difference between a pixel and its neighbours for a given direction in order to detect whether this pixel lies in a homogeneous area. This technique has been successfully applied in urban areas (Huang, Zhang, & Li, 2007).

The main drawback of these approaches lays in their intense time consumption and their requirement for a high level of parameterization that render them intractable for large-area analysis. That is why the tests will rather be conducted on different methods:

- **Granulometry by Mathematical morphology:** Mathematical morphology is the analysis of the image constructions and their distribution at different scale. It consists in simplifying the image progressively though the preservation of bright elements (with closing operators) or dark elements (opening operators)**.** Amongst numerous existing techniques, the following one is particularly interesting et is testing in the frame of the times series classification methods:

Differential Attribute Profiles (DAP): Multiscale features often appear as a relevant alternative, with Gabor filters and Differential Morphological Profile (DMP) having achieved great classification performances. However, even such features come with a significant cost. DMP is relying on a series of morphological filters by reconstruction and it has shown for more than a decade its ability to deal with VHSR images (Pesaresi & Benediktsson, A new approach for the morphological segmentation of high-resolution satellite imagery, 2001). Recently, an alternative multiscale feature, called Differential Attribute Profile (DAP) (Dalla Mura, Benediktsson, Waske, & Bruzzone, 2010) has been built upon DMP to achieve more discriminative power, a higher flexibility, for a lower computational cost. DAP is very appealing since it is computed from a tree-based image representation that can be built with very efficient algorithms (Carlinet & Géraud, 2014). Once the tree is built, the description of each pixel (or object, node) is straightforward and relies on the analysis of all its ancestors up to the root. As such, it has been embedded in large-scale analysis performed by Joint research Center (JRC) such as the Global Human Settlement Layer (GHSL) (Pesaresi, et al., 2013) and European Settlement Map (Florczyk, et al., 2015).

The training data chosen must therefore be representative of the whole study area in order to cover all the reflectance variations of the classes, as well as to go further and take into account the local variability of the environmental classes due to the soil type, moisture, etc. The training sites must be exempt from anomalies and must be a suitable statistical representation of the area. There must be a substantial number of them. That is why, the High Resolution Layers have been used as training data:

- HRL Imperviousness 2015;

- HRL Forest 2015;
- HRL Grassland 2015;
- HRL Water and Wetness 2015;
- HRL Small Woody Features 2015.

The sampling design refers to the protocol whereby the training samples are selected. A probability sampling design is preferred for its objectivity. “Simple random, stratified random, clustered random and systematics designs are all examples of probability sampling designs” (Stehman & Czaplewski, 1998). For the purpose of the tests, a stratified random approach, based on the HR Layers, has been preferred.

The candidate methods

The time series classification methods can be divided into two categories:

- The mono-temporal pixel-based classification, which is performed for each image of the time series;
- The multi-temporal pixel-based classification, which is performed on a full stack of each individual images to reconstruct a one-year time composite time series to take advantage of the phenology of inter-yearly and intra-yearly seasonal dynamics. The algorithms are based on statistical metrics derived from this yearly time series (median, min, max, standard deviation).

Multiple algorithms could be used to map artificial lands. Classification methods range from unsupervised algorithms such as K-means to parametric supervised algorithms such as maximum (Jensen, 2005); to machine learning algorithms such as artificial neural networks (Mas & Flores, 2008), decision trees (Breiman, Classification and Rgression Trees, 1984), Support Vector Machines (Mountrakis, Im, & Ogola, 2011) and ensembles of classifiers such as Random Forest (Breiman, Random Forests, 2001). A selection of these best algorithms for classification has been made, specially adapted for the imperviousness topic:

- K-means;
- Support Vector Machine (SVM);
- Random Forest (RF);
- Artificial Neural Networks (ANNs);
- Active learning (AL).

The methods selected are pixel-based classifications based on two fundamental principles: all the objects (or pixels) of the same class are characterized by identical spectral signatures and all the signatures of the object classes are perfectly distinct from each other. Commonly, there are two classification methods based on the pixel from which all the variants are derived. These are supervised (SVM, RF and NNs) and unsupervised classifications (K-means).

Specifically, Random Forest (RF), Support Vector Machines (SVM) are supervised tree based classification approaches. In our study case, these methods will be applied to create the updated built-up mask 2017. Their objectives are to find and recognize patterns in data in order to analyze and classify it as seen in studies like (Gilsason, Benediktsson, & Sveinsson, 2006), (Rodriguez-Galiano, Ghimire, Rogan, Chica-Olmo, & Rigol-Sanchez, 2012), (Tan, Steinbach, & Kumar, 2006), (Lary, Alavi, Gandomi, & Walker, 2015) and (Camp-Valls & Bruzzone, 2009) to take a few examples.

K-means

The K-mean clustering algorithm is one the most popular classifier in remote sensing. It assumes that features associated with each class are distributed according to a Gaussian distribution. Results are then easy to understand but it can lead to spurious results if the data is not normally distributed. This method is a pixel-based unsupervised and iterative classification algorithm based on spectral information and similarity. In fact, in order to reduce the variability within each cluster (based on sums of square distances (errors) between each pixel), the algorithm performs two steps iteratively:

- Reassign data points to the cluster whose centroid is closest;
- Calculate new centroid for each cluster.

K-means classification automatically identify groups (or classes) on the basis of the spectral information of the pixels. These classes are then associated with types of land use in order to produce the map. This classification is made without any information a priori on the nature of the objects to be classified. The k-means assumes that the number of clusters is known a priori.

Therefore, multispectral data is most commonly used for this type of classification as it enables the differences of the signatures between the objects to be best exploited.

Even if this algorithm is used in studies for the detection of built-up (Jensen, 2005) (Lu & Trinder, 2006), the K-means classification tends to be not completely suitable as unsupervised classification requires a post priori interpretation of the terrain or other reference data signified by the classes obtained. K-means method is therefore not worthwhile to be tested. Supervised classifications are much more adapted.

Indeed, the following three methods require a set of training data to be defined and established. Basically, this set of training data enables a library to be established based on the spectral signature types for each class which needs to be extracted. The spectral signature of each pixel of the image is analysed and compared to the signature types established initially for each class. Assigning a pixel to a given class is based on criteria which complies with the decision rules and algorithms (whether parametric or non-parametric), ultimately resulting in the image to be split into groups.

Studies tend to show that these methods are more accurate and efficient compared to conventional algorithms such as K-means. These algorithms can deal with large multi-dimensional and complex data. Moreover, these methods have been used for large area mapping including human settlement and imperviousness areas (Hansen, Dubayah, & DeFries, Classification trees: an alternative to traditional land cover classifiers, 1996), (Pesaresi, Gerhardinger, & Kayitakire, A Robust Built-Up Area Presence Index by Anisotropic Rotation-Invariant Textural Measure, 2008), (Pesaresi, et al., 2013), (Kemper, Mudau, Mangara, & Pesaresi, 2015).

Support Vector Machine (SVM)

Support vector machines is a supervised non-parametric statistical learning technique; therefore, no assumption is made on the underlying data distribution, contrary to the previously mentioned method. This is an advanced classifier representing input data in a specific feature space within which each class is ‘easily’ separable. The prime advantage of the SVM classification is that it requires very few parameters. However, SVM is complicated to implement due to the large number of parameters that need to be adjusted and is difficult to automate (Mountrakis, Im, & Ogola, 2011). Additionally, this algorithm has a tendency to over-fit the data.

Random Forest (RF)

Random Forest combines many decision trees to obtain better predictive performance. Each decision tree is calibrated on a selection of random subset. Such algorithms such as RF have recently received increasing interest (Rodriguez-Galiano, Ghimire, Rogan, Chica-Olmo, & Rigol-Sanchez, 2012), (Breiman, Random Forests, 2001) because they are reputed more accurate and robust to noise than single classifiers (Shang & Breiman, 1996). The philosophy behind classifier ensembles is based upon the principles that a set of classifiers perform better than an individual classifier can. Breiman introduced RF in 2001 which presents many advantages for its application in remote sensing:

- efficiently on large data bases;
- thousands of input variables without variable deletion;
- estimation of what variables are important in the classification;

- relatively robust to outliers and noise;
- computationally lighter than other tree ensemble methods (e.g. Boosting);
- not sensitive to overtraining.

A RF consists of a combination of classifiers where each classifier contributes with a single vote to the assignation of the most frequent class detected for the input vector. The fact that it is a combination of many classifiers confers RF some special characteristics which make it substantially different to a traditional classification trees (CT). A RF increases the diversity of the trees by making them grow from different training data subsets created through.

Artificial Neural Networks (ANNs)

Neural networks consist of a set of adaptive functions (neurons) able to approximate a non-linear system. Neural networks algorithms are supervised classifiers particularly suitable when a large quantity of samples is available (Benediktsson, Swain, & Ersoy, 1990). Indeed, Artificial Neural Networks (ANNs) are computing systems inspired by the biological neural networks that constitute animal and human brains. Such systems progressively improve performance on tasks by considering examples, generally without task-specific programming, but carefully tailored to achieve one sole goal. These methods work without any a priori knowledge and evolve their own set of relevant characteristics from the learning material that they process – as explored in (Kemper, Mudau, Mangara, & Pesaresi, 2015) or (Lefebvre, Sannier, & Corpetti, Monitoring Urban Areas with Sentinel-2A Data: Application to the Update of the Copernicus High Resolution Layer Imperviousness Degree, 2016).

However, such as SVM, neural networks are complicated to implement due to the large number of hyperparameters that need to be adjusted and are thus difficult to automate. Those algorithms are also prone to over-fit the data.

Active Learning (AL) and Differential Attribute Profiles (DAP)

Production of Land Cover maps is usually achieved with selection of reference (or training) data, supervised classification, and manual map refinement/correction. The classification accuracy is directly related to the quality of the training samples, i.e. their ability to represent the data to be classified. Collecting training samples is done through a costly operation consisting of manually labelling the pixels. Furthermore, such pixels may not be representative of the land cover classes, thus requiring important corrections in the post processing step. To alleviate these issues, active learning has been introduced a couple of decades ago, and used in remote sensing since more than 5 years (Tuia, Ratle, Pacifici, Kanevski, & Emery, 2009)It works in both interactive and batch mode. In the former case, the user is given some specific pixels to label (e.g. by photo-interpretation), while in the latter case only relevant samples from the training sets will be used (leading to a better modelling of land cover classes as well as a more efficient classification process). It has been a very active field of research (Tuia, Volpi, Copa, Kanesvski, & J., 2011) reaching similar accuracies than supervised classifiers but with only 5 to 10% of the training samples. It is now considered as a well-established framework(Crawford, Tuia, & Ynag, 2013). Recent developments are related to large-scale analysis and domain adaptation (Alajlan, Bazi, AlHichri, Melgani, & Yager, 2013) or multiscale classification (Zhang, Zhu, Zhang, & Du, 2016).

Following the approach mono-temporal for which a classification is performed for each image of the time series, it is required to fuse them to provide a unique map, synthetizing all information.

Nevertheless, because of the different parameters of associated images (spectral and spatial resolution, acquisition date, cloud cover, etc.) and algorithms, their classification may provide results associated with various levels of quality. Although selecting the best result among all available classifications would seem a rational approach, combining them by taking into account their qualities should make it possible to reach

an even higher level of accuracy. This is the idea behind the concept of data fusion. A large number of techniques is available to fuse data. Two main groups of techniques can be distinguished, based on:

- The probability theory, such as Kalman filter and other data assimilation techniques depending on the presence of models for sensors;
- The evidence theory, where each decision is represented with a belief function associated with uncertainties. In this family, we find Dempster–Shafer Theory (DST).

In a remote sensing context, we rely on evidence theory and in particular on the Dempster–Shafer Theory of evidence (DST). The DST is based on a Bayesian approach and fuses a set of mass functions issued from various sources of observations associated with a weighted belief on some hypotheses. A key advantage is that uncertainty (the union of all hypotheses for a given pixel) is accurately managed by the Dempster's fusion rule.

Regarding the principles behind the algorithm, for each pixel, the class label for which the belief function is maximal is selected. This belief function is calculated by the Dempster Shafer combination of degrees of belief, also referred to as masses, and indicates the belief that each input classification map represents for each label value. Moreover, the masses of belief are based on the input confusion matrices of each classification map, either by using the "precision" rates, "recall" rates, "overall accuracy", or the "kappa" coefficient. Thus, each input classification map needs to be associated with its corresponding input confusion matrix file for the Dempster Shafer fusion.

DLR Settlement Extent and Growth Classifier

The 2017 settlement extent product is generated by properly exploiting multi-temporal Sentinel-1 (S1) radar and Sentinel-2 (S2) optical data intersecting the investigated Area of Interest (AOI) in 2017. Specifically, the rationale of the adopted methodology is that given a series of satellite images for the AOI, the temporal dynamics of human settlements are sensibly different than those of all other land-cover classes (e.g., vegetated and cultivated areas are prone to multiple changes over 1/2-year timeframe, whereas this generally does not occur for built-up structures).

As regards radar data, Ground Range Detected S1 scenes acquired at high resolution in Interferometric Wide Swath Mode (IW GRDH) are used. Each scene is pre-processed by means of the SNAP software available from ESA; specifically, this task includes: orbit correction, thermal noise removal, radiometric calibration, Range-Doppler terrain correction and conversion to dB values. Scenes acquired with ascending and descending pass are processed separately due to the strong influence of the viewing angle in the backscattering of built-up areas. Moreover, the joint employment of VV/VH imagery did not provide any sensible improvement with respect to the solely use of VV data; accordingly, VH data are excluded. As a means for characterizing the behavior over time, the backscattering temporal maximum, minimum, mean, standard deviation and mean slope is derived for each pixel. Texture information is also extracted to ease the identification of lower-density residential areas; in particular, the coefficient of variation (COV) of the temporal mean backscattering is computed, which is defined for each pixel as the ratio between the local standard deviation and the local mean calculated over a NxN spatial neighborhood. In the light of the 10m spatial resolution of the considered S1 data, a neighborhood of 5x5 pixels proved to be an effective choice.

Concerning optical data, only Sentinel-2 scenes with cloud cover lower than 60% are taken into consideration (indeed, further raising this threshold often results in accounting for images with non-negligible misregistration error). Data are calibrated and atmospherically corrected using the Sen2Cor software available from ESA. Next, a series of 6 spectral indices suitable for an effective delineation of settlements (identified through extensive experimental analysis) are extracted; these include – among others – the Normalized Difference Built-Up Index (NDBI), the Modified Normalized Difference Water Index (MNDWI) and the Normalized Difference Vegetation Index (NDVI). For all of them, the same set of 5 key temporal statistics used in the case of S1 data are generated for each pixel in the AOI. Moreover, to improve the detection of suburban areas, for each of the 6 temporal mean indices also the corresponding

COV is computed in a neighborhood of 3x3 pixels, which proved being the most effective choice in the light of the 10m spatial resolution of Sentinel-2 data.

To identify reliable training points for the settlement and non-settlement class, a strategy has been designed which jointly exploits the temporal statistics computed for both S1 and S2 data, along with additional ancillary information. In the case of optical data, in general the most of settlement pixels can be effectively outlined by properly jointly thresholding the corresponding NDBI, NDVI, and MNDWI temporal mean; likewise, this holds also for non-settlement pixels. Nevertheless, being all 3 spectral indices affected by the presence of vegetation, absolute threshold values are not universally effective since vegetation strongly varies depending on climate. To overcome this drawback, by accounting for the well-established updated Köppen Geiger climate classification, for each zone specific thresholds have been determined for outlining both candidate settlement and non-settlement training samples. Furthermore – in the reasonable hypothesis that the higher is the number of cloud/cloud-shadow free acquisitions, the more robust are the corresponding temporal statistics – all pixels whose number of Sentinel-2 clear-sky acquisitions is lower than 5 are excluded.

Regarding radar data, it generally occurs that the temporal mean backscattering of most settlement samples is sensibly higher than that of all other non-settlement classes. Accordingly, samples whose temporal mean backscattering (either in the case of data acquired in ascending and descending pass) computed from more than 4 scenes is: i) lower than -8.5 dB are not eligible to be labelled as settlement training samples; and ii) greater than -11 dB are not eligible to be labelled as non-settlement training samples. Finally, it is worth noting that in complex topography regions: i) radar data show high backscattering comparable to that of urban areas; and ii) bare rocks are present, which often exhibit a behaviour similar to that of settlements in the multispectral based temporal statistics. Accordingly, to exclude these from the analysis, all pixels are masked whose slope - computed based on SRTM 30m DEM for latitudes between -60° and +60° and the ASTER DEM elsewhere - is higher than 10 degrees.

Support Vector Machines (SVM) with Radial Basis Function (RBF) Gaussian Kernel are used in the classification process. However, as the criteria defined above for outlining training samples might results in a high number of candidate points, for AOIs up to the a size of ~10000 km² the most effective choice proved extracting 1000 samples for both the settlement and non-settlement class. However, since results might vary depending on the specific selected training points, as a means for further improving the final performances and obtain more robust classification maps, 20 different training sets are randomly generated and given as input to an ensemble of as many SVM classifiers. Then, a majority voting is applied and each pixel is finally associated with the settlement class only in the case it is labeled as settlement in at least 11 over 20 of them.

It worth noting that the stacks of S1- and S2-based temporal features are classified separately as this proved more effective than performing a single classification on their merger.

In both cases, a grid search with a 5-fold cross validation approach is employed to identify for each training set the optimal values for the learning. The values resulting in the highest cross-validation overall accuracy is selected and used for classifying the corresponding AOI. In particular, this is carried out by employing the largely employed open source C++ library libSVM.

A final post-classification phase is dedicated to properly combining the S1- and S2-based classification maps and automatically identifying and deleting potential false alarms. To this purpose, an updated version of the post-editing object-based approach adopted in the production of the GUF2012 has been used. Specifically, it consists of two phases. First, segmentation is performed for categorizing each cluster of connected pixels in the two classification maps as individual image objects; then, a ruleset is employed for selecting whether: i) to combine the S1- and S2-based objects; ii) to keep just one; or iii) to discard both of them. The final classification map is given by the merger of the objects preserved in the S1- and S2-based classification maps.

DLR Imperviousness processor

Urban growth is associated not only to the construction of new buildings, but – more in general – to a consistent increase of all the impervious surfaces (hence also including roads, parking lots, squares, pavements or railroads), which do not allow water to penetrate, forcing it to run off. To effectively map the extent of all such areas is then of high importance being it related to the risk of urban floods, the urban heat island phenomenon as well as the reduction of ecological productivity. Moreover, monitoring the change in the imperviousness over time is of great support for understanding, together with information about the temporal evolution of the extent of urban areas, also more details about the type of urbanization occurred (e.g., if areas with sparse buildings have been replaced by highly impervious densely built-up areas or vice-versa).

To this purpose an imperviousness product is generated by properly exploiting S2 multi-temporal imagery acquired over the study area within a given time interval of interest in which no relevant changes are expected to occur (typically a time period of 1-2 years allows to obtain very accurate results). For all the considered scenes, cloud masking and, optionally, atmospheric correction are performed. Next, the normalized different vegetation index (NDVI) is extracted for each image. Here, being it inversely correlated with the amount of impervious areas (i.e., the higher the NDVI, the higher the expected presence of vegetation, hence the lower the corresponding imperviousness) the core idea is to compute per each pixel its temporal maximum which depicts the status at the peak of the phenological cycle. It is worth noting that for different pixels in the study area, different number of scenes might be available. However, in the hypothesis of a sufficient minimum number of acquisitions for computing consistent statistics, this does not represent an issue. Moreover, in this framework it is also possible to obtain spatially consistent datasets to be employed for the desired analyses even when investigating large areas. Areas associated with impervious surfaces are then extracted at high spatial resolution [e.g., by photointerpretation of VHR imagery (e.g., from Google Earth), the analysis of OpenStreetMap layers or information derived from in-situ campaigns] in various parts of the study region and then rasterized and aggregated at the Sentinel-2 10m spatial resolution. A support vector regression SVR module is then employed for properly correlating the resulting training information with the temporal maximum NDVI to finally derive the PIS for the entire area of interest.

3.2.1.2 Benchmarking criteria

Benchmarking is conducted in two steps:

- Validation of the products based on visual check also known as “look-and-feel” to eliminate and exclude obvious methods/algorithms that present poor results and then,
- Assessment of layers using validation sites to perform a thematic accuracy measurement using the current metrics such as: user, producer accuracies or omission and commission errors.

The look-and-feel is a visual comparison between the resulting classification and a reference map: here the HRL IMD 2015 is selected, as seen with validation points on the Figure 3-21, since few changes is expected between the year 2015 (sometimes using data from 2016) and the year 2017.

The validation approach provides guidance on how the classification results will be validated by defining suitable indicators or metrics. Classification correctness should be evaluated using misclassification rate and/or misclassification matrix. Thematic accuracy cannot be subjected to an exhaustive check. A thorough thematic assessment would imply a very time-consuming work and therefore high costs. Misclassification rate is estimated by sampling and product information is compared to reference data. The aim is to provide a description of suggested procedures for a scientifically and statistically sound sampling scheme for assessing the thematic quality of the Imperviousness products obtained in the tests.

Thus, thematic accuracy assessment has three components: (i) the sampling design, (ii) the response design and (iii) the estimation and analysis procedures.

(i) The stratification and the sampling design primarily consist in selecting an appropriate sampling frame and sampling units. These sampling units can either be "defined on a cartographic representation of the surveyed territory" (Gallego, Area Frames for Land Cover Estimation: Improving the European LUCAS Survey, 2004), in which case it is an area frame, or on a list of the features. According to this study, area frames give a better representation of the population as the spatial dimension is kept.

In an area frame, sample units can be points, lines (often referred to as transects) or areas – often referred to as segments, described in (Gallego, Sampling Frames of Square Segments, 1995). The first step is to define the AOI for which the accuracy assessment is to be reported and the type of sample units. For the majority of cases, point samples will be used, but areas or segments may be used in specific cases such as when not only thematic accuracy needs to be reported, but also the geometry of mapped objects. Polygons have also the drawback of being specific to a single map. In case of changes, the sample may not be adapted anymore. Points are considered as the most appropriate unit for our tests.

Sampling design refers to the protocol whereby the samples are selected. A probability sampling design is preferred for its objectivity. "Simple random, stratified random, clustered random and systematics designs are all examples of probability sampling designs" (Stehman & Czaplewski, 1998). Even though a simple random design is easy to implement, its main drawback lies in the fact that some portions of the population may not be adequately sampled. Cluster sampling is often used to reduce the costs of the collection of reference data, but does not resolve geographic distribution problems. A systematic approach would solve this problem, yet it is not appropriate if the map contains cyclic patterns. A stratified approach consists in allocating a pre-defined number of samples per land-cover class. As explained in Stehman's paper, stratification ensures that each class is correctly represented.

The validation approach chosen combines random and stratified approaches and benefits from the advantages of both of them.
For the purpose of the tests, a stratification is applied based on a series of omission and commission strata:

- Commission: Imperviousness Degree 1-100% in the layer 2015 (historical layers)
- Omission: Imperviousness Degree 0% in the layer 2015

The HR Layers from previous productions of 2015 are used in order to perform the stratification, as seen in Figure 3-21.

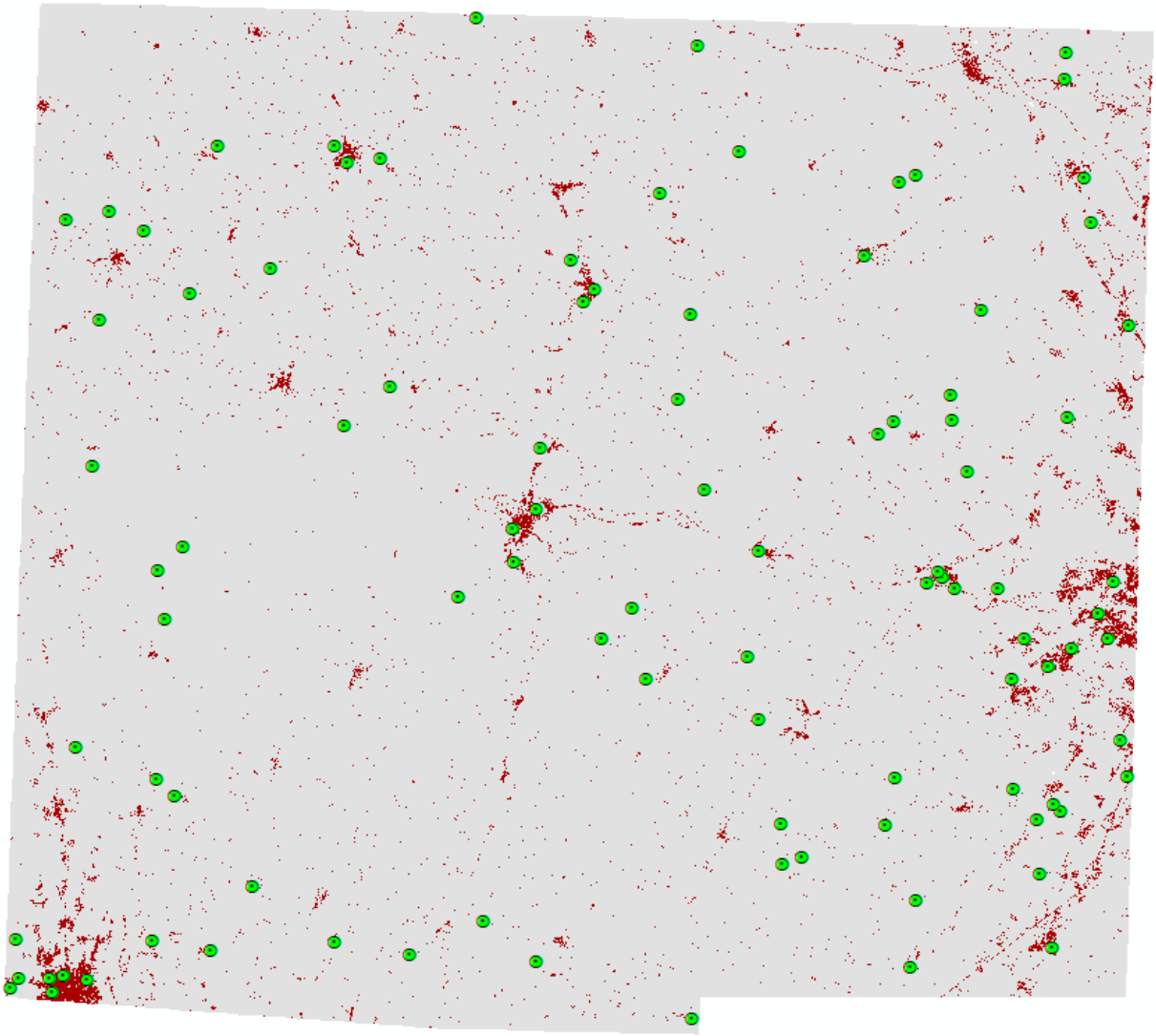


Figure 3-21. Validation samples overlaid on the HRL IMD 2015, reference map.

(ii) The response design "is the protocol for determining the reference land cover classification of a sampling unit" (Stehman & Czaplewski, 1998).

The datasets against which the interpretation is performed are divided in two main groups, guiding data and reference data. The guiding data used in the production of the classifications are the re-processed HR Sentinel-2 data. The reference data provide more spatial details and stronger landscape context to the assessment. The available reference data are:

- Bing maps image / cartography layer
- Google Earth image / cartography data

(iii) The density values are not directly assessed, only the binary built-up mask.

Thematic accuracy is usually assessed based on the construction of a confusion or error matrix made out of the results of the samples interpretation.

3.2.1.3 Implementation and results of benchmarking

As describe before, the benchmarking is only done on Sentinel-2 cloud-free images and they offer a high resolution (spectral and spatial). The implementation has been done on the test site in South-West site of France, over the tiles 30TYP and 31TCJ.

We saw that various classification methods, input data set, fusion algorithms can be explored regarding the thematic classification.

The following tests proposed for the determination of the algorithms used are related to:

- the various Dempster-Shafer fusion algorithms to merge the classifications, as listed in the Table 3-5;
- the classification algorithms themselves, as listed in Table 3-6;
- the various input data that can be feed to the classification algorithms, as listed in Table 3-7.
- the various input sensor (Sentinel-1 or 2) that can be used for the classification algorithms, as listed in Table 3-8.

Table 3-5. Tests related to the Dempster-Shafer fusion algorithm choice.

Test	Input Data	Fusion of classifications	Metrics used for the Dempster-Shafer fusion
1	Full dataset – all bands	Support Vector Machine	Overall Accuracy
2	Full dataset – all bands	Support Vector Machine	Kappa coefficient
3	Full dataset – all bands	Support Vector Machine	Precision rate
4	Full dataset – all bands	Support Vector Machine	Recall rate

Table 3-6. Tests related to the classification algorithm selection.

Test	Input Data	Fusion of classifications	Classification algorithm
1	Full dataset – all bands / Subset dataset (36 images) – all bands	Dempster-Shafer	Random Forest
2	Full dataset – all bands / Subset dataset (36 images) – all bands	Dempster-Shafer	Support Vector Machine
3	Full dataset – all bands / Subset dataset (36 images) – all bands	Dempster-Shafer	Artificial Neural Network

Table 3-7. Selection of the best input dataset based on the results given by various classifications.

Test	Input Data	Fusion of classifications	Classification algorithm
1	Full dataset – all bands	Dempster-Shafer	Support Vector Machine/Random Forest/ Artificial Neural Network
2	Selection of the 36 images – all bands	Dempster-Shafer	Support Vector Machine/Random Forest/ Artificial Neural Network
3	Selection of the 36 images – bands subset (Bands 2, 3, 4, 7 and 9)	Dempster-Shafer	Support Vector Machine/Random Forest/ Artificial Neural Network
4	Selection of the 36 images – indices NDVI and NDBI	Dempster-Shafer	Support Vector Machine/Random Forest/ Artificial Neural Network Vector Machine

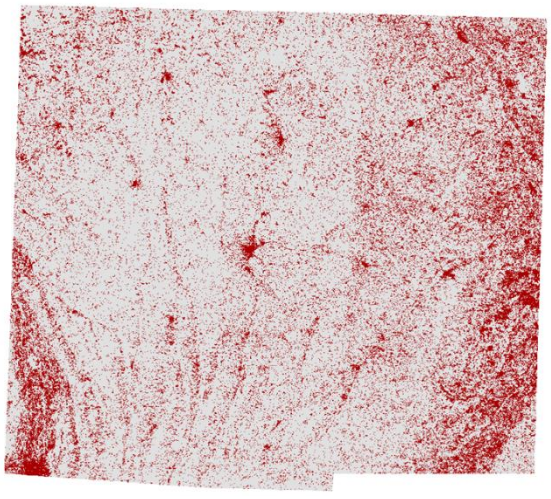
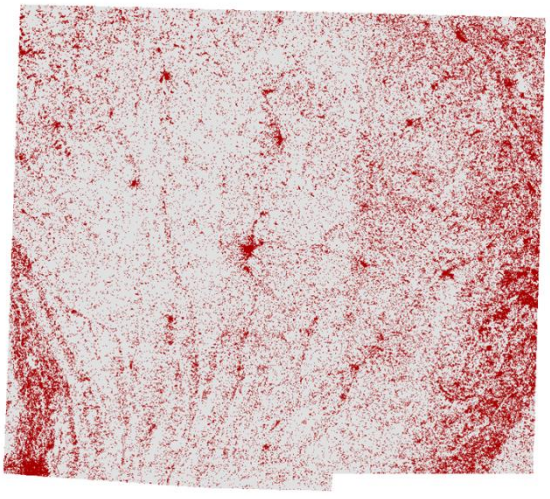
5	Selection of the 36 images – bands dataset & indices	Dempster-Shafer	Support Vector Machine/Random Forest/ Artificial Neural Network
6	One-year time series – indices metrics	Dempster-Shafer	Support Vector Machine
7	One-year time series – bands subset metrics	Dempster-Shafer	Support Vector Machine
8	One-year time series – bands dataset & indices metrics	Dempster-Shafer	Support Vector Machine

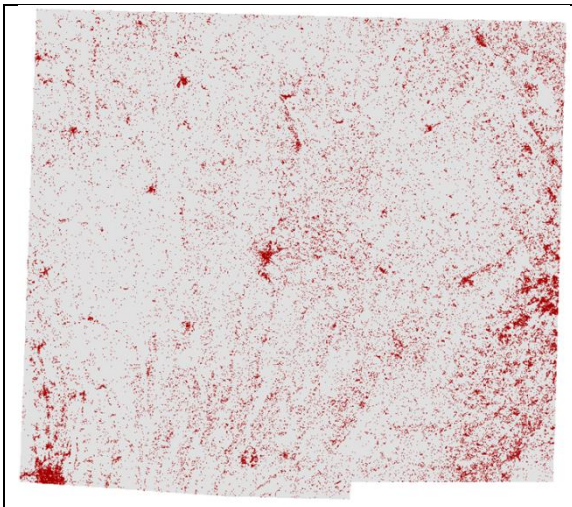
Table 3-8. Selection of the best sensor dataset based on the results given by SVM.

Test	Input Data	Fusion of classifications	Classification algorithm
1	Selection of Sentinel-2 images – all bands	Dempster-Shafer	Support Vector Machine
2	Selection of Sentinel-1 images – all bands	Dempster-Shafer	Support Vector Machine
3	Selection of Sentinel-1 and 2 images – all bands	Dempster-Shafer	Support Vector Machine

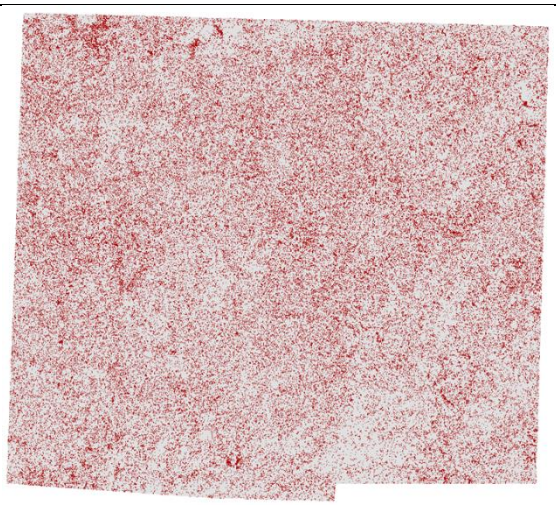
The results of the tests for the determination of the algorithms used for the Dempster-Shafer fusion of the classifications are quantified in the Table 3-10 and visually assessed in the Table 3-9.

Table 3-9. Visual check for the Demspter-Shafer fusion algorithms based on the precision rate, the recall rate, the overall accuracy and the kappa coefficient – the D-S fused result using the overall accuracy is the closest to the HRL IMD for 2015.

	
Precision rate	Recall rate



Overall accuracy



Kappa coefficient

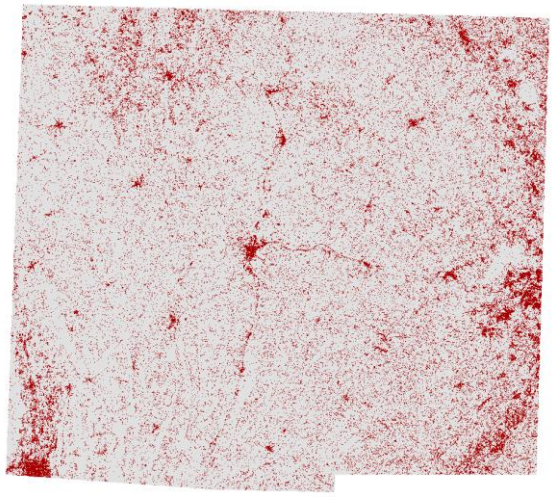
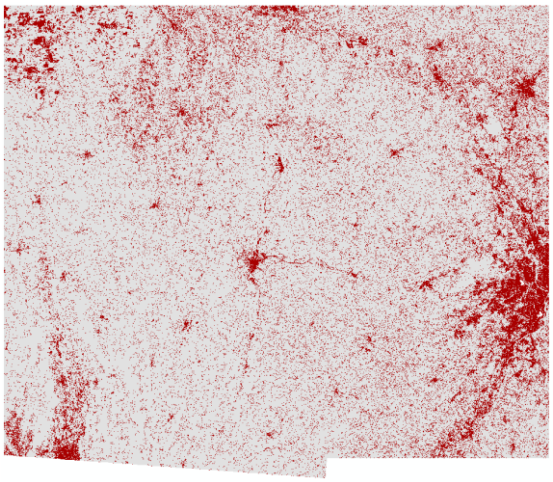
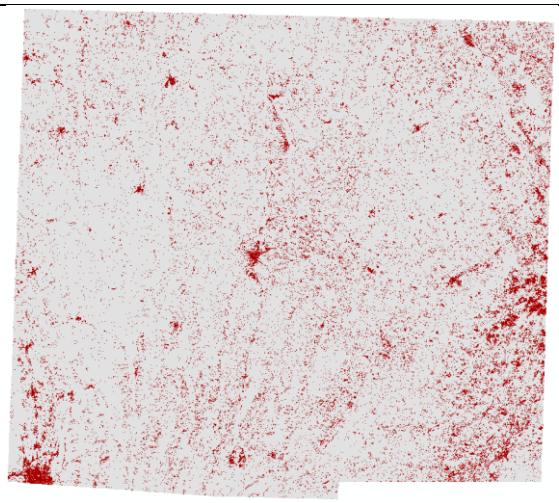
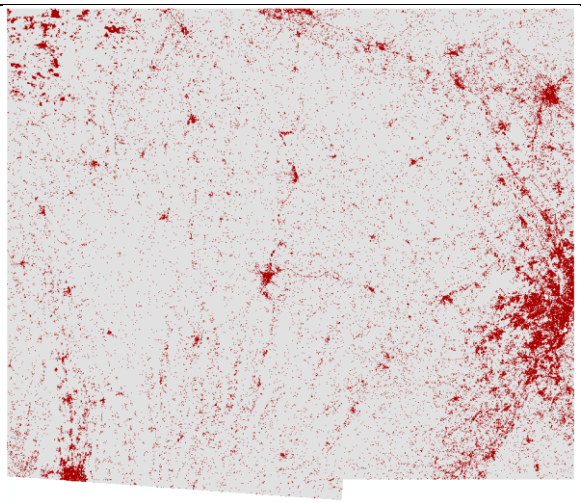
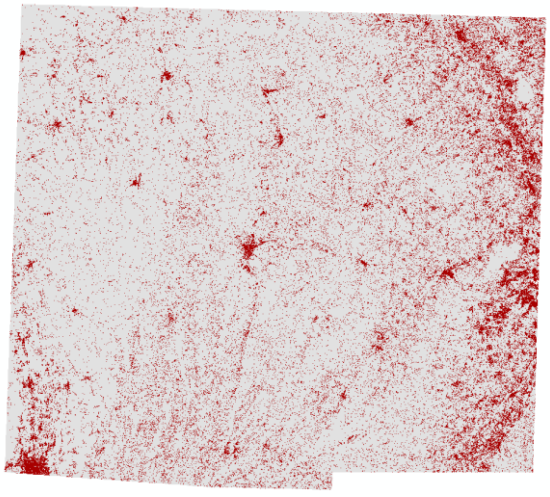
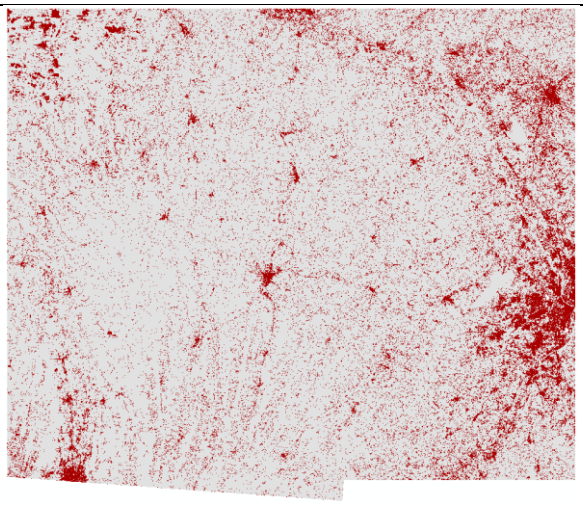
Table 3-10. User and producer accuracy for the diverse Dempster-Shafer algorithms.

Test	Fusion algorithm	Visual checking	User Accuracy	Producer Accuracy
1	Dempster-Shafer - Overall Accuracy	Yes	74.19%	88.46%
2	Dempster-Shafer – Kappa coefficient	No	42.86%	34.62%
3	Dempster-Shafer – Precision rate	Yes	63.41%	100.00%
4	Dempster-Shafer – Recall rate	Yes	65.79%	96.15%

The best algorithm for the DST data fusion tends to be the one using the “overall accuracy” metric. Indeed, there is a good balance between the user and the producer accuracies (e.g. commission and omission errors). In terms of user accuracy (commission error), the best algorithm seems to be obtained with the “overall accuracy” component. On the contrary, in terms of producer accuracy (omission error), the best algorithms are obtained with the “precision” and “recall” rates. However, it is important to note that these technics show very high level of commission errors clearly not suitable.

The results of the tests for the determination of the best classification are quantified in the Table 3-12 regarding the use of the full dataset for one year and in the Table 3-14 for a reduced dataset input while being visually assessed in the Table 3-11.

Table 3-11. Visual check for the various classification algorithms and different input datasets – the SVN classifier gives the best result compared to the HRL IMD layer for 2015.

	
RF applied on full dataset	RF applied on subset
	
SVM applied on full dataset	SVM applied on subset
	
ANNs applied on full dataset	ANNs applied on subset

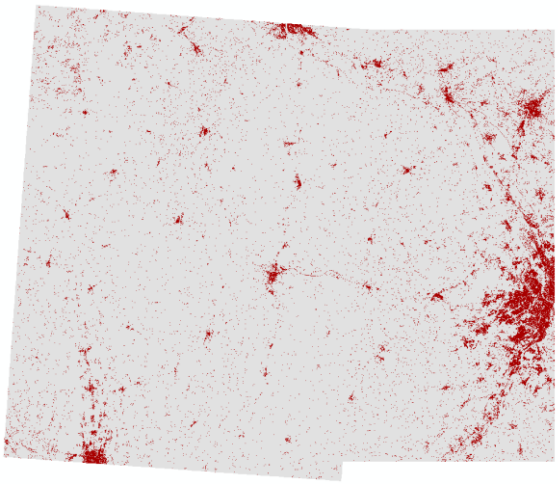
	
AL applied on subset	

Table 3-12. Full dataset of images for the yearly time series with all spectral bands results

Test	Classification algorithm	Visual checking	User Accuracy	Producer Accuracy
1	Random Forest	Yes	60.47%	100.00%
2	Support Vector Machine	Yes	74.19%	88.46%
3	Artificial Neural Network	Yes	61.76%	80.77%

Table 3-13. DLR Settlement Extent and Growth Classifier

Test	Classification algorithm	Visual checking	User Accuracy	Producer Accuracy
1	Support Vector Machine	Yes	86.21%	89.29%

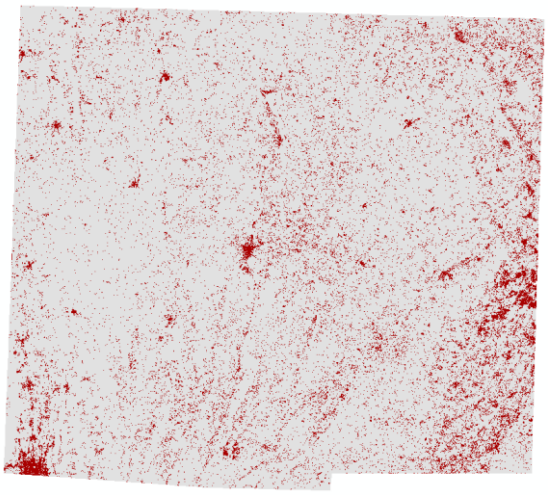
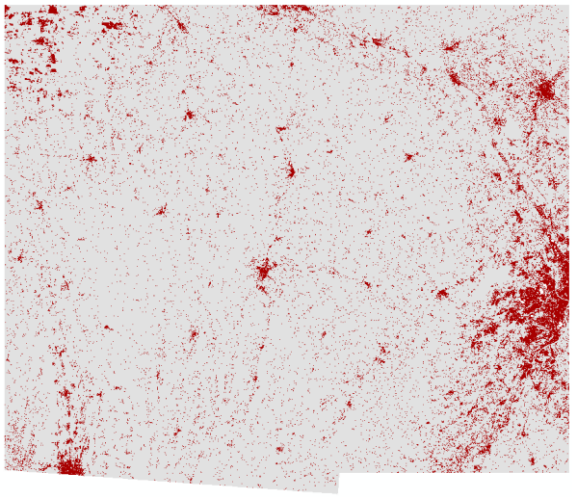
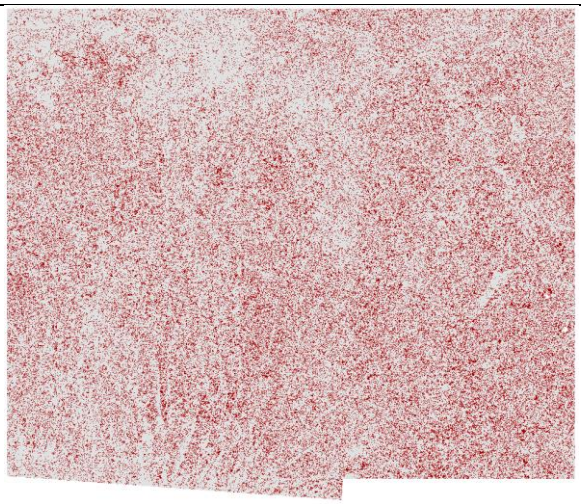
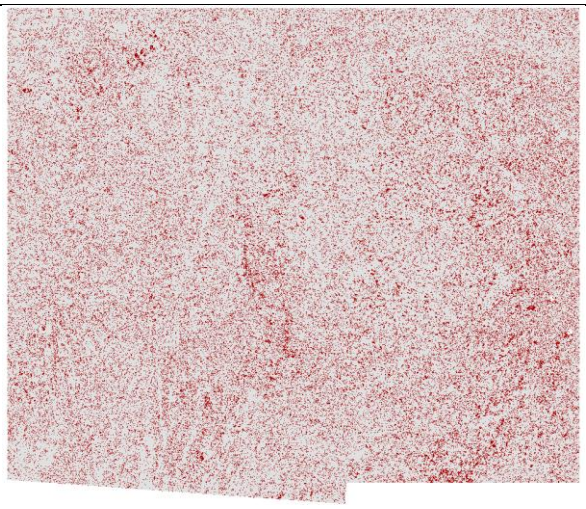
Table 3-14. Subset dataset (36 best images) with all spectral bands results

Test	Classification algorithm	Visual checking	User Accuracy	Producer Accuracy
1	Random Forest	Yes	65.85%	100.00%
2	Support Vector Machine	Yes	70.59%	88.89%
3	Artificial Neural Network	Yes	66.67%	96.30%
4	Active Learning	Yes	85.19%	85.19%

The best classifier appears to be the Active Learning which shows a good balance between the user and the producer accuracies. Then, the Support Vector Machine shows the next best results but with high commission errors. The random forest and neural network classifiers present high producer accuracy but very high rate of commission errors.

The results of the tests for the determination of the best input datasets fed to mono-temporal classifications, fused with the DS algorithm based on the overall accuracy, are quantified in Table 3-16 while being visually assessed in the Table 3-15.

Table 3-15. Visual check for different input datasets – the full dataset input gives the best result compared to the HRL IMD layer for 2015.

	
Full Sentinel-2 pre-processed dataset with all spectral bands	Selection of the best 36 Sentinel-2 pre-processed images with all spectral bands
	
Selection of the best 36 images with a spectral subset (bands 2-3-4-7-9)	Selection of the best 36 images based on spectral indices combined (NDVI and NDBI)

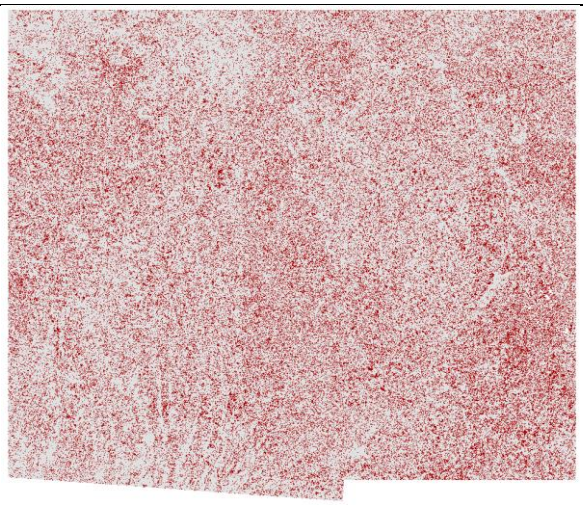
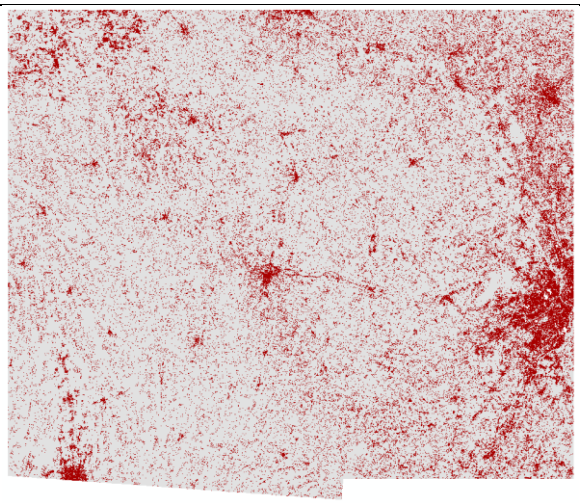
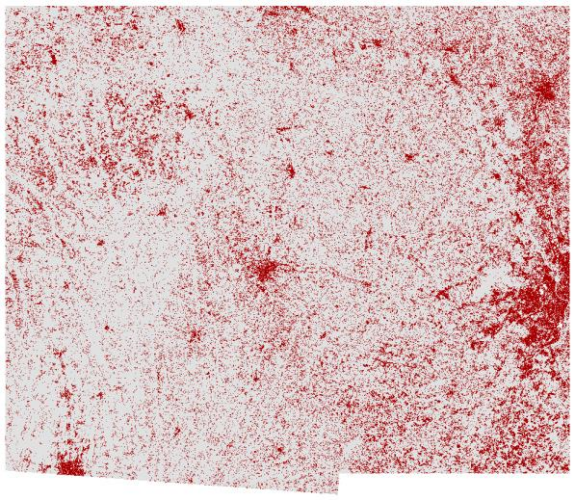
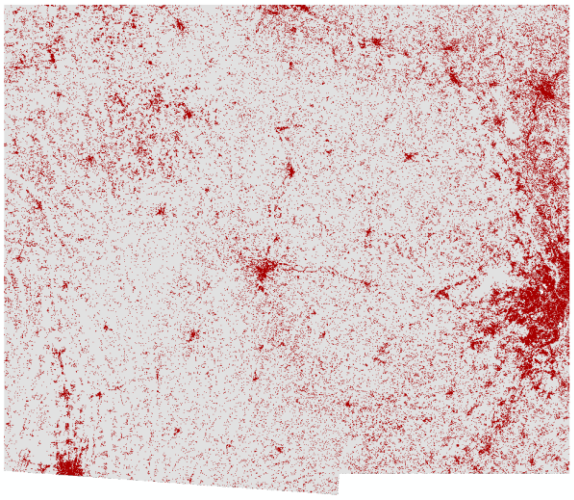
	
Selection of the best 36 images based on spectral subset (bands 2-3-4-7-9) and indices combined (NDVI and NDBI)	One-year time series based on bands subset metrics (maximum, minimum, median, standard deviation)
	
One-year time series based on indices metrics	One-year time series based on bands subset and indices metrics

Table 3-16. Overall results for the selection of the proper input data

Test	Input Data	Visual checking	User Accuracy	Producer Accuracy
1	Full dataset – all bands	Yes	74.19%	88.46%
2	Selection of the 36 images – all bands	Yes	70.59%	88.89%
3	Selection of the 36 images – bands subset	No		
4	Selection of the 36 images – indices	No		
5	Selection of the 36 images – bands dataset & indices	No		
6	One-year time series – bands subset metrics	Yes	61.90%	100.00%
7	One-year time series – indices metrics	Yes	64.10%	96.15%
8	One-year time series – bands dataset & indices metrics	Yes	65.00%	100.00%

Regarding the input data for classification, the tests show that the best set for the classification is the one with all the data pre-processed available, closely followed by the data subset with a selection of the best available cloud-free images.

The result of the imperviousness mapping is presented in Figure 3-22. The layer is a continuous raster with values between 0 and 100 indicating high (red) and low (green) density of impervious surface area.

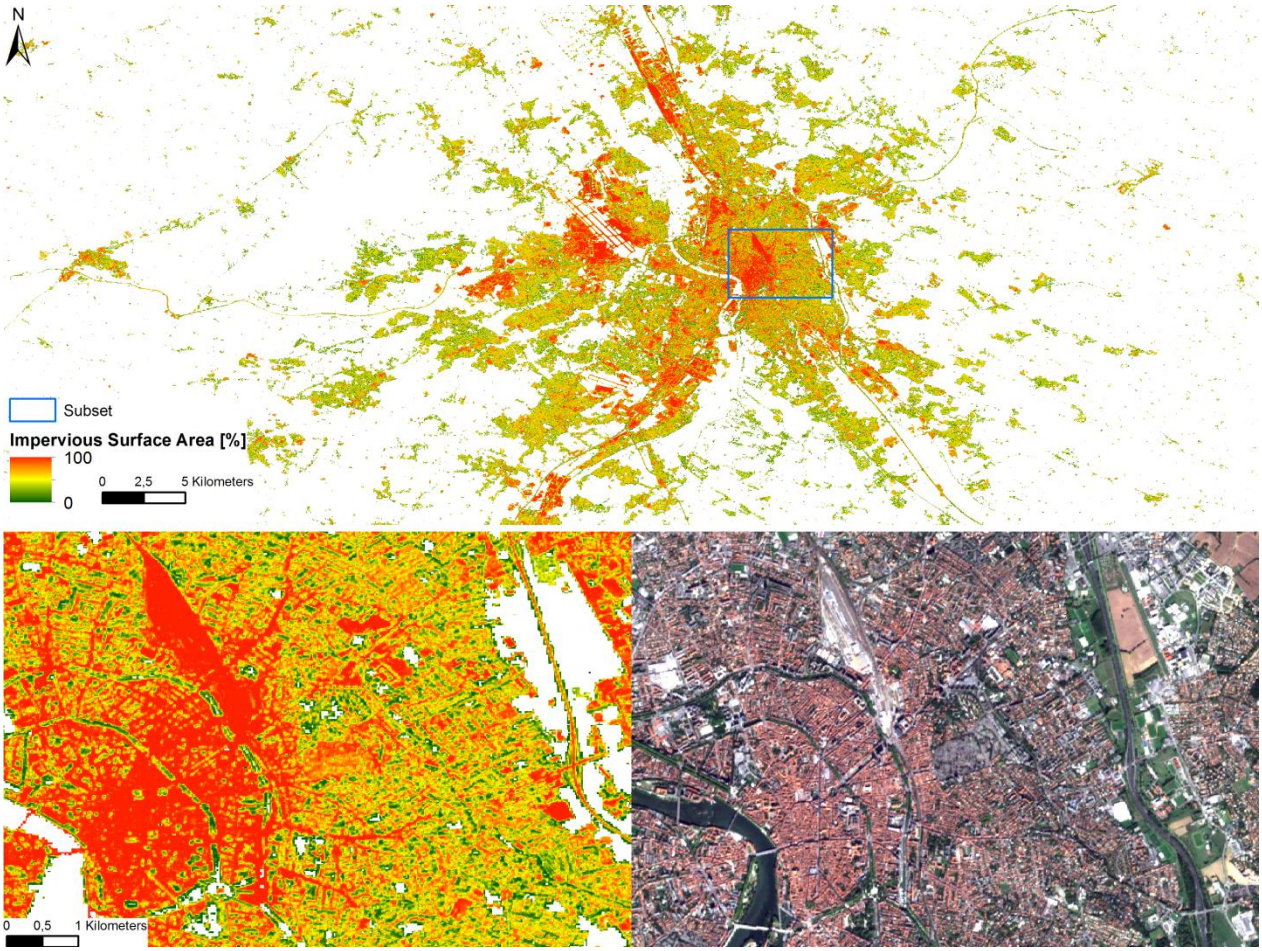


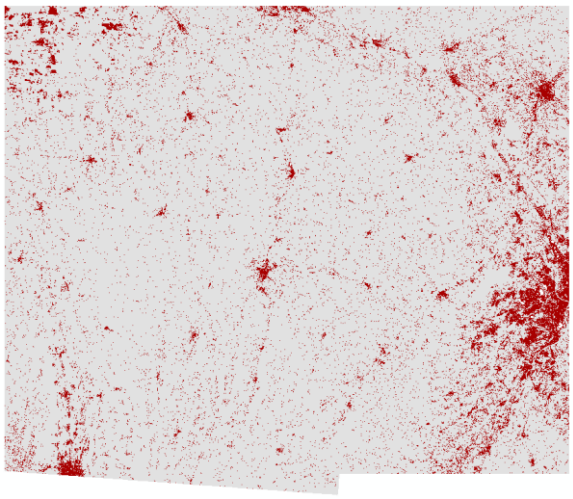
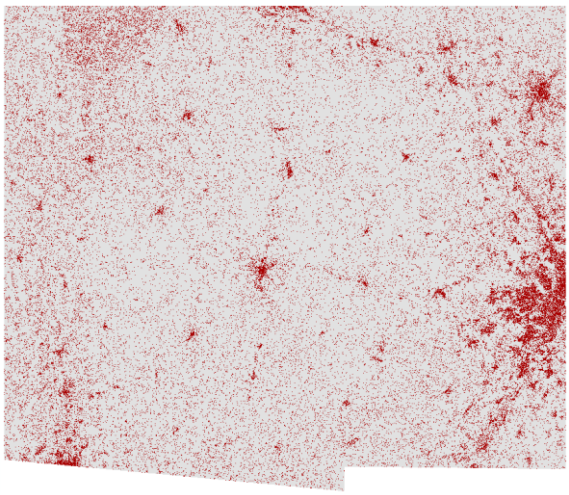
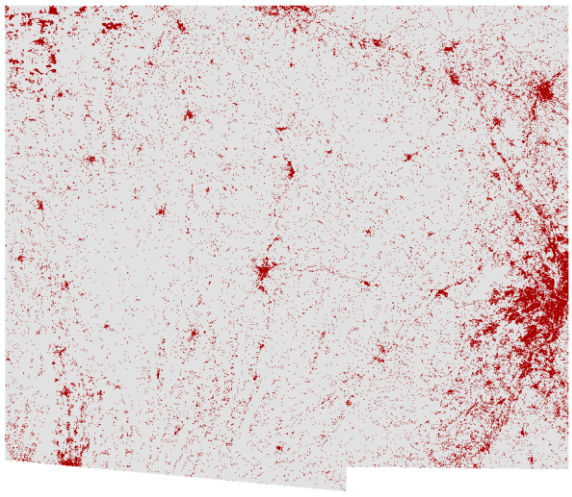
Figure 3-22. Subset of Imperviousness Layer compared with Sentinel-2 imagery.

The results of the tests for the determination of the best sensor are quantified in the Table 3-17 regarding the use of either Sentinel-1 data, or Sentinel-2, or even a combination of both time series while being visually assessed in the Table 3-18.

Table 3-17. Impact of the sensor used for the SVM classification

Test	Sensor	Visual checking	User Accuracy	Producer Accuracy
1	Sentinel-2	Yes	70,59%	88,89%
2	Sentinel-1	Yes	72,41%	77,78%
3	Fusion Sentinel-1 & 2	Yes	74,19%	85,19%

Table 3-18. Visual check for different input datasets – the combination of both time series, from S1 and S2, as input gives the best result compared to the HRL IMD layer for 2015.

	
Sentinel-2 pre-processed dataset with all spectral bands	Sentinel-1 pre-processed dataset (gamma 0)
	
Fusion Sentinel-1 and 2 pre-processed dataset	

3.2.1.4 Summary and conclusions

The analysis shows better results for the following set of parameters:

- a mono-temporal approach, image-by-image;
- the use of an active learning or SVM classifier;
- the input being all data available (or subset based on the best available cloud-free images) with both sensors Sentinel-1 and Sentinel-2.

The results are not fully compliant with the actual specifications (both 90% user and producer accuracies). Nevertheless, the results nearly meet the threshold. It should be notice that very few post-processing (mostly manual enhancement) has been applied and the results can be easily increased.

The active learning algorithm shows great classification performances whilst being very computer efficient, thus substantially reducing processing time overall and dealing with large dataset. The SVM classifier shows interesting results as an alternative method.

The approach based on both sensors Sentinel-1 and Sentinel-2 shows the interest to use data fusion. The mono-source approach, based on one HR sensor, Sentienl-1/2, doesn't seem in fact sufficient. The optical time series, in particular, is not dense enough to take advantage of the phenology of inter-yearly and intra-yearly seasonal dynamics.

Further investigations need to be undertaken for the next steps and the following WP in order to enhance the results as well as the efficiency of the classifiers. The multi-sourcing approach should be explored with not only one sensor, Sentinel-2, but also other sensors including Sentinel-3. Different studies (Pesaresi, et al., 2013), (Hansen, et al., 2013) exploit this multi-source approach to create global built-up maps with remarkable success.

These results will have a major impact on the following activities (WP 34 and 35).

3.2.2 Forest

Accurate and timely forest type mapping is essential for the assessment of a forest's biological and ecological state and the management of forest resources. The Copernicus HRL Forest has been previously produced for the reference years 2012 and 2015, and shall be updated in future productions. The following sections explore methods for improving the HRL Forest classification by exploiting the use of dense optical and SAR time series. The aim of this work is the automated classification of the dominant leaf type (DLT) in the ECoLaSS north test site, Sweden. The use of temporal-spectral metrics as classification input features is assessed and compared between Sentinel-2 and Sentinel-1 sensor data.

3.2.2.1 Description of candidate methods

Using time features (see section 3.1.4) of Sentinel-2 and Sentinel-1 data and reference samples based on an existing dominant leaf type product (HRL 2015), a new classification of broadleaf and coniferous forest for the year 2017 was delineated. Time features can capture the intensity of significant change information and statistical time series properties (section 3.1.4). A random forest classifier was applied in five experiments using different combinations of sensor data and time periods to benchmark their respective feasibility, effort and accuracy:

- Sentinel-2, full year
- Sentinel-2, spring period
- Sentinel-1, spring period
- Combined Sentinel-2 (full year) and Sentinel-1 (spring period)
- Combined Sentinel-2 (spring period) and Sentinel-1 (spring period)

3.2.2.2 Benchmarking criteria

In addition to a traditional classification accuracy assessment (Overall accuracy, class specific producer and user accuracy, kappa coefficient) several other criteria were used to evaluate the trade-off between optimal results and suitable effort or "cost" of the different experiments. These cost criteria include the estimated processing time and advantages or disadvantages specific to the sensors.

3.2.2.3 Implementation and results of benchmarking

The following section focuses on the implementation of the benchmarking process, starting with the classification input data (section 3.2.2.3.1), followed by explaining the class separability analysis (section 3.2.2.3.2), the results of the classification (section 3.2.2.3.3) and the outcome of the benchmarking process (section 3.2.2.3.4).

3.2.2.3.1 Classification input data

The ECoLaSS north test site in Sweden is comprised of the footprints of two adjacent Sentinel-2 tiles (33VVF and 33VWF) for which Sentinel-2 and Sentinel-1 data were processed. The Sentinel-2 imagery was atmospherically corrected and topographically normalized using the ESA Sen2Cor software (Louis et al. 2016). Only scenes with cloud cover lower than 50% were used for the classification and analysis. The cloud cover metric does not rely on the official metadata cloud score provided by the original Sentinel-2 Level 1C product. A cloud mask was calculated as part of the pre-processing chain using Sen2Cor to derive Level 2A data. Figure 3-23 shows the Sentinel-2 scene cloud cover distribution in the test site. Figure 3-24 shows the respective data score (inverted cloud score) for each pixel in the area of interest, which is the number of available Sentinel-2 observations with average cloud cover <50% per pixel, within the full year 2017.

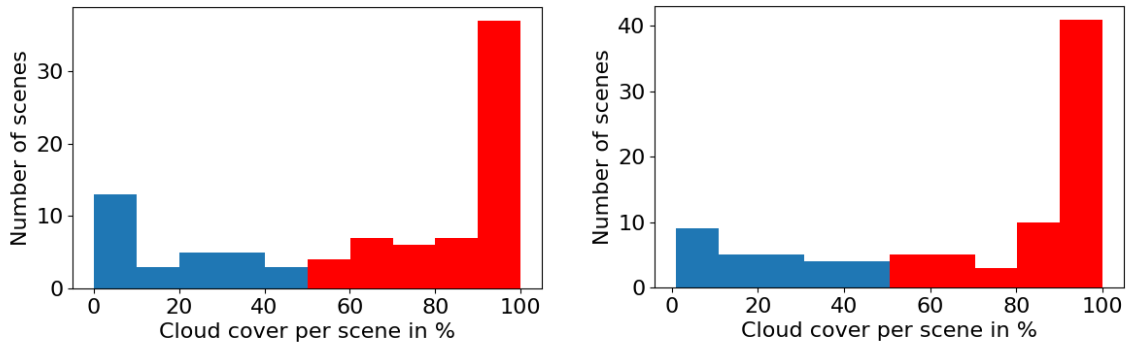


Figure 3-23. Cloud coverage of Sentinel-2 tile VVF (left) and VWF tile (right). Blue: Scenes with < 50% cloud cover.

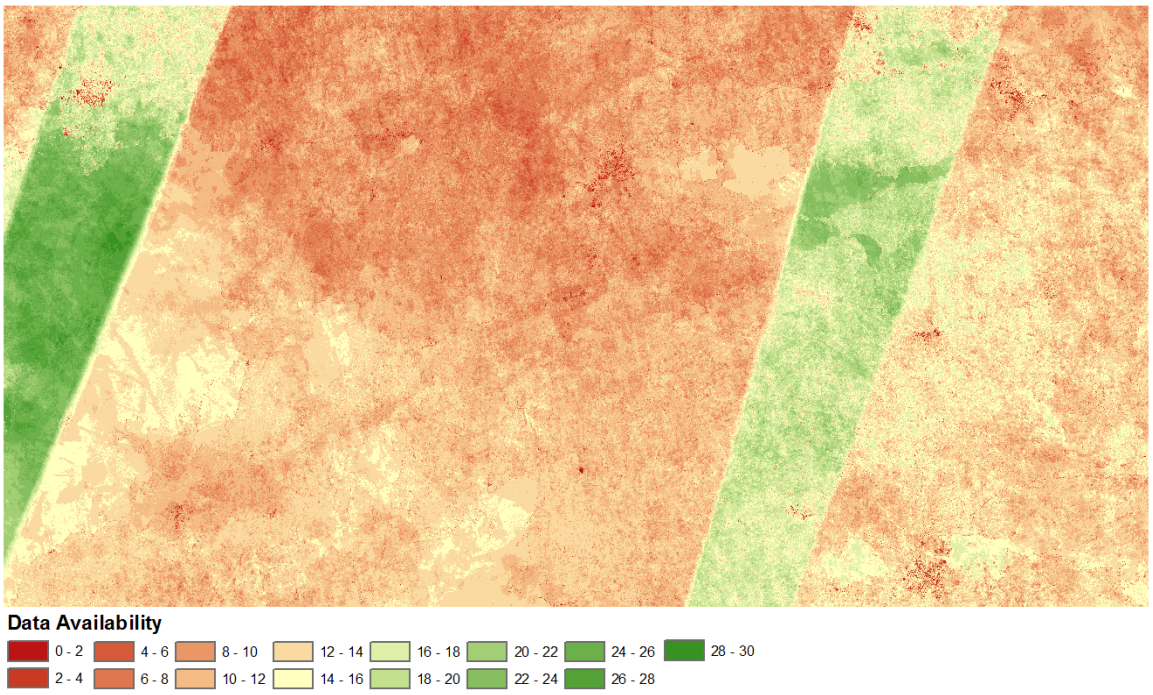


Figure 3-24. Sentinel-2 data score (number of cloud-free images) of scenes with average cloud cover <50% for ECoLaSS north test site (VWF/VVF tiles), within the full year 2017.

The large amount of scenes with strong cloud cover in the time series reinforces the need for the use of image composite-like time features (section 3.1.4). For Sentinel-2, the time series over the full year was processed, and analyzed comparatively with using Sentinel-2 data of the spring period (15. March - 15. June 2017) only. Due to preliminary research on vegetation phenology showing the limited potential for leaf type separation with Sentinel-1 data outside the spring period, the dataset was limited to the period 15. March - 15. June 2017. The Sentinel-1 GRD data (VV and VH polarization) was preprocessed to gamma naught (radar backscatter coefficient for an assumed ellipsoidal ground surface) and a multi-temporal filter was applied on the time series.

The time features described in chapter 3.1.3 were calculated for the NDVI, NDWI, Brightness and IRECI indices using the Sentinel-2 data, once for the full year 2017 and once for the spring period (15. March - 15. June 2017). For Sentinel-1 the regular time features were calculated for the same spring period, for gamma nought of the VV and VH polarizations and the normalized difference as well as the ratio of VV and VH. Additionally, the change trend features between March and June and between April and June (the expected minimum and maximum canopy cover of broadleaf forest in the spring timeframe) were calculated.

Two independent sample datasets were used for the classification and validation. The training samples for coniferous and broadleaf forest were extracted from the combined HRL 2015 Dominant Leaf Type Forest product and the HRL 2015 Grassland product (HRL 2015). Certain measures were undertaken to reduce the number of outliers and errors in these samples:

1. Reduction of edge effects and mixed pixels through negative buffering (60 m) of the HRL DLT product classes (coniferous forest, broadleaf forest and non-forest). The remaining forest patches usually represent patches of relatively homogenous leaf type.
2. Removal of patches smaller than 1 ha
3. Stratified random point sampling within the remaining forest areas
4. Removal of sampling errors through visual checks of samples
5. Iterative resampling and visual check of samples for the broadleaf forest class to match the number of coniferous forest samples
6. Creation of rectangle polygons (corresponding to 3x3 10 m pixels) from the point samples by positive buffering by 15 m

The measures applied for the creation of the training data set lead to a certain bias in the data. Samples of transitional or more heterogeneous forest cover are not well represented in the data set, limiting the validity to assess the classification success. In order to be able to evaluate the classification accuracy and consequently compare different sensors, time periods and input time features, an independent validation data set was created. This guarantees an evaluation independently from the quality of the HRL 2015 DLT product and the sample enhancement process. For that, the 2017 DLT classification layer was masked with the 2017 forest / non-forest mask (derived as part of work package 34 – Forest-Change, using the same input time features) and for each class, a sample of 110 points was randomly selected and visually interpreted. Table 3-19 shows the sample and response design for the creation of the validation dataset, Table 3-20 the distribution of sample points of the training and validation dataset.

Table 3-19. Validation dataset specifications.

Sample Design:	Stratified random point sampling (per class)
Sample Units:	Points with a 1-pixel distance to class border (to avoid border effects)
Stratification pattern	50% inside VHR-reference data extent, 50% outside; fixed # of samples for each class
Response Design	within VHR data extent: Interpretation of each sample using VHR data as primary source; Google Earth/Bing Maps as secondary data source Outside VHR extent: Google Earth/Bing Maps; selected S2 data pair (spring/summer(in leaf)) as secondary source

Table 3-20. Sample distribution of training and validation dataset.

Class ID	Class name	Training data # polygons	Validation data # points
0	Non-forest	500	127
1	broadleaf	200	62
2	coniferous	200	141

3.2.2.3.2 Class separability analysis

The ability of the different time features to separate the forest classes were evaluated by visual interpretation of box plots and the calculation of the random forest feature importance. Figure 3-25 and Figure 3-26 show boxplots of the reference pixel distribution for four important Sentinel-2 respectively Sentinel-1 time features.

Multiple Sentinel-2 time features allow for relatively good separation of broadleaf and coniferous forest, with the complex difmin features (see chapter 3.1.4.1) of several indices dominating the feature importance. This significant difference in the strongest positive change within the time series agrees with the characteristic seasonal patterns of the broadleaf forest compared to the more stable vegetation cover of coniferous forest. The various indices' difmin features are directly followed by the importance of multiple simple features, e.g. percentiles, std and max statistics of the NDVI, NDWI and IRECI indices, whereas the brightness features are less significant.

Compared to Sentinel-2, the box plots of the Sentinel-1 time feature show inferior separability, especially for the VV-polarization. The highest importance by far can be attributed to the VH change trend and the closely following VH difmin features, confirming the high importance of the strong seasonal value difference between coniferous and the more seasonal broadleaf forest. The difmin time feature considers the full time series, and the change trend feature selected scenes. Both capture similar information (the strongest value delta in the time series), however, the latter based on selected scenes shows a slightly higher feature importance. This can be attributed to the temporal window size of the difmin feature, as only scenes with a limited time distance are compared. Other simple Sentinel-1 time features show vastly lower feature importance for the forest class separation.

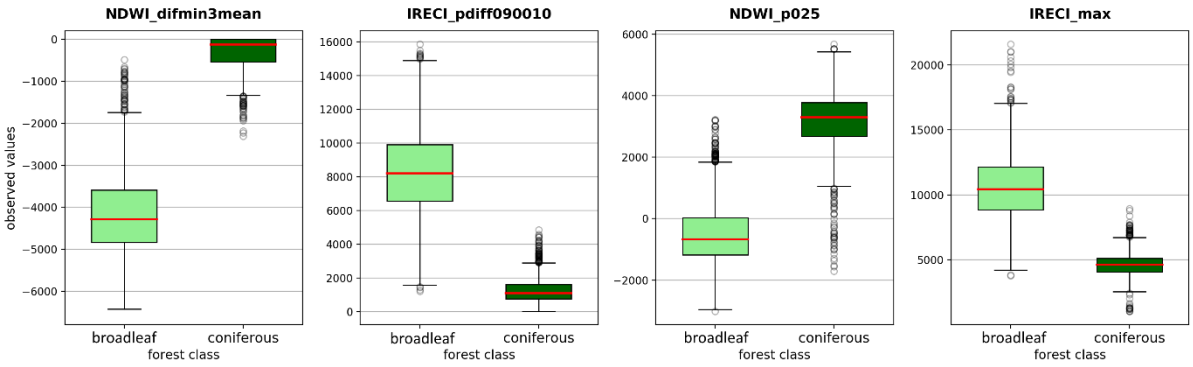


Figure 3-25. Forest class separability box plots for selected Sentinel-2 time features.

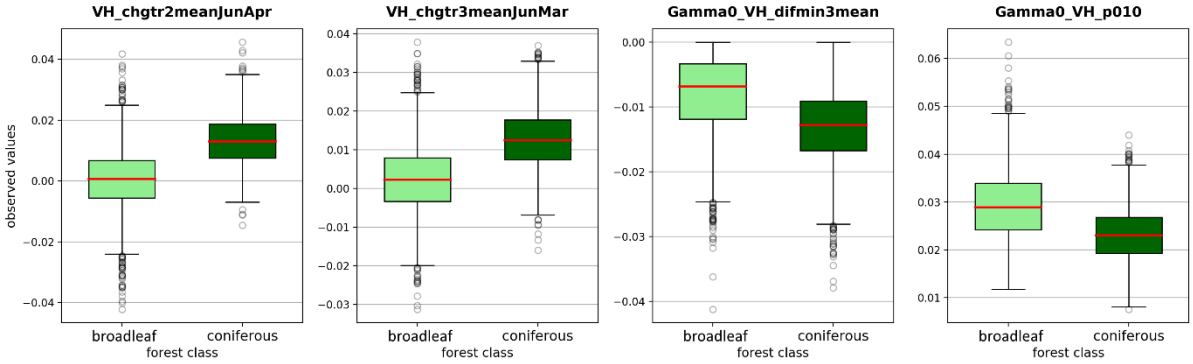


Figure 3-26. Forest class separability box plots for selected Sentinel-1 time features.

3.2.2.3.3 Classification results

The classification was carried out using a random forest classifier with preceding recursive feature elimination. The general accuracy metrics of the different classification configurations can be seen in Table 3-21 and Figure 3-27, while Table 3-22 shows the class specific results.

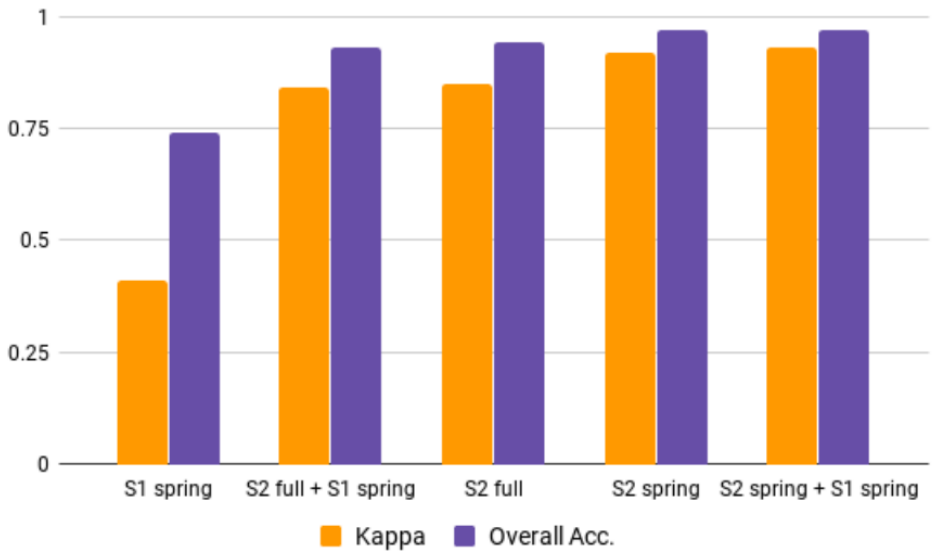


Figure 3-27. Kappa and overall accuracy for the five DLT input data configurations.

Table 3-21. Accuracy metrics for the five DLT input data configurations.

	Users Acc.	Producers Acc.	Overall Acc.	Kappa
S1 spring	0.75	0.74	0.74	0.41
S2 full + S1 spring	0.93	0.93	0.93	0.84
S2 full	0.94	0.94	0.94	0.85
S2 spring	0.97	0.97	0.97	0.92
S2 spring + S1 spring	0.97	0.97	0.97	0.93

Table 3-22. User and producer accuracy of broadleaf and coniferous forest for the five DLT input data configurations.

	Users Acc.		Producers Acc.	
	Broadl. Forest	**Con. Forest**	**Broadl. Forest**	**Con. Forest**
S1 spring	0.57	0.83	0.63	0.79
S2 full + S1 spring	0.88	0.96	0.90	0.94
S2 full	0.88	0.96	0.92	0.94
S2 spring	0.91	0.99	0.98	0.96
S2 spring + S1 spring	0.92	0.99	0.98	0.96

The classification using Sentinel-2 time features, both for the full year and spring period, are able to successfully differentiate between broadleaf and coniferous forest with an overall accuracy of 97%. This is also considering the partially challenging forest geography in the area of interest with a multitude of forest stand ages and densities due to frequent tree harvesting. Interestingly, the shorter spring data period offers slightly better results than using the full year 2017. The differentiation of broadleaf and coniferous forest vastly depends on the seasonal pattern of broadleaf forest, and the spring period captures this period of biggest variance.

As expected from the separability analysis, the Sentinel-1 classification is far less successful with 74% overall accuracy and a very low Kappa of 0.41 representing a strong mismatch between the class specific accuracies. This is due to frequent misclassifications of broadleaf forest being incorrectly detected as coniferous forest, resulting in lower producer's and user's accuracies. The combination of Sentinel-2 and Sentinel-1 features does not add any significant gain in accuracy. Figure 3-28 shows a detailed view of the Sentinel-2 and Sentinel-1 spring period classification map.

Figure 3-28. Classification result detail view of 33VVT tile for Sentinel-2 spring (mid), Sentinel-1 spring (right) compared to Sentinel-2 NIR-R-G false colour composite (left).

3.2.2.3.4 Benchmarking

The following table gives a summary of how classification results (accuracy) relate to processing costs. Furthermore, scenario-specific chances and problems are listed. The two input-data scenarios with the highest achieved accuracies are "S2 spring" and "S2 spring + S1 spring". Both reach very high Kappa values of 0.92 and 0.93, respectively. While accuracies are comparable, the processing cost for "S2 spring" is roughly one half of the other scenario. It therefore can be concluded that "S2 spring" offers a very good balance between cost and benefit. The preprocessing of the Sentinel-2 data via the automated processing chain took about 2 days per Sentinel-2 tile (including atmospheric correction, topographic normalization, resampling, indices calculation and time feature calculation) and about 4 days for Sentinel-1 (calibration, terrain flattening and correction, multitemporal filtering, ratio calculation and time feature calculation). In case of the test site, the problem of clouds and cloud shadows in the imagery is mitigated by the use of time features: for 100% of the forested area in the test site a classification decision could be made. This, however, is always dependent on cloud cover situations specific to a region and the particular year. The addition of S1 data to the data scenario might only be required when the cloud cover/data availability situation is even more difficult than for the test site.

Table 3-23. Benchmarking criteria, and chances and problems of the different experiment setups

	Accuracy (Kappa)	Processing cost	Chances	Problems
S1 spring	0.41	+	Independent from cloud cover	SAR inherent properties (foreshortening, layover in strong relief, speckle)
S2 full + S1 spring	0.84	+++++	Dependent on cloud cover, but SAR and S2 time features mitigate problematic	Clouds/cloud shadows, SAR inherent properties (foreshortening, layover in strong relief, speckle)
S2 full	0.85	++++	Dependent on cloud cover, but time features mitigate problematic	Clouds/cloud shadows
S2 spring	0.92	+	Dependent on cloud cover, but time features mitigate problematic	Clouds/cloud shadows

S2 spring + S1 spring	0.93	++	Dependent on cloud cover, but SAR and S2 time features mitigate problematic	Clouds/cloud shadows, SAR inherent properties (foreshortening, layover in strong relief, speckle)

3.2.2.4 Summary and conclusions

This work investigates the potential of combining Sentinel-2 and Sentinel-1 data for the delineation of a dominant forest leaf type product in Sweden. The random forest classification uses spatio-temporal input features that capture important time series properties and patterns. Considering the limited availability of cloud free satellite scenes and heterogeneous character of the analysed forest types in the area of interest, the results are very promising for future application on larger areas. The use of Sentinel-2 data limited to the spring period provided the best ratio of high accuracy and lowest benchmarking cost. This finding could potentially benefit future dominant leaf type product generation, but requires further research to validate the transferability to areas of different geographic conditions and seasonal patterns. Although the nominally highest DLT accuracy was provided by the combined use of Sentinel-2 and Sentinel-1 time features, the gain compared to only focusing on Sentinel-2 data was insignificant. Regarding this test case, it would not justify the enormous overhead for the preprocessing, time features calculation and data handling of additional Sentinel-1 data. However, Sentinel-1 data on its own shows moderate predictive performance and would be a viable input data complement if the test area is even stronger affected by cloud cover in optical satellite imagery.

The presented methodology offers room for further improvements and additional research, e.g. the detailed feature analysis of (i) misclassified areas or (ii) pixels with low classifier probabilities. This could lead to the development of additional time features more specifically tuned to differentiating forest types and respective edge cases and thus improve the classification accuracy.

The examined methodology could also benefit future HRL forest products, especially in areas strongly affected by cloud cover, e.g. northern European countries. Compared to systematically combining a number of scene classifications in the HRL 2015 project, the use of time features could offer a more streamlined workflow and potentially a spatially more consistent classification result over multiple satellite image tiles.

3.2.3 Grassland

Methods for large area mapping of grasslands at an operational level often do not provide a sufficiently high accuracy level because of the strong variation of grassland surface (natural, semi-natural, agricultural), its diversity in grassland management practices as well as a spectral overlap with croplands. With the availability of Sentinel-1 and Sentinel-2, providing data in short revisit intervals and large coverage, grassland mapping will profit from the availability of the dense time series. The ECoLaSS consortium is addressing this topic and is developing a supervised classification approach based on dense time series data from Sentinel-1 and Sentinel-2, performed separately for main biogeographic regions in Europe and using in-situ data such as LUCAS (Land Use/Cover Area frame Statistical Survey) and visually interpreted reference plots.

This section deals with automated grassland mapping based on integrated Sentinel-1 SAR and Sentinel-2 multispectral optical time series data. In this context grasslands considered, are covered by Herbaceous vegetation with at least 30% ground cover, which includes at least 30% graminoid species such as Poaceae, Cyperaceae and Juncaceae (see Table 2-2). One of the major challenges of past pan-European high resolution optical satellite image coverages has been data gaps due to high-frequency cloud cover and/ low solar incidence angles. The availability of Sentinel-2 satellite(s) significantly improves the data situation. Nevertheless, due to heavy cloud cover over specific regions alternative image data sources such as SAR are included. Therefore, in this chapter, the usage of Sentinel-1 as alternative image data and how

to combine and integrate SAR (Sentinel-1) and optical (Sentinel-2) are addressed. Methods are developed and tested how to use Sentinel-1 SAR data to close data gaps from optical image sources and as complementary information (to Sentinel-2) for increasing the thematic classification accuracy. In task 4 of the project, the methods will be applied on a larger scale over the demo sites. The results are compared to other existing pixel-based approaches in terms of classification accuracy and processing time.

3.2.3.1 Description of candidate methods

CLASSIFICATION METHODS

THRESHOLD SCHEMES

In order to classify grasslands a set of training areas were ascertained and from these samples, the signatures were extracted. The signatures of the classes grassland/no-grassland cluster show a good differentiation, enabling a successful grassland/no-grassland classification based on thresholding to the different features. Grasslands detection by Sentinel-1 data is based on VV polarization annual statistics applying minimum and maximum thresholds for VV annual mean and for VV annual coefficient of variation for years 2016 and 2017. The thresholds are derived by a 95% fitting of 700 grassland reference plots manually selected from 2017 VHR imagery at the demo site.

RANDOM FOREST

The Random Forest (RF) classifier first proposed by BREIMAN 2001 belongs along with other boosting and bagging methods as well as classification trees in general to the ensemble learning methods, which generate many classifiers and aggregate their results to calculate their response (Liaw and Wiener, 2002; Horning et al., 2010; T. Li et al., 2016). The random forest algorithm generates multiple decision trees with randomly drawn subsets, instead of using all variables from the available data. The subsets are drawn with replacement, meaning that one sample can be selected several times, while others may not be selected at all (Belgiu and Dragut¸, 2016; Ali et al., 2012). Regarding each random sample, a classification or regression tree is grown to the largest possible extent without pruning. At each node, a random sample of a predictor variable is extracted; among those, the best split is chosen. To predict new data the prediction among all trees are aggregated using majority votes. The class with the maximum vote overall decision trees is the one selected for the output product (Liaw and Wiener, 2002; Ali et al., 2012). One advantage of the classifier is the calculation of the variable feature importance. In this context, the relative importance of variables is calculated for each feature available for both optical and SAR data.

For training and validation, 3408 LUCAS points covering the Belgium site are visually interpreted based on the Sentinel-2 time series data from 2016 till 2017. The land cover classes interpreted are shown in Table 3-27. The interpreted points were randomly split into training and validation data sets at a ratio of 66% training to 33% validation. Furthermore, the eight land cover classes are aggregated to grassland / non grasslands classes. With the aggregated classes the random forest model has been trained using temporal and spectral variables with the same input parameters with the number of trees set to 500 and the number of variables to the square root of the total number of input variables. From the training models the Mean Decrease Impurity measure is calculated for each feature based on the aggregated classes. Finally, the output classifications are treated as thematic layers and validated against the remaining points not used for training using a point-based method. The accuracy is assessed with confusion matrices and accuracy metrics.

SAR time series products are based on Level-1 products in Interferometric Wide swath (IW) mode and Level-1 Ground Range Detected (GRD). The IW mode is considered the main acquisition mode over land and satisfies the majority of service requirements. For each Sentinel-1 orbit, the pre-processing is calculated separately as multi-temporal filtering can only be applied to images of the same orbit. In addition, a local incidence file is calculated for each orbit stack and delivered with the data [see WP32 report]. Additionally, temporal image stack statistics have been calculated which are used as input data

for the time series classification processing chains. The temporal SAR features generated are described in detail in Table 3-24.

Table 3-24. Derived annual features based on the SAR time series.

feature	description
MIN	Minimum
MAX	Maximum
MEAN	Mean
STD	Standard deviation
CoV	Coefficient of Variation
DIFF	difference between the mean of the first three images and the mean of the last three images of the defined time period – useful for assessing phenological changes in seasonal/monthly image stacks or annual changes for low variance land cover classes (e.g. forests) in annual stacks

OPTICAL TIME SERIES PRODUCTS

The Sentinel-2 sensor system has an overall number of 12 bands from 10m to 60m spatial resolution. For the ECoLaSS processing only the 10m and 20m bands are used, which are in total 10 bands. The list of the used bands with their central wavelengths and abbreviations is shown in Table 3-25.

Table 3-25. Used Sentinel-2 reflectance bands (adapted from Suhet, 2015).

Sentinel-2 Bands	Description	Central Wavelength (µm)
Band 2	Blue	0.490
Band 3	Green	0.560
Band 4	Red	0.665
Band 5	Vegetation Red Edge (VRE1)	0.705
Band 6	Vegetation Red Edge (VRE2)	0.740
Band 7	Vegetation Red Edge (VRE3)	0.783
Band 8	NIR	0.842
Band 8A	Narrow NIR (NNIR)	0.865
Band 11	SWIR (SWIR1)	1.610
Band 12	SWIR (SWIR2)	2.190

Following vegetation indices are derived from the Sentinel-2 data sets for each image and are used for further processing steps, as e.g. the calculation of the median for a specific reference period.

Table 3-26. Used vegetation indices. Xue, J., & Su, B. (2017); Lagunas et. al. (2015)

Index abbreviation	Index name
CI_green	Green Chlorophyll Index
CI_red_edge	Red Edge Chlorophyll Index
EVI	Enhanced Vegetation Index
MCARI_705_740	Modified Chlorophyll Absorption Ratio Index
MTCI	MERIS Terrestrial Chlorophyll Index
NBR	Normalized Burn Ratio
NDMI	Normalized Difference Moisture Index
NDRE1	Normalized Difference Red Edge Index (1)
NDRE2	Normalized Difference Red Edge Index (2)
NDVI	Normalized Difference Vegetation Index
OSAVI_705_740	Optimized Soil-Adjusted Vegetation Index
REP	Red-Edge Position
SAVI	Soil-Adjusted Vegetation Index
TCB	Tasseled Cap Brightness
TCG	Tasseled Cap Greenness
TCW	Tasseled Cap Brightness

Based on the reflectance bands and the vegetation indices annual features like median, mean, maximum, minimum and standard derivation are derived used as classification input. In task 4 of ECoLaSS the feature importance will be calculated for all vegetation indices mentioned in Table 3-26 to evaluate which vegetation indices are most suitable for the grassland detection.

3.2.3.2 Benchmarking criteria

REFERENCE DATA

The first reference data set used is "Landbouwgebruikspercelen ALV, 2016 "(LGP) provided by the Departement Landbouw en Visserij. The dataset presents a polygon-wise assessment for the year 2016, differentiating between several agricultural areas including cultivation crops and grasslands. Since the reference data set was composed for agricultural purposes this reference data set does not include following features, which are included within the grassland definition (see Table 2-2).

- Grasslands in urban areas: parks, urban green spaces in residential and industrial areas, sport fields, golf courses
- Natural grasslands on military sites, airports
- Grasslands on land without use
- Semi-arid steppes with scattered Artemisia scrub
- Coastal grasslands, such as grey dunes and salt meadows located in intertidal flat areas with at least 30% graminoid species of vegetation cover

Another reference dataset has been created by Joanneum Research through visual interpretation. The dataset is based on the LUCAS points located on the Belgium site. The reference for the interpretation is the Sentinel-2 data from 2017 and 2016. A Minimum Mapping Unit (MMU) of 30m x 30m has been applied in the interpretation process. Additionally, high resolution data like Bing maps (ArcGIS Basemap layer, RGB imagery) or Arc2Earth imagery (Google commercial ArcGIS plugin, RGB imagery) have been used. Following classes have been interpreted (Table 3-27):

Table 3-27. VIRP reference dataset codes.

Class code	Class label
1	Cropland
2	Grassland
3	Forest and Trees
4	Shrubs
5	Artificial Surfaces & Associated Area(s)
6	Bare Area(s)
7	Waterbodies, Snow and Ice
8	Wetlands

The visual interoperation reference plots (VIRP) and the LGP polygons overlap partly in the granules UFS and UFR of the demo site).

FEATURE IMPORTANCE/SELECTION:

The random forest algorithm offers two methods for feature selection and importance measurements. The first is the mean decrease impurity measure and second the mean decrease accuracy measure (Breiman, 2001).

MEAN DECREASE IMPURITY

Within the forest generation every node in the decision trees is a condition on a single feature to split the dataset. The Mean Decrease Impurity (also known as Gini importance) measure, calculates the sum of the total impurity reductions at all tree nodes where the variable appears (Breiman, 2001). Therefore, each feature importance represents the sum over the number of splits across all trees that include the feature, proportionally to the number of samples it splits (Louppe et. Al, 2013). One drawback of this method is that the mean decrease impurity measure is biased towards preferring variables with more categories. Another drawback is when the dataset is composed of correlating features, which can be assumed to have the same importance. Nevertheless, the first feature analysed reduces the importance of other correlating features (Louppe et. Al, 2013).

MEAN DECREASE ACCURACY

Another feature selection method is the Mean decrease accuracy, which measures the accuracy reduction on out-of-bag samples when the values of the variable are randomly permuted (Breiman, 2001). In other words, the relative change in classification accuracy between the permuted values is calculated. After each permutation, the mean decrease accuracy measures the effect of the permutation on the model accuracy. Regarding less important variables, the mean decrease accuracy measurements should show no effect on the model accuracy in contrast to the important features. One drawback is that the estimates are biased if the predictor variables are highly correlated (Genuer et. al., 2010).

THEMATIC ACCURACY

The thematic accuracy assessment is performed by comparing the classified grassland products with one of the above mentioned reference data sets. The main purposes of the accuracy assessment and error analysis are to permit quantitative comparisons between several methods (Congalton, 1991). Maps produced from different input images classified with different methods will be evaluated using a point-by-point comparison. The thematic accuracy of the classification results will be assessed with an error matrix and following accuracy metrics:

- Overall Accuracy and Error
- User's accuracy
- Producer's accuracy
- Kappa Coefficient

3.2.3.3 Implementation and Results of Benchmarking

This chapter is focusing on benchmarking the time series classification methods for grasslands. The classification methods applied are threshold schemes and Random Forest classifier. The main focus of the benchmarking lies in the evaluation of different temporal input features for the classification approaches which are based on spectral information. These input features are derived from SAR and optical time series data.

3.2.3.3.1 Comparison of reference data sets (2016)

It is necessary to compare both reference data sets because it is important to assess the quality of these sets. Therefore, only VIRP points located within the LGP polygons are compared with each other as shown in Table 3-28.

Table 3-28. Reference data comparison (LGP2016 vs VIRP2016).

		LGP2016		
		Grassland	Others	Total
VIRP2016	Grassland	144	6	150
	Others	13	283	296
	Total	157	289	446

Overall Agreement [%]	95.74
Kappa	0.91

Differences can be observed between the two data sets due to different grassland definitions. The LGP polygons do not include urban green areas like gardens or parks, whereas the interpreted VIRP points follow the grassland definition described in Table 2-2, including urban grasslands.

Both reference data sets are representing different geometry types. The newly interpreted LUCAS points (VIRP) present pointwise assessment, whereas the LGP shapefile present a polygon/parcel based assessment. Within the pointwise assessment method shrubs within a grassland parcel are labelled as grassland if the major part of the MMU (900m^2) is covered by grassland. There is a 96% overall agreement and 94% grassland class agreement between VIRP2016 and LGP2016. For both reference data sets, misclassifications could be observed at parcel borders with mixed pixels in the satellite imagery.

3.2.3.3.2 Threshold-based grassland classification with SAR Data

VIRP2016 vs SAR2016

The SAR2016 features are derived from the entire year 2016 including the Sentinel-1B scenes, which are available since May 2016 within total of 36 images. All images are representing one orbit (ascending 161) with a VV polarization. The stack represents six different features (Minimum, Maximum, Mean, Standard derivation, Coefficient of variation and the difference between the first three images and the last three images of the time period).

The classification is based on thresholding for the features "Mean" and "Coefficient of Variation" of the annual stack. The thresholds were derived by a 95% fitting of 700 grassland reference plots for the year 2017.

Table 3-29. Error matrix: VIRP2016 VS SAR2016

		Classification			
		Grassland	Others	Total	PA [%]
Ground Truth	Grassland	439	167	606	72.44
	Others	244	2516	2760	91.16
	Total	683	2683	3366	
	UA [%]	64.28	93.78		

Overall Accuracy [%]	87.79
Kappa	0.61

LGP2016 vs SAR2016

Table 3-30. Error matrix: LGP2016 VS SAR2016

		Classification			
		Grassland	Others	Total	PA [%]
Ground Truth	Grassland	87	63	150	58
	Others	19	277	296	93.58
	Total	106	340	446	
	UA [%]	82.08	81.47		

Overall Accuracy [%]	82.96
Kappa	0.61

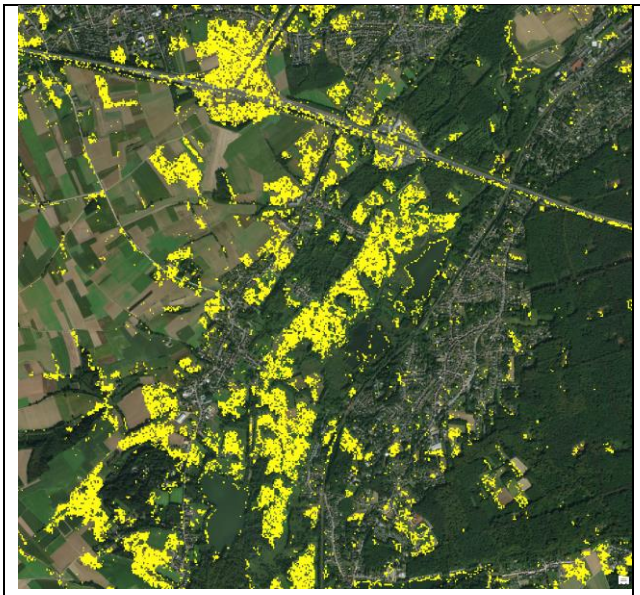


Figure 3-29. SAR grassland threshold-based classification for 2016 (grassland in yellow).

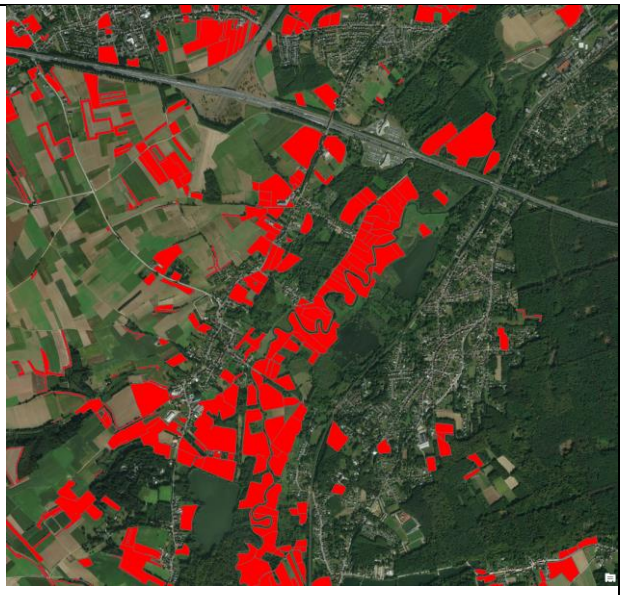


Figure 3-30. LGP grassland areas in red. Basis layer: ArcGIS Basemap.

The thematic accuracy for the SAR threshold classification 2016 compared with the visually interpreted reference plots is slightly higher (see

TABLE 3-29 and Table 3-30) due to the fact that the LGP reference dataset does not include urban grasslands.

Figure 3-29 and Figure 3-30 both show the classification compared with the LGP polygons from 2016. The threshold based approach results on inhomogeneous patches resulting from the speckle noise in the SAR data. Confusion with roads and shore areas are present.

VIRP2017 vs SAR2017

The SAR2017 image stack is derived over the year till 15.11.2017 embracing 52 different images. All images are representing one orbit (asc161) and the VV polarization. The stack represents six different features (Minimum, Maximum, Mean, Standard derivation, Coefficient of variation and the difference between the first three images and the last three images of the time period).

Again the classification is based on thresholding for the features "Mean" and "Coefficient of Variation" of the annual stack. The thresholds were derived by a 95% fitting of 700 grassland reference plots.

Table 3-31. Error matrix: VIRP2017 VS SAR2017

		Classification			
		Grassland	Others	Total	PA [%]
Ground Truth	Grassland	523	166	689	75.91
	Others	170	2507	2677	93.65
	Total	693	2673	3366	
	UA [%]	75.47	93.79		

Overall Accuracy [%]	90.02
Kappa	0.69

LGP2016 vs SAR2017

Table 3-32. Error matrix: LGP2017 VS SAR2017

		Classification			
		Grassland	Others	Total	PA [%]
Ground Truth	Grassland	105	45	150	70
	Others	31	265	296	89.53
	Total	136	310	446	
	UA [%]	77.2	85.48		

Overall Accuracy [%]	82.96
Kappa	0.61

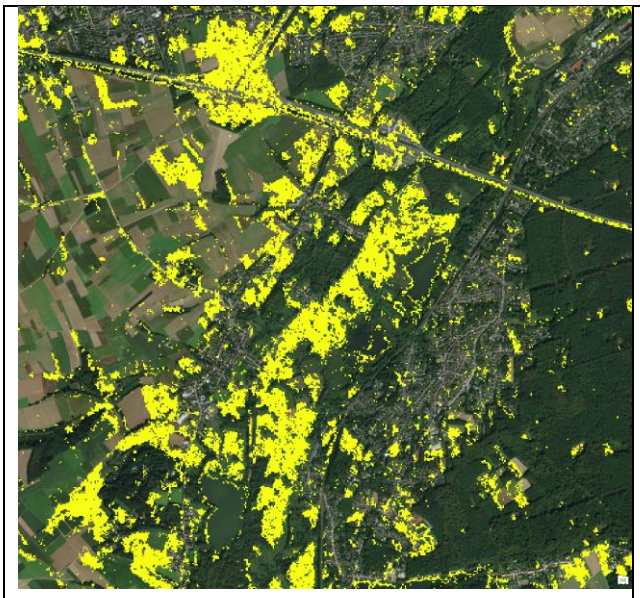


Figure 3-31. SAR grassland threshold-based classification for 2017 (grassland in yellow).

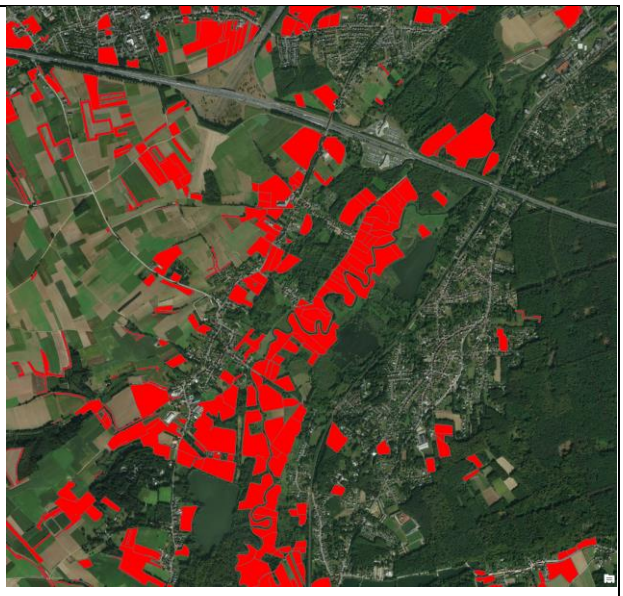


Figure 3-32. LGP grassland areas in red. Basis layer: ArcGIS Basemap.

Figure 3-29 and Figure 3-31 show the threshold based classification approaches results compared with the agricultural grassland features. As both figures show the SAR data classification approach tends to less homogeneous patches due to the speckle noise in the SAR data and to misclassify streets and other roads. Nevertheless, there is only a small confusion between grasslands and agricultural fields.

The SAR2016 threshold based grassland classification is less accurate (Table 3-31 and Table 3-32 compared with and Table 3-30) due to fewer data sets in the growing season. Winter scenes are not used for the classification purpose. Furthermore, the used thresholds are derived based on 2017 data sets and transferred to 2016 dataset without adjustment. Whereas, the SAR2017 threshold based grassland classification shows better classification results. The dataset for 2017 includes more images than 2016 and fewer winter scenes. Furthermore, the thresholds are derived and optimized for the 2017 data set.

For all reference data sets, many misclassifications are at parcel borders with mixed pixels in the satellite imagery. Largest misclassifications occur for waterbodies (minimum threshold for annual SAR VV mean is too low), bare soil, and artificial surfaces which also feature low mean backscatter and little variance over time. These areas can however easily be removed with optical data (e.g. all features are characterized by very low NDVI values).

TEST FOR CONFUSIONS

For better understanding, the confusion between grassland and other classes, the classification result of SAR2017 is compared with the VIRP plots 2017. Therefore, those plots were evaluated which are classified as grassland in SAR2017 and not grassland in VIP2017 resulting in 166 overall wrongly classified samples (see Table 3-33).

Table 3-33. SAR threshold based grassland classification confusions.

Reference class definition	with the percentage of total in the class	
Cropland	55	of total 1189 = 4,6%
Forests and Trees	41	of total 945 = 4,3%
Shrubs	4	of total 78 = 5,1%
Artificial Surfaces & Associated Area(s)	32	of total 382 = 8,4%
Bare Area(s)	6	of total 29 = 20,7%
Waterbodies, Snow and Ice	27	of total 49 = 55,1%
Wetlands	1	of total 1 = 100%

Largest misclassifications occur for waterbodies (minimum threshold for annual SAR VV mean is too low), bare soil, and artificial surfaces which also feature low mean backscatter and little variance over time. These areas can however easily be removed with optical data (e.g. all features are characterized by very low NDVI values).

3.2.3.3.3 Random forest based grassland classification with SAR Data

The feature importance for the in Table 3-24 mentioned annual SAR features including both polarizations (VV, VH) are estimated with the Mean Decrease Impurity measure (also known as Gini importance). The feature importance is estimated for a *grassland/non grassland* separation. Earlier tests differentiating between 8 land cover classes have shown that the feature importance for the separation of all 8 classes is not significantly lower or higher if using one or more features. Figure 3-33 shows that both polarizations have more than one feature that is important for the grassland separation. Using the VH polarization the Standard Deviation and the Coefficient of Variation ranked highly, whereas using the VV polarization the Mean and the Standard deviation show the highest ranking. Keeping in mind the drawback of the Mean Decrease Impurity measure, reducing the importance of following correlating features, both polarizations are correlating with each other and, therefore, the VV polarization shows a lower feature importance. Nevertheless, the combination of both polarizations using the Mean VV feature and the Standard deviation VH feature is tested further on.

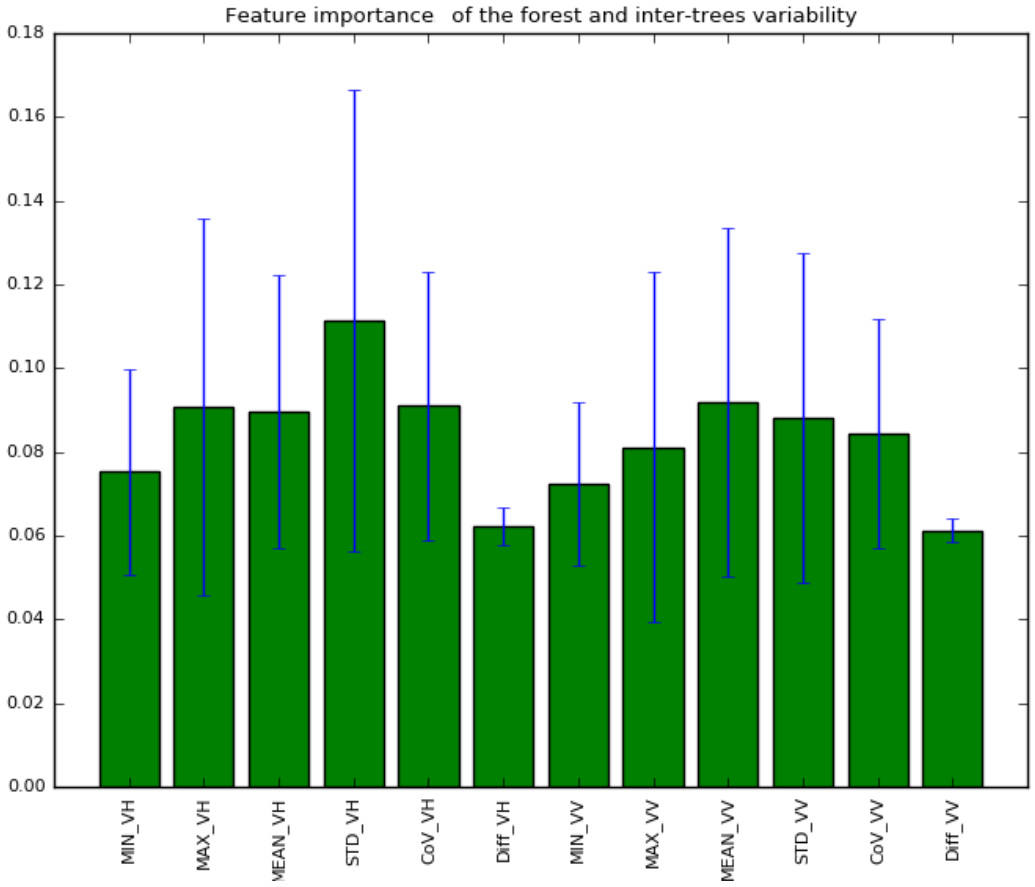


Figure 3-33. Feature Importance for the annual SAR features in both polarisations (VV, VH).

Using the results of the feature analysis a classification map is produced using the aggregated classes grassland and others with the following annual features: STD_VH, MEAN_VV, CoV_VH, MAX_VH, MEAN_VH and STD_VV. Regarding the feature importance, it should be noted that the first analysed feature shows a higher importance than other correlating features although they have the same importance. Therefore, it seems that the VV polarisation is less important, although it can be assumed that they have a similar importance.

Figure 3-34 and Figure 3-35 show the first results of the grassland classification based on selected annual SAR features (STD_VH, MEAN_VV, Cov_VH, MAX_VH, MEAN_VH, STD_VV). The random forest classification results confirm the conclusion derived from the threshold-based classification results based on the SAR data sets.

Table 3-34. SAR 2017 - thematic accuracy with different probability thresholds.

	SAR 2017 p>60%	SAR 2017 p>50%	SAR 2017 p>40%
Producer Accuracy	47.41	59.05	68.53
User Accuracy	78.01	75.69	67.37
Overall Accuracy	86.30	87.56	86.57

Different threshold are applied on the grassland probability maps to derive grassland/non grassland masks. Those masks are statistically evaluated as shown in Table 3-34. The results show that with higher probability thresholds the producer accuracy decreases whereas the user accuracy increases. The overall accuracy does not change significantly.

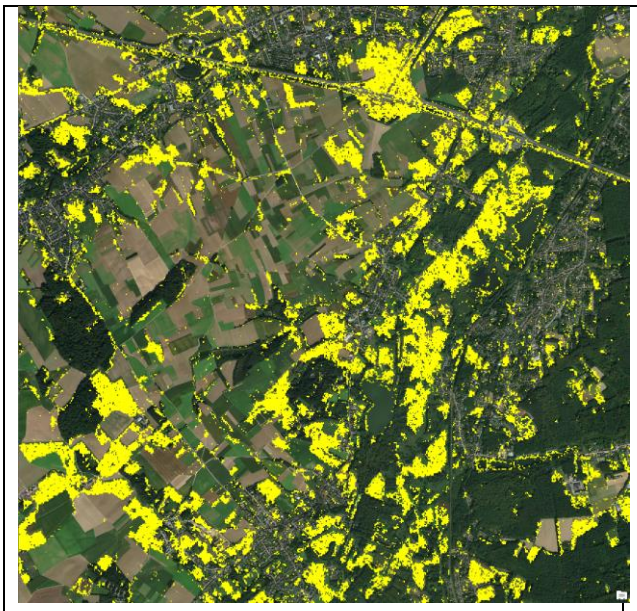


Figure 3-34. SAR grassland classification with random forest and selected features for 2017 (p>50%). (grassland in yellow)

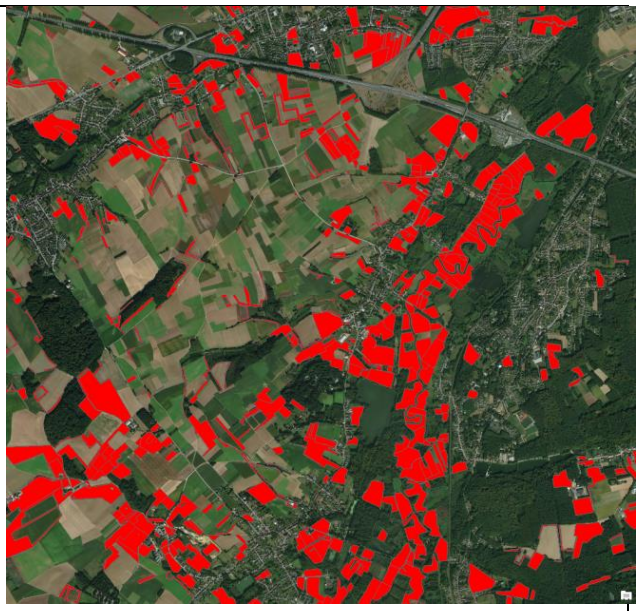


Figure 3-35. LGP grassland areas in red. Basis layer: ArcGIS Basemap.

Table 3-35. Error matrix grassland mapping with SAR2017 vs VIRP2017.

		Classification			
		Grassland	Others	Total	UA [%]
Ground Truth	Grassland	137	44	181	75.69
	Others	95	841	936	89.58
	Total	232	885	1117	
	PA [%]	59.05	95.03		

Overall Accuracy [%]	87.56
Kappa	0.59

The SAR classification result reaches a producer accuracy of 56% (see Table 3-35) tending to less homogeneous patches due to the speckle noise in the SAR data and misclassifications are detected at streets and other roads. Nevertheless, there is only a small confusion between grasslands and agricultural fields. The producer accuracy might be higher if the classification results are aggregated according to the MMU and whole within patches are closed.

3.2.3.3.4 Random forest based grassland classification with optical Data

The feature importance for the in Table 3-25 mentioned annual reflectance features is estimated with the Mean Decrease Impurity measure (also known as Gini importance). Figure 3-36 presents the feature importance of the listed features and shows that the importance slightly varies between the reflectance features. With regard to the feature importance calculation, it should be noted that the first analysed feature shows a higher importance than other correlating features although they have the same importance. Figure 3-36 show that the green and the red channel are more important than the blue. The first red edge band shows a higher importance, but the channel is correlating with other red edge and NIR

channels and therefore their importance can be assumed as equal. Furthermore, the SWIR1 channel shows a higher importance than the SWIR1 band.

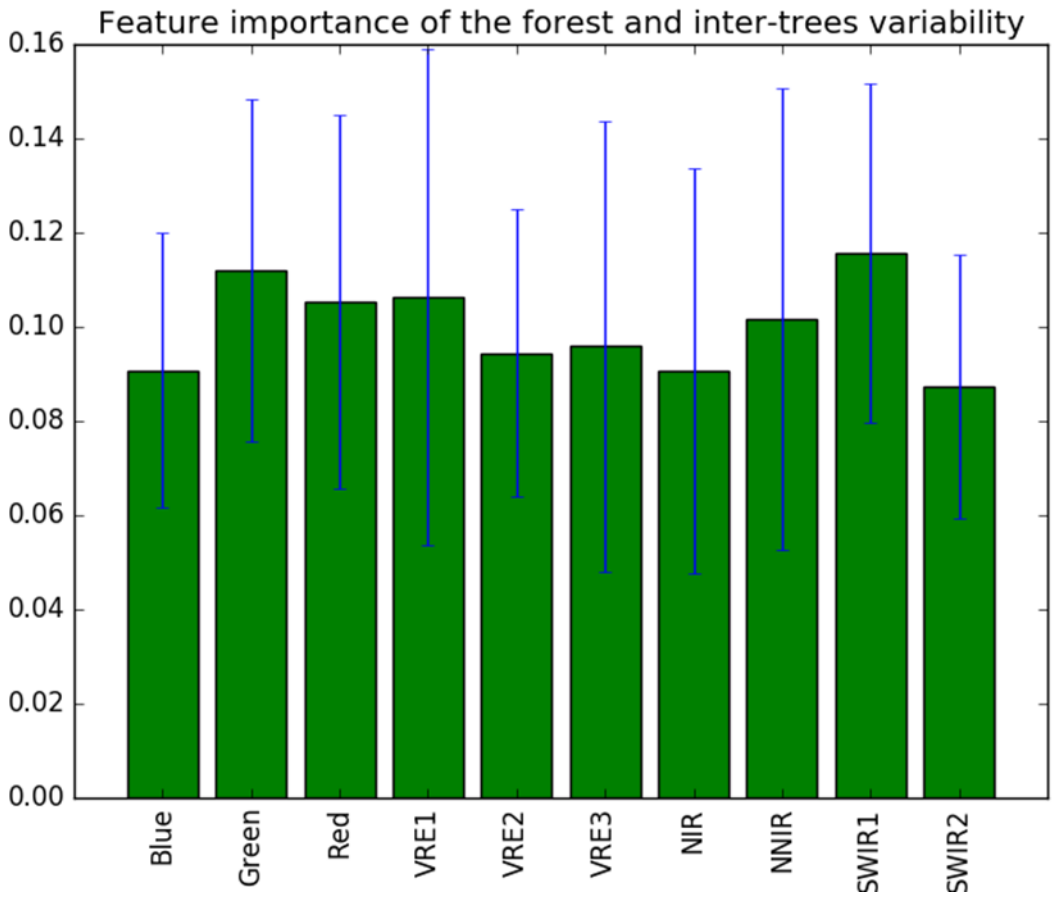


Figure 3-36. Feature importance of Sentinel-2 optical data.

Using the results of the feature selection analysis a classification map is produced using the aggregated classes grassland and others with the following features: Median over the vegetation period (March till October) including Green, Red, VRE1, VRE3, SWIR1 and SWIR2 reflectance.

Table 3-36. OPT 2017 - thematic accuracy with different probability thresholds.

	OPT 2017 p>60%	OPT 2017 p>50%	OPT 2017 p>40%
Producer Accuracy	51.72	62.93	73.31
User Accuracy	79.47	74.49	66.80
Overall Accuracy	87.20	87.82	86.93

The selected features serve as input for the random forest classifier to derive a grassland probability map. Different thresholds are applied on the grassland probability maps to derive grassland/non grassland masks. Those masks are statistically evaluated using the reference plots described in chapter 3.2.3.2. The results are presented in Table 3-36 showing that with higher probability thresholds the producer accuracy decreases whereas the user accuracy increases. As already shown with the SAR classification the overall accuracy does not change significantly.

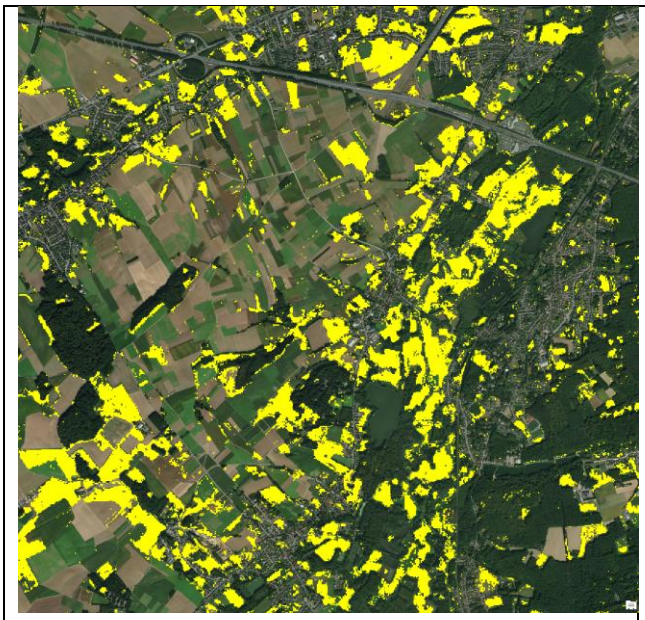


Figure 3-37. Optical grassland classification with random forest and selected features for 2017 (p>50%). (grassland in yellow)

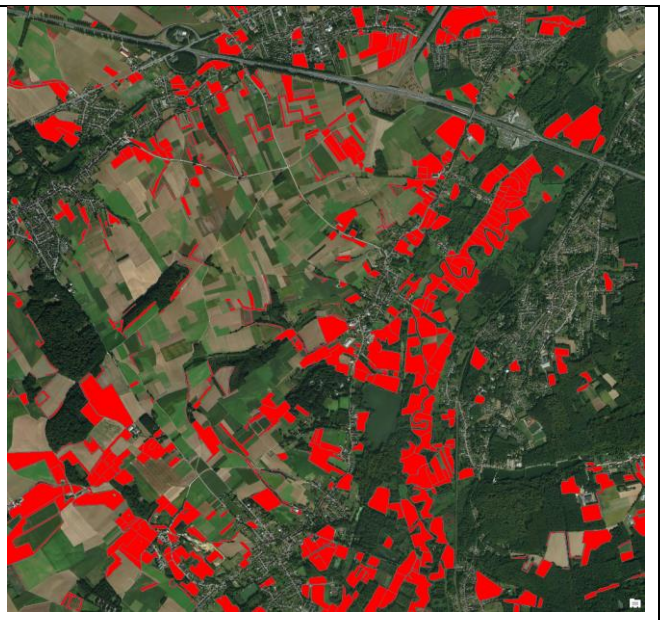


Figure 3-38. LGP grassland areas in red. Basis layer: ArcGIS Basemap.

As expected there is still confusion of grasslands with cropland areas which have high vegetation cover over the year. Compared to the SAR classification the grassland patches are more homogenous and show fewer gaps.

Table 3-37. Error matrix grassland mapping with OPT2017 vs VIRP2017

		Classification			
		Grassland	Others	Total	UA [%]
Ground Truth	Grassland	146	50	196	74.49
	Others	86	835	921	90.66
	Total	232	885	1117	
	PA [%]	62.93	94.35		

Overall Accuracy [%]	87.82
Kappa	0.61

Compared to the SAR classification the producer accuracy increased whereas the user accuracy decreased (see Table 3-37 and Table 3-35). Which leads to the conclusion that a combination of SAR and optical should improve the result.

3.2.3.3.5 Random forest based grassland classification with combined Data

The combined data set includes 12 features, 6 SAR features and 6 optical features. Using the results of the feature analysis following annual SAR features are used: STD_VH, MEAN_VV, CoV_VH, MAX_VH, MEAN_VH and STD_VV. Furthermore, following optical features are included in the combined stack: Median over the vegetation period (March till October) including Green, Red, VRE1, VRE2, SWIR1 and SWIR2 reflectance.

Table 3-38. OPT/SAR 2017 - thematic accuracy with different probability thresholds.

	SAR/OPT 2017 >60%	SAR/OPT 2017 >50%	SAR/OPT 2017 >40%
Producer Accuracy	59.48	69.40	78.02
User Accuracy	83.13	76.67	71.83
Overall Accuracy	89.08	89.26	89.08

Different thresholds are applied on the grassland probability maps to derive grassland/non grassland masks. Those masks are statistically evaluated using the reference plots described in chapter 3.2.3.2. The results are presented in Table 3-36 showing that with higher probability thresholds the producer accuracy decreases whereas the user accuracy increases. The results show that with higher probability thresholds the producer accuracy decreases whereas the user accuracy increases. The overall accuracy does not change significantly. As already shown with the optical and SAR classification the overall accuracy does not change significantly.

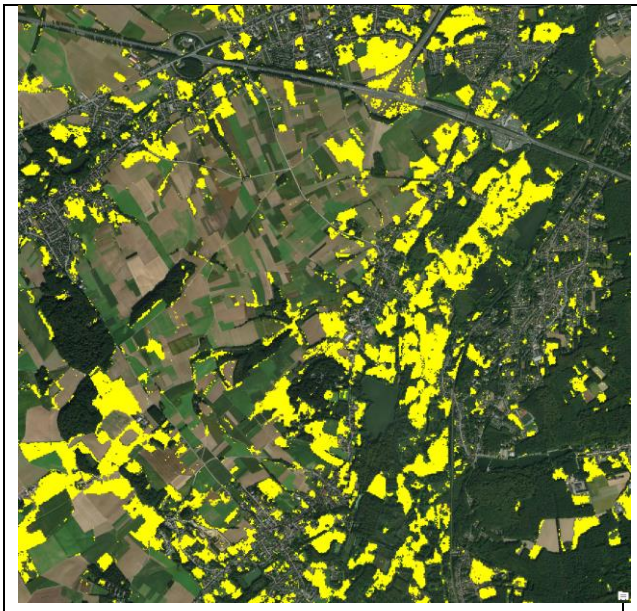


Figure 3-39. SAR + OPT grassland classification with random forest and selected features for 2017 (p>50%). (grassland in yellow)

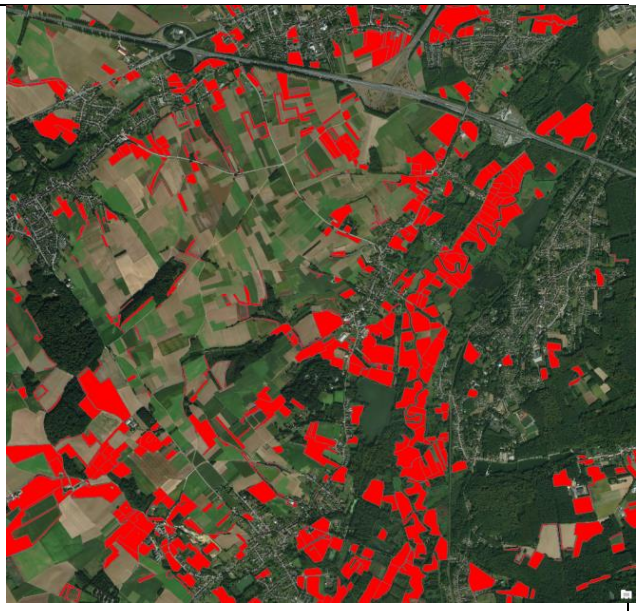


Figure 3-40. LGP grassland areas in red. Basis layer: ArcGIS Basemap.

Table 3-39. Error matrix grassland mapping with SAR/OPT 2017 p>50% vs VIRP2017.

		Classification			
		Grassland	Others	Total	UA [%]
Ground Truth	Grassland	161	49	210	76.67
	Others	71	836	907	92.17
	Total	232	885	1117	
	PA [%]	69.40	94.46		

Overall Accuracy [%]	89.26
Kappa	0.66

The combination of optical and SAR data showed the slightly better results with a producer accuracy of 69.40% (p>50%) / 78.02% (p<40%) and a user accuracy of 76.67% (p>50%) / 71.83% (p<40%). The classification with combined datasets reduces SAR specific misclassification with roads and optical specific misclassifications with cropland (see Figure 3-37 and Figure 3-40).

Further post-processing steps are applied to improve the classification accuracy. Patches smaller than the MMU should be excluded and wholes within patches should be filled.

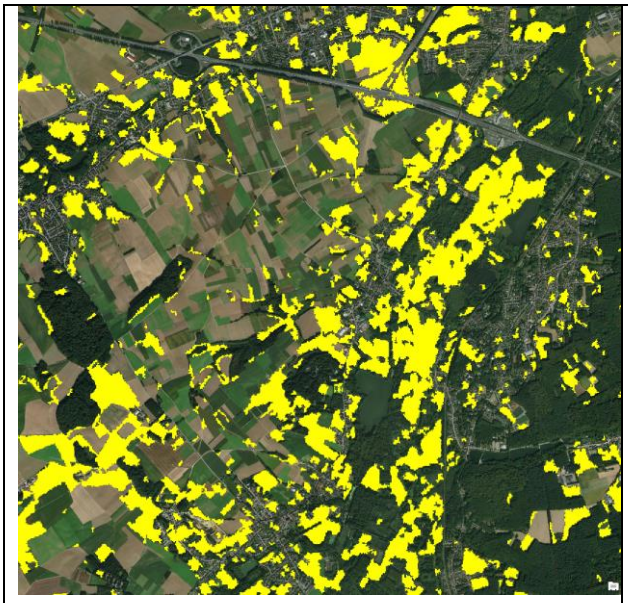


Figure 3-41. Aggregated (MMU 0.09ha) SAR + optical grassland classification with random forest and selected features for 2016. (grassland in green)

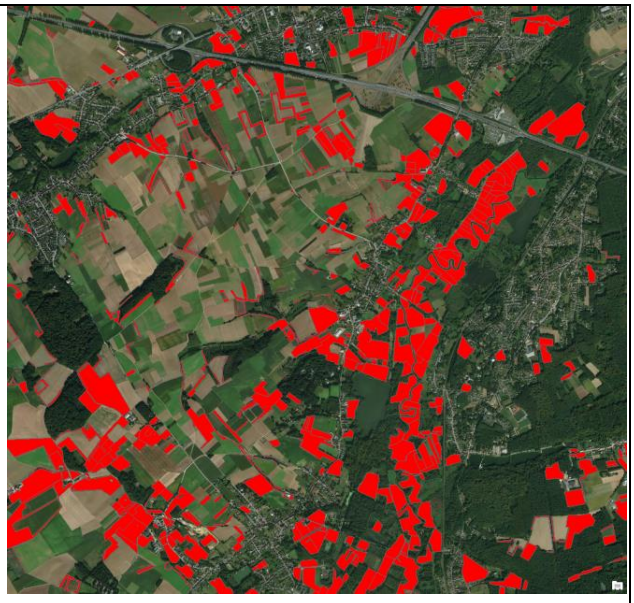


Figure 3-42. LGP grassland areas (2016) over Basemap VHR data.

Table 3-40. Error matrix grassland mapping with SAR-OPT2017 aggregated (MMU 0.09) vs VIRP2017.

		Classification			
		Grassland	Others	Total	UA [%]
Ground Truth	Grassland	183	60	243	75.31
	Others	49	825	874	94.39
	Total	232	885	1117	
	PA [%]	78.88	93.22		

Overall Accuracy [%]	90.24
Kappa	0.71

Further aggregation tests with an MMU of 0.25ha resulted in higher commission errors, due to the exclusion of smaller grassland areas. Those grassland features are located mainly in urban areas like parks and urban green spaces in residential and industrial areas. The aggregation with a MMU of 0.09ha showed the improved the thematic accuracy (see Table 3-40) by removing small grassland patches and single pixels and filling smaller gaps within grassland patches (see Figure 3-41 and Figure 3-42).

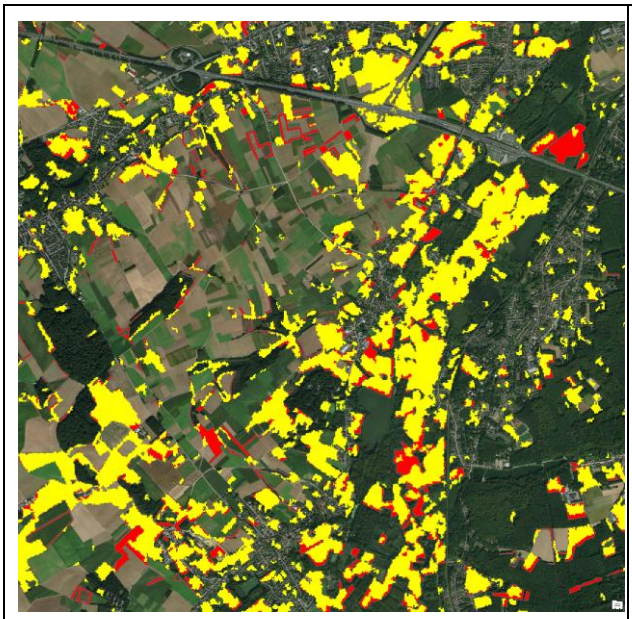


Figure 3-43. SAR + OPT classification 2017 aggregated: Omission errors (red). Classification (yellow) vs. LGP polygons (red).

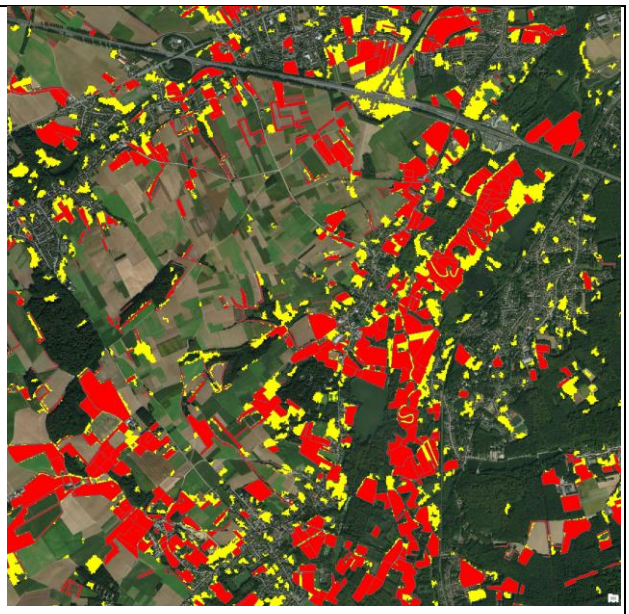


Figure 3-44. SAR + OPT classification 2017 aggregated: Commission errors (in yellow). Classification (yellow) vs. LGP polygons (red).

As Figure 3-43 presents, the omission errors include grassland patches with trees small grassland patches around agricultural fields and pasture which show a low grass cover.

As interpreting Figure 3-44 which presents the commission errors the difference between the grassland definitions should be kept in mind. In other words, not all green features are commission errors since the grassland definition from the LGP polygons does not include grasslands apart from agricultural areas. The actual commission errors are quite low. Some agricultural field are mistaken for grassland if the vegetation cover is high over the whole year.

Table 3-41. Thematic accuracy comparison of different features.

	SAR 2017 p>50%	OPT 2017 p>50%	SAR/OPT 2017 p>50%	SAR/OPT 2017 p>50% aggregated
Producer Accuracy	59.05	62.93	69.40	78.88
User Accuracy	75.69	74.49	76.67	75.31
Overall Accuracy	87.56	87.82	89.26	90.24

The classification result with SAR and OPTICAL combined datasets are quite encouraging with a producer and user accuracy of 79% and 75% (see Table 3-41). After a visual interpretation of all classifications, it can be observed that using optical data more confusion between grasslands and cropland are present, whereas using SAR data only more misclassification between grassland and roads are present. The combined approach shows more homogenous patches than using SAR data only. Another approach which will be tested in the task 4 will be the combination of SAR features with vegetation indices derived from the optical data set.

3.2.3.4 Summary and conclusions

The SAR2016 threshold based grassland classification is less accurate compared to the random forest approach, but it shows the potential of SAR data for the grassland classification. Due to fewer data sets in the growing season, the SAR2017 threshold based grassland classification shows better classification results than for the year 2016. This shows that the SAR threshold based grassland classification highly

depends on dense time series. Furthermore, the used thresholds were derived based on 2017 data sets and transferred to 2016 dataset without adjustment.

For all reference data sets, many misclassifications are at parcel borders with mixed pixels in the satellite imagery. Largest misclassifications occur for waterbodies (minimum threshold for annual SAR VV mean is too low), bare soil, and artificial surfaces which also feature low mean backscatter and little variance over time. These areas can however easily be removed with optical data (e.g. all features are characterized by very low NDVI values).

The aggregated classification result with SAR and OPTICAL combined datasets are quite encouraging with an overall accuracy of 90,24%, a producer accuracy of 78.88% for the grassland category, and a user accuracy of 75.31% for the grassland category. After the visual interpretation of all classifications, it can be observed that using optical data only more confusion between grasslands and cropland are present, whereas using SAR data only more misclassification between grassland and roads are present. The combined approach shows more homogenous patches than using SAR data only. Another approach which will be tested in the task 4 of the project will be the combination of SAR features with vegetation indices derived from the optical data set. Based on the applied tests and the experience of consortium partners from other projects, we recommend to apply the supervised random forest based approach within the demonstration sites in task 4, however, also other classification approaches such as support vector machine, can be applied. A main requirement however is the precise pre-processing of the dense time series including a topographic normalisation for hilly to mountainous terrain. For SAR time-series we recommend to apply multi-temporal filtering on gamma naught corrected imagery.

It can be stated that the results achieved in the first project phase are encouraging, but there is still potential to improve the accuracy which will be assessed in task 4 of the project. The general approach of applying supervised classification will thereby be followed.

Further research is specifically required to determine the optimal combination of features and indices derived from the optical as well as SAR dense time series. In phase 2 of the project the optimal combination of features and indices from optical as well as SAR imagery will be performed at the biogeographical region level, as for example different combinations will be optimal e.g. for the Mediterranean, the alpine and boreal ecosystem region.

After further tests the developed processing line on grassland identification will be implemented on larger demonstration sites in task 4 together with an approach for yearly incremental updates.

3.2.4 Agriculture

The following subchapters comprise the testing and benchmarking of the time series classification methods for Agriculture, in the Central test site (Germany) and the Belgium site.

3.2.4.1 Central test site – Germany

The potential of time series analysis for crop mask extraction and crop type monitoring via automated, supervised classification was examined in a variety of data-scenarios in the ECoLaSS central test site located in Baden-Wuerttemberg, Germany. The following chapters describe and discuss the results that were achieved with a selection of reasonable data configurations.

The central goal of this method testing is the generation of a potential future pan-European HRL on Agriculture, for which the specifications (e.g. variables, crop types, time intervals) are not yet defined (see AD05), and are up to the European Entrusted Entities (EEEs), the European Environment Agency (EEA) and the Joint Research Center (JRC).

There are ongoing efforts towards a Sentinel-based "Monitoring" approach (JRC, 2016) as part of the subsidies control in the framework of the Common Agricultural Policy (CAP) of the European Union. Supporting the control is an important potential application requiring spatial crop type information. In such an operational monitoring application it is not sufficient to deliver a crop classification at the end of the crop growing cycle. Instead, intermediate classifications have to be available during the season, with iterative updates improving the results throughout the year. Then, the crop type map can potentially increase the efficiency of the subsidy controls, where the reported crop types of the farmers are verified by on-site inspections. Copernicus core services, such as an Agricultural Service, could bring added value and be integrated into downstream services, such as CAP.

On the one hand, ECoLaSS aims at deriving methods for a potential future HRL Agriculture, and on the other hand, a Copernicus HRL on cropland could potentially provide information on a yearly basis which could be used as additional input to CAP, if desired.

3.2.4.1.1 Description of candidate methods

As described in the first three paragraphs of Section 3.1, one of the most important components of large-area land cover classification are the predictors or (time) features. These can be derived from different time series, such as S1 or S2 time series data. Furthermore, the features from both sensors can be combined. From a cost/benefit perspective, benefit arises mainly from higher product qualities (i.e. a higher accuracy of the produced map) while the amount of required processing is a matter of expense. The main purpose of the investigations presented in this section (3.2.4) is to investigate the suitability of the different datasets (S1, S2, S1 & S2) and different feature sets.

Particularly, using temporal-spectral features of Sentinel-2 and Sentinel-1 data (as described in chapter 3.1.4.1), multiple input data periods and configurations (pixel/field based) are evaluated with respect to the classification accuracy (Overall Accuracy 'OA' and Kappa Coefficient 'K'). Particularly, the following research questions were analyzed for the Central Site:

- Q1: Which accuracies can be achieved for the crop mask and the crop types classification based on input data from S-1, S-2 and the combination of the two?
- Q2: Can the accuracies of the crop types classification be improved by aggregating the results on field basis?
- Q3: Can the number of features be significantly limited with respect to the full feature set, without a significant accuracy decrease?
- Q4: Can the accuracies of the crop types classification be improved by including data and features from the late season of the previous years, e.g. in case of winter crops.
- Q5: How well can the crops be classified during the growing period?
- Q6: Is it possible to provide comprehensive information that enables users to assess the reliability of a prediction?

These questions have important implications with respect to the suggested input dataset selection and workflow definitions. For example, if the combination of features of both sensors does not improve the accuracy significantly, it is obviously preferable not to pre-process both S1 and S2. Also, if the pixel results are similar to the field based results, then a pixel classification is sufficient and the additional processing cost of a segmentation can be saved. It is important to stress, that no segmentation of the image data has been performed so far. However, the crop type reference data was available on parcel level. Thus, it was possible to aggregate the pixel-based classification outcomes per parcel in order to derive one prediction per parcel. This aggregation has been performed by calculating the class-wise mean probabilities over all pixels of a parcel. The parcel's crop type prediction has been assigned according to maximum (aggregated) class probability. Of course, the outcome of this procedure cannot be compared directly to the outcome of a segment-based classification in the sense that the quality of the derived segments would be less optimal compared to the parcels in many cases. However, the results derived by the parcel-based aggregation can serve as a proxy to results that could be possible with a segmentation-based classification and are thus useful for focusing future research resources.

3.2.4.1.2 Benchmarking criteria

At first glance, the approach with the optimum cost-benefit ratio is preferable. Cost factors can be manual labor, data availability, processing load and other sensor and scenario specific data properties, advantages and problems. The trade-off between optimal accuracies and low cost is always application dependent. To give a comprehensive impression of the different experiments and possible outcomes, these criteria are reported as well.

3.2.4.1.3 Implementation and results of benchmarking

CLASSIFICATION INPUT DATA

The area of interest consists of two Sentinel-2 tiles (ECoLaSS Central test site, Baden-Wuerttemberg, Germany) out of the nine tiles of the demonstration site. Sentinel-2 and Sentinel-1 data from October 2016 to December 2017 were downloaded and pre-processed for the two Sentinel-2 tiles T32UNV and T32UNU. The coverage of the area of interest is shown in Figure 3-45, while the number of available scenes for each tile is shown in Table 3-42.

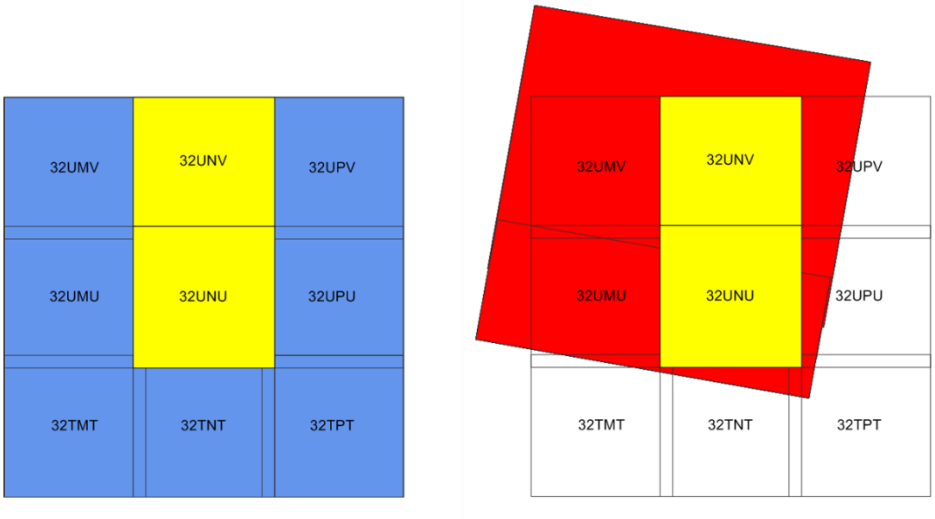


Figure 3-45. Sentinel-2 data coverage (left, blue) and Sentinel-1 coverage (right, red). The two test tiles are highlighted in yellow.

Table 3-42. Number of Sentinel-2 (< 50% Cloudcover) and Sentinel-1 scenes for the period October 2016 - December 2017.

	32UNU	32UNV
Sentinel-1	46	46
Sentinel-2	38	39

The Sentinel-2 imagery was atmospherically corrected and topographically normalized using the ESA Sen2Cor software. Only scenes with cloud cover < 50% were used for classification and analysis. The cloud cover metric does not rely on the official metadata cloud value provided by the original Sentinel-2A product, but is derived from the Scene Classification produced by Sen2Cor. Figure 3-46 shows the Sentinel-2 data score (inverted cloud value count) in the test site. The data score is calculated by counting the available cloud-free pixels and represents the number of available Sentinel-2 scenes with a cloud cover < 50%. The lower number of available scenes in the eastern part of the test site is caused by the product tiling of the Sentinel-2 data, leading to more available data for the western part of the test site.

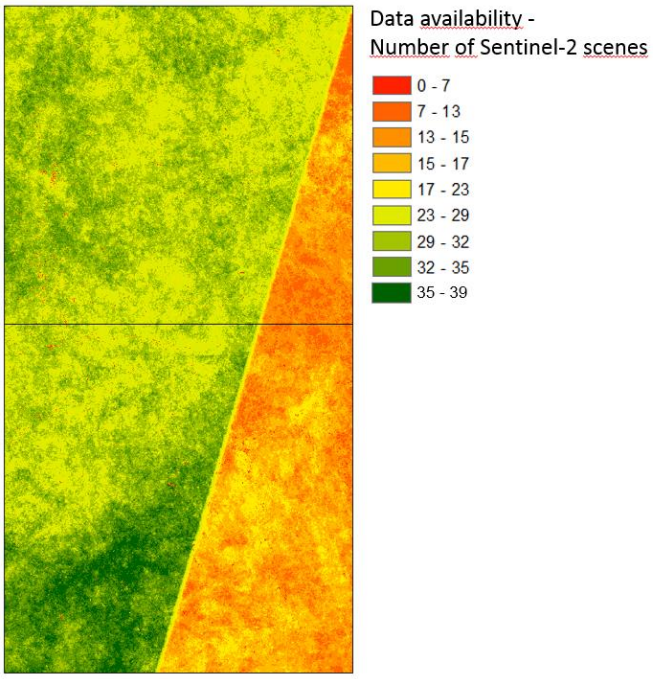


Figure 3-46. Sentinel-2 data score (inverted cloud value count) for ECoLaSS central test site (T32UNV/T32UNU tiles) for the time period Mar-Nov 2017.

The Sentinel-1 Ground Range Detected (GRD) data (VV and VH polarisation) were pre-processed to Gamma0 values and a multi-temporal filter was applied on the time series. The pre-processing was done using the ESA SNAP toolbox. Only data of the descending orbit 66 was used for the analysis.

Figure 3-47 describes the monthly data availability of Sentinel-2 and Sentinel-1 data for the two test tiles. The amount of scenes is calculated for the whole test site, meaning that data from the two Sentinel-2 tiles were included with a cloud cover of < 50%. Note that there is an intentional data gap between December 2016 and March 2017 as the growing season in central Europe begins at the earliest in March and winter crops are likely to be covered by snow during the winter months. Whereas this data interval is adjusted for central European conditions, further testing and prototyping in e.g. southern European regions might require regional adaptions for the selection of time periods. The low amount of Sentinel-1 scenes from 2016 is caused by the limited availability of Sentinel-1B data (in completion to Sentinel-1A). This also applies to the Sentinel-2 data, since Sentinel-2B data is only available starting from July 2017, meaning that there is a limited availability of Sentinle-2 scenes from October 2016 to June 2017.

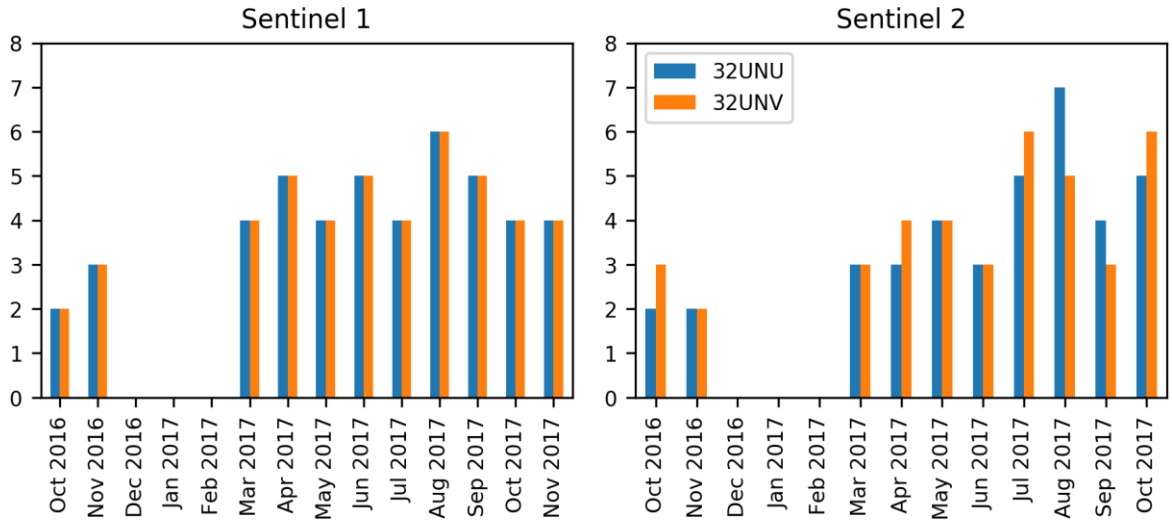


Figure 3-47. Monthly data availability for the two test tiles of Sentinel-1 (left) and Sentinel-2 (right) with cloud cover <50%.

TIME FEATURES

For the crop mask and crop type classification, all the time features described in section 3.1.4.1 were calculated for the full time period (from March to November 2017, referred to as 201703m9 which stands for the start year and month and the total number of months comprised by the time period). Additionally, the simple time features mean and median of consecutive two-month periods (March and April 2017, May and June 2017, July and August 2017 and September and October 2017, referred to as 201703m2, 201705m2, etc.) were calculated. An overview over the number of features created per period and sensor is given in Table 3-43 and some selected features are exemplarily shown in Figure 3-48. The benchmarking results of the research questions Q1, Q2 and Q3 are based on these features.

Table 3-43. Overview of the number of features for the different sensor and period combinations.

Sensor(s) Name	**Period**	S1	S2	S1&S2
201703m9	Mar-Nov 2017	60	63	123
201703m2	Mar-April 2017	8	8	16
201705m2	May-Jun 2017	8	8	16
201707m2	Jul-Aug 2017	8	8	16
201709m2	Sep-Oct 2017	8	8	16
SUM		92	95	187

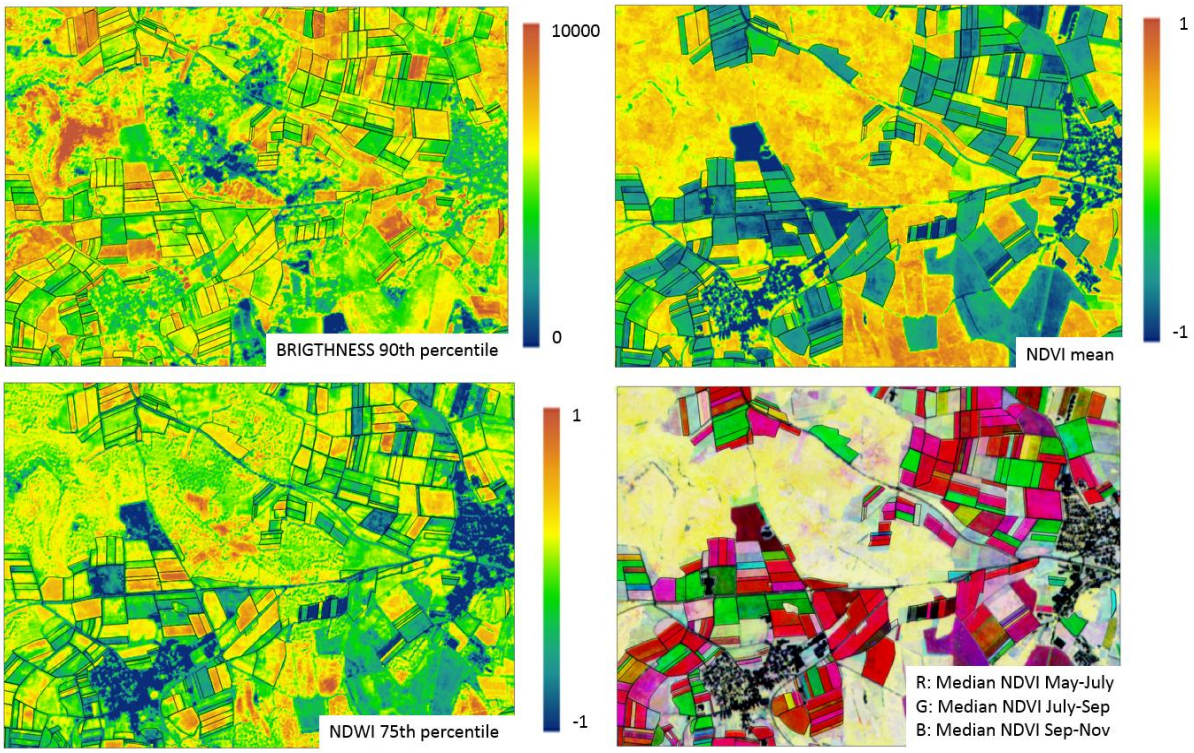


Figure 3-48. Exemplary selected time features from the Mar-Nov 2017 period (brightness 90th percentile, NDVI mean, NDWI 75th percentile) and an RGB composite of different two-month periods.

For Q4 it was analyzed if it is useful to include data and features from the end of the previous year in the classification for a possible better differentiation of the winter crops. Therefore, the full time features for the period October 2016 to November 2017 (however without data from December-February) and the mean and median for October and November 2016 were additionally computed. Additional features are derived from the two month period (8 for S1 and S2 each) and from the complete period Oct 2016 to Nov 2017 (92 and 95 for S1 and S2 respectively). The two classification scenarios thus differ by the inclusion of the 2016 data in the feature sets. Table 3-44 shows the number of features used in these two settings.

Table 3-44. Comparison of the number of features when excluding and including the October and November 2016 data.

Sensor(s) Dataset	S1	S2	S1&S2
2016 excluded	92	95	187
2016 included	160	166	326

Instead, for Q5 the features were extracted by limiting the data availability to two specific due dates: In the mid-June Scenario all images acquired between March 1, 2017 to June 19, 2017 were considered. The end date is the preliminary cross checks deadline of farmers geospacial aid applications (GSAA). In the mid-July scenario all images acquired between March 1, 2017 and July 15, 2017 were considered. The end date is the last day for the examination of crop diversification. Both dates are relevant for the subsidies control in the framework of the Common Agricultural Policy of the European Union (see section 3.2.4). For the two shorter periods the same features as in case of 201703m9 where calculated but for the respective time interval only. The two-month period features were not considered in case they included scenes acquired after the respective end date. However, for the first short period scenario the same features as

for 201705m2 were calculated but only with the data of the 6 weeks ranging from May 1, 2017 to the end date.

Table 3-45: Number of features available for specific time period data scenarios.

Sensor(s) Period	S1	S2	S1&S2
Mid-June	76	79	155
Mid-July	76	79	155
Full Period	92	95	187

Since the processing cost increases with the number of features it is desirable to reduce the number of features without sacrificing accuracy. For this purpose, a feature selection algorithm can be applied in the classification workflow as described in Section 3.1.4.2 (Feature Selection). According to this workflow a large set of features is generated for the training samples. Then a subset of most informative features can be selected for the final classification model. Finally, only the selected features need to be computed on the complete raster data. In the benchmarking a recursive feature selection (Guyon et al. 2002) was used for selecting a subset of features. This algorithm selects features by iteratively considering smaller and smaller sets of the most informative features. The feature ranking is based on the cross-validated accuracy derived only from the training data in order to keep the test data independent.

REFERENCE SAMPLES

In the test area, applications for EU subsidies have been submitted by local farmers for 123 different agricultural land use classes. By far, the largest proportion covers grassland which is not examined in the present analysis. For the study, only major crop types were identified and grouped into 13 categories. Table 3-46 lists these relevant crops by category and number of available reference parcels.

Table 3-46. Overview of the reference samples used for crop type classification.

Cropcode	Category	Abbreviation	# of Parcels
115	Winter Wheat (Containing Winter Bread Wheat, Winter Spelt, Einkorn/Emmer Grain)	W-Wheat	2882
116	Spring Bread Wheat	S-Wheat	29
121	Winter Rye	W-Rye	56
131	Winter Barley	W-Barley	1728
132	Spring Barley	S-Barley	1360
143	Spring Oat	S-Oat	475
156	Winter Triticale	W-Triticale	859
210	Peas	Peas	97
311	Winter Oilseed Rape	W-Rapeseed	729
411	Maize	Maize	2814
424	Agrarian Grassland (Containing Clover, Grass-Clover, Alfalfa-Grass- & Clover Mix, Alfalfa)	Agr-Grass	1442
590	Fallow	Fallow	266
602	Potatoes	Potatoes	268

While cereal production is dominated by winter crop types, there is also a large proportion of fields used for maize and agricultural grass cultivation (Figure 3-49, left). As for spring cereals, only barley shows a certain frequent occurrence. Mean parcel sizes range between 0.46 ha for potatoes which tend to be grown on rather small strips of land and 2.11 ha for winter rape fields (Figure 3-49, right).

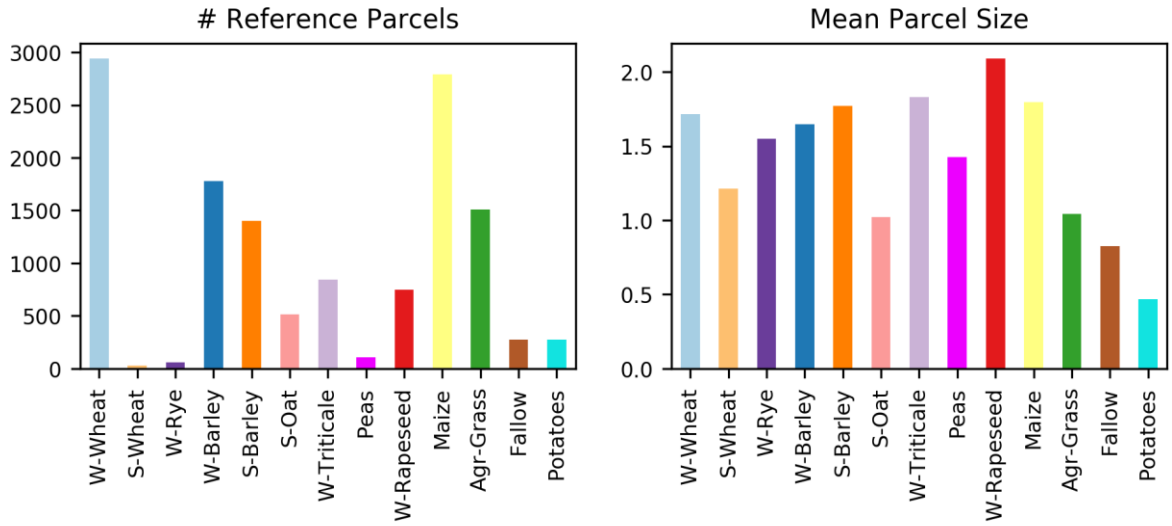


Figure 3-49. Frequency (left) and mean parcel size (right) of the reference samples used for crop type classification.

For the training and evaluation of the crop type classifier, 2000 pixels were chosen randomly for each of the predefined crop categories. The initial split of the data into training- and test sets was based on the input geometries (proportion = 7:3) to ensure unambiguous reference data.

For the crop mask classification, additional reference samples that cover the basic landcover types were available from the HRL 2015 layer production. They were re-used to complete the crop-sample base (described above) to support the differentiation between non-crop and crop. The class 'forest' consists of coniferous and broadleaf forest samples, the class 'grassland' includes data from the above mentioned reference parcels as well as grassland samples of the HRL 2015 grassland layer production. Table 3-47 shows the number of samples for each class. Splitting the dataset into a training- and test set was also performed for the crop mask classification, meaning that approx. 70% were used for training and 30% for testing.

Table 3-47. Overview of the reference samples used for the crop mask classification.

Class code	Classname	# of samples	Source
1	Forest	734	HRL 2015
2	Crops	13719	Farmer's Application
3	Grassland	5801	Farmer's Application, HRL Grassland 2015
4	Urban areas	300	HRL 2015
5	Waterbodies	163	HRL 2015

CLASSIFICATION AND RELIABILITY LAYERS

A Random Forest Classifier was used for all classifications due to its generally good performance and ease of use. Apart from the class predictions, the classifier also provides the output of (pseudo-) probabilities, i.e. the mean predicted class probabilities of the decision trees. From these probabilities it is possible to derive reliability information: three layers are calculated in addition to the class predictions and individual class probabilities:

- largest probability (maximum probability)
- largest probability - second largest probability (breaking ties) (Luo et al. 2015)
- entropy $-\sum_{c=[1, \#Classes]}(p_c \log p_c)$ Where the p_c is the probability of class c

The range of the first two layers is naturally between [0, 1], or, as in our case when multiplied by 100, [0, 100]. Of the three layers, the 'largest probability layer' is least significant, but can eventually be useful in specific analyses when combined with the other layers. The breaking ties layer is based on the two highest class probabilities: Two samples may have the same largest probability, e.g. 60, but a differing second best probability, e.g. 40 in one case and 5 in another case. As a result, the breaking ties reliability is 20 and 55 respectively. The entropy layer is another reliability measure which takes into account the probabilities of all classes. Originally, the potential range of the entropy depends on the number of classes. In order to align the value interpretations of the entropy according to the other two reliability layers, the values are rescaled to the range [0,100]. Low values correspond to unreliable and high values to reliable predictions.

3.2.4.1.4 Results of benchmarking

Q1 - CROP MASK: COMPARISON BETWEEN SENSORS AND PIXEL/FIELD-SAMPLE UNIT FOR THE REFERENCE YEAR 2017

The following results were generated using only data from reference year 2017. In the central site the classification accuracies based on S1 data are significantly lower than the respective accuracies based on S2 data. The combination of the two sensors does not significantly improve the classification accuracy when compared with the accuracies using simply the S2 data (Table 3-48 and Figure 3-50Figure 2-1). Since the reference data is available as polygons, it was possible to aggregate the results of the pixel classification on field level and perform an accuracy assessment on both pixel and field level, with fields being the sample unit. To do so, the mean class-probabilities per field were calculated. Then, the new class prediction and reliabilities were computed based on the aggregated probabilities. In case of all input data sets, but particularly for S1, the overall accuracies increase with the field size. This is an expected pattern due to, e.g., the presence of speckle in the SAR data. This important finding indicates that - even if the real field polygons are not available, segmentation should be considered in the case that optical data availability is not sufficient (due to high cloud cover) and SAR data must be used as primary data source.

Table 3-48. Kappa Coefficient (K) and Overall Accuracy (OA) for the different crop mask experiment setups (S1, S2, and S1&S2 on pixel and field level).

	K * 100	K * 100	OA	OA
	pixel	field	pixel	field
S1	68.1	78.8	89.5	89.3
S2	79.2	84.9	93.0	92.3
S1&S2	80.4	85.2	93.9	92.6

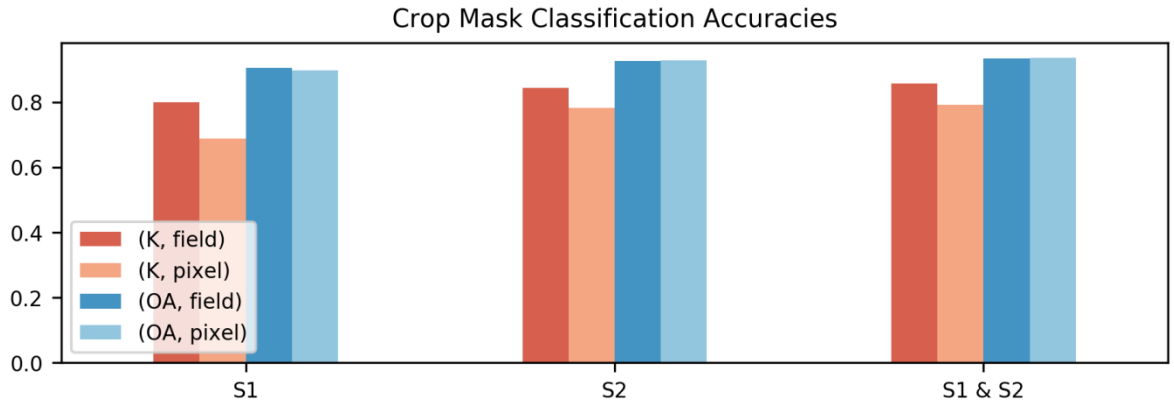


Figure 3-50. Barplot of Kappa (K) and Overall Accuracy (OA) for the different experiment setups (S1, S2, S1&S2 on field and pixel level).

The following Table 3-49 gives a summary of the results with regard to the processing cost, accuracy and other benchmarking criteria. The listed sensor and data specific problems can lead to misclassification effects or class confusion in the final classification raster. For example, the speckle noise of the Sentinel-1 data can lead to strong 'salt and pepper' effects, which would require a preceding, time consuming segmentation to avoid this effect. Other problems are the partially inconsistent cloud and cloud shadow masks of the Sentinel-2 L2A data that are not always able to capture all of the clouds and cloud shadows, which leads also to misclassification in the final raster.

Table 3-49. Benchmarking criteria and specific problems of the different experiment setups.

	Accuracy (K*100)	Processing Cost	Specific Problems
S1 pixel level	68.1	+	Foreshortening, layover in strong relief, speckle
S2 pixel level	79.2	+	Clouds/cloud shadows
S1&S2 pixel level	80.4	++	As in S1/S2 at pixel level; the strength of one sensor type can compensate for the weaknesses of the other and vice versa.
S1 field level	78.8	++	Foreshortening, layover in strong relief, segmentation
S2 field level	84.9	++	Clouds/cloud shadows, segmentation
S1&S2 field level	85.2	++++	As in S1/S2 at field level; the strength of one sensor type can compensate for the weaknesses of the other and vice versa. Segmentation

Q1 & Q2 - Crop Types: Comparison between Sensors and Pixel/Field-Sample Unit for the Reference Year 2017

The patterns of the crop type classification accuracies are similar to those of the crop mask (Table 3-50 and Figure 3-51). In case of the pixel classifications the S2 based classification yields better results than the S1 based classification (Kappa*100: + 18.2) while the combination of the two sensors does not improve the S2 based classification significantly (Kappa*100: + 2.6).
As expected, the field-based accuracies are all significantly higher compared to the corresponding pixel-based accuracies. For the S2 based classification, Kappa*100 increases from 74 to 77.5 and for the S1/S2 based classification there is an increase of 1.9. In case of the S1-based classification there is an improvement of Kappa*100 from 55.8 to 64.

Table 3-50. Kappa Coefficient (K) and Overall Accuracy (OA) for the different experiment setups (S1, S2, S1&S2 on pixel and field level).

	K * 100	K * 100	OA	OA
	pixel	field	pixel	field
S1	55.8	64.0	61.3	68.5
S2	74.0	77.5	77.8	80.6
S1&S2	76.6	78.5	80.0	81.4

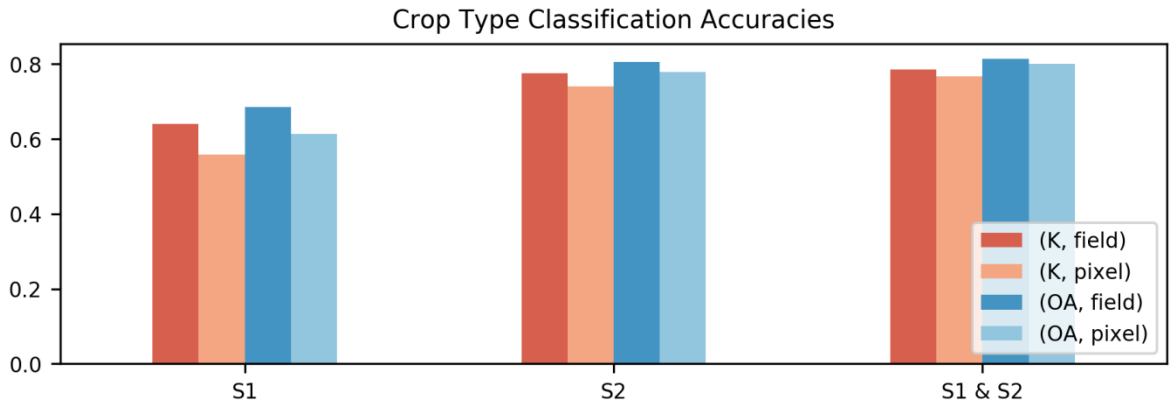


Figure 3-51. Barplot of Kappa (K) and Overall Accuracy (OA) for the different experiment setups.

gives again a short overview on the results with specific problems that can occur and with regard to the processing cost and accuracy.

Table 3-51. Benchmarking criteria and specific problems of the different experiment setups.

	Accuracy (Kappa*100)	Processing Cost	Specific Problems
S1 pixel level	55.8	+	Foreshortening, layover in strong relief, speckle
S2 pixel level	74.0	+	Clouds/cloud shadows
S1&S2 pixel level	76.6	++	As in S1/S2 at pixel level; the strength of one sensor type can compensate for the weaknesses of the other and vice versa.
S1 field level	64.0	++	Foreshortening, layover in strong relief, segmentation
S2 field level	77.5	++	Clouds/cloud shadows, segmentation
S1&S2 field level	78.5	++++	As in S1/S2 at pixel level; the strength of one sensor type can compensate for the weaknesses of the other and vice versa. Segmentation

The class-wise F1-Scores (mean of User's and Producer's Accuracy) depicted in Figure 3-52 shows that maize and winter rapeseed, which account for respectively 21 % and 6 % of the parcels, can be classified with very high accuracies. Winter rapeseed can be classified almost as good with S1 as with S2 and the field level aggregation does not improve the classification accuracy importantly. In case of maize, the accuracy of the S1 classification is also very high (with an F1 score of ca. 0.8) and the field based aggregation improves the classification importantly. Nevertheless, compared to S1, for maize significantly higher accuracies can be achieved with S2. In general, the accuracies of the cereals are not as high compared to rapeseed and maize. As expected, there is a higher confusion between cereal types belonging to the spring and winter group, respectively. Particularly, the confusion between winter wheat and winter triticale and between spring barley and spring oat is high. In general, some of the classes with a very small amount of fields present in the study site cannot be well separated, particularly not without the field level aggregation and/or sensor aggregation. This is particularly true for spring wheat (116), spring oat (143), peas (210), fallow (590) and potatoes (602). These classes are relatively rare in the study site as can be seen in the sample distributions of Table 3-46 (section 3.2.4.1.3, Reference Samples). However, if such smaller classes play an important role for a given application, then sensor combination and segmentation should be considered for improving their accuracies.

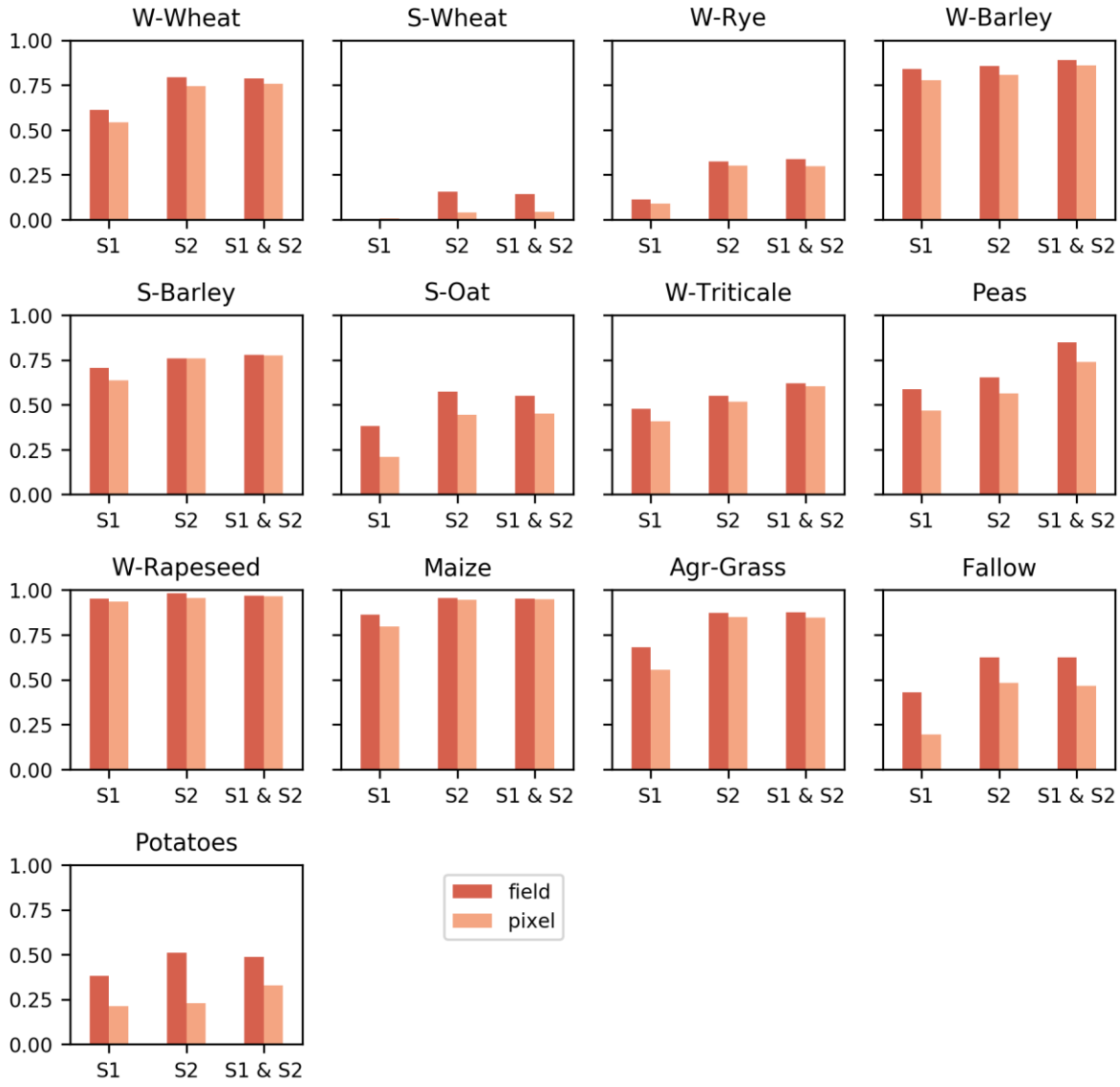


Figure 3-52. Class-wise F1-Score (mean of User's and Producer's Accuracy) for field vs. pixel-based classifications, reference year 2017.

The Figure 3-53 shows the confusion matrix of the crop type classification on field level based on the combination of Sentinel-1 and Sentinel-2 time features. The ability of the different time features to separate between the 13 different crop types depends strongly on the similarity of the classes. For example, of the 882 actual samples for 115 - winter wheat, ca. 29% were misclassified, mainly as 156 - winter triticale. Vice versa, of the 252 available samples for winter triticale, approx. 17% were falsely assigned to the class winter wheat. And of the 421 actual reference samples for 132 - spring barley, approx. 28% were misclassified, mainly as 143 - spring oat. This shows that it was not possible to differentiate these highly similar classes with the calculated time features. Further investigation and testing is necessary, e.g. one could aggregate similar classes (e.g. winter wheat and winter triticale) if the required application allows it. Or if the classes are too similar, different periods within the growing season should be investigated for differentiation, for example at what time the crop was sown or harvested.

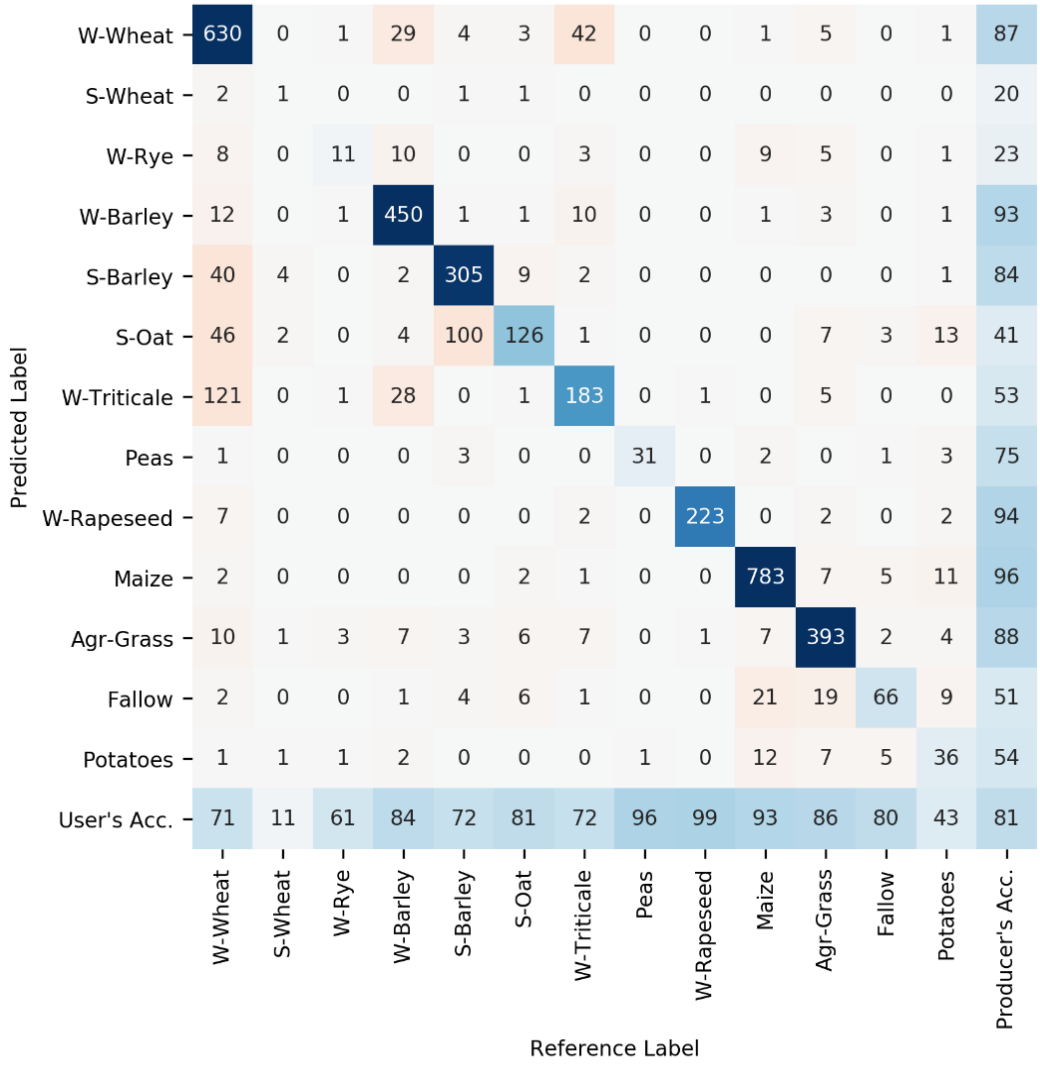


Figure 3-53. Confusion Matrix of the crop type classification on field level based on the combination of Sentinel-1 and Sentinel-2 time features.

Q3 - Crop Types: Including Data from the Previous Year 2016 in the Classification

Figure 3-54 shows that the classification accuracies improve slightly when data and features from 2016 (October and November) are included. It should be evaluated if the integration of the additional data (i.e., from 2016) is worth in terms of the trade-off between accuracy requirements on the one hand and costs with respect to data processing on the other hand. Due to the small improvement in accuracy, this work concludes that it might not be worth to include 2016 data. Therefore, the following experiments were conducted with 2017 data only.

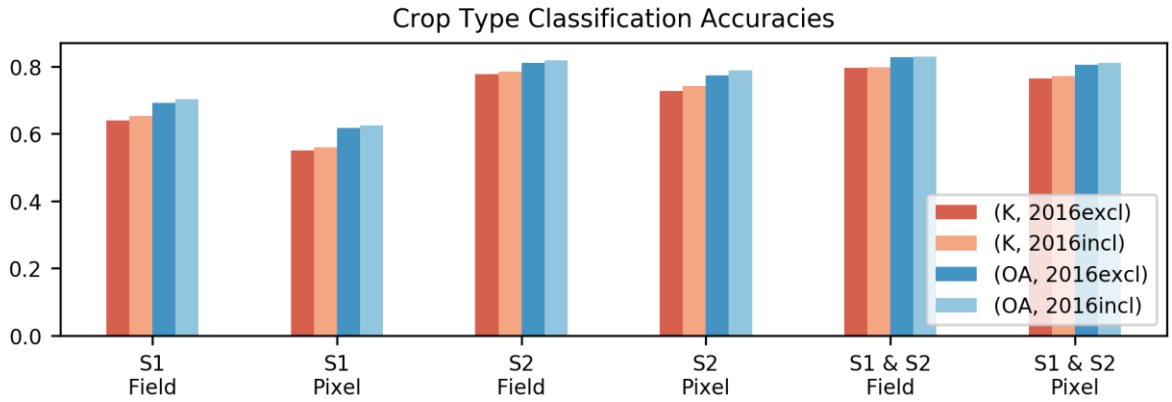


Figure 3-54. Barplot of Kappa (K) and Overall Accuracy (OA) for the different experiment setups.

Q4 - Crop Types: Can the number of Features be reduced without the Loss of Accuracy?

Given that there is no significant reduction in the accuracy, it is desirable to reduce the number of calculated features to reduce the processing cost. For the crop type classification the recursive feature elimination returned an optimal set of 50 features from the almost 187 considered features. As can be seen in the plot below, the cross-validation score saturates early and peaks at 50 features. Even though a close to the peak accuracy is already achieved earlier with ca. 25 features, the number of features at the absolute maximum was selected.

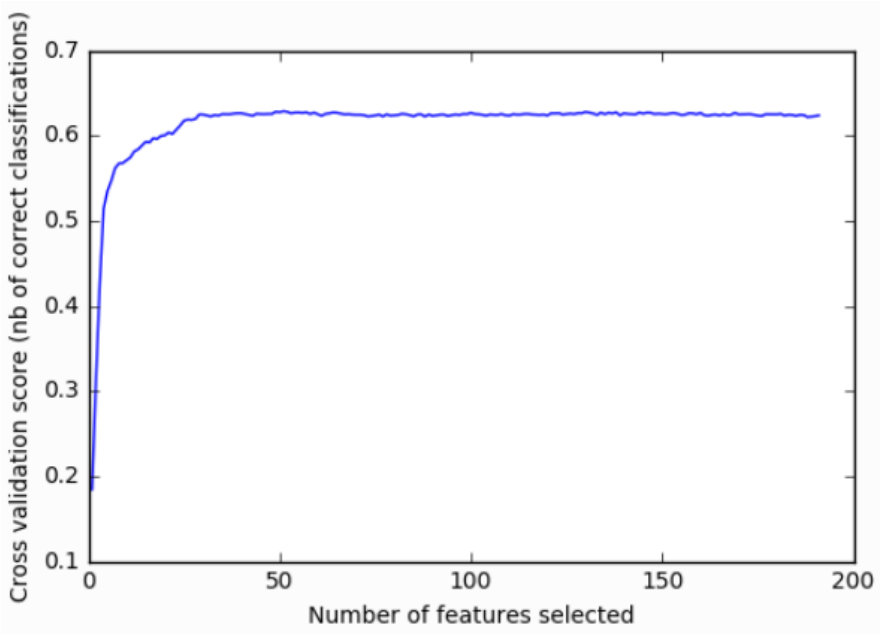


Figure 3-55. Overall accuracy (OA) based on the cross-validated training samples dependent on the number of selected features.

Based on the independent test data it can be confirmed that the accuracies with the selected feature subset are similar to the ones achieved with the full feature set (Figure 3-56). As a consequence, there is a high potential to reduce the processing cost without reducing the accuracy by, firstly, computing all spectro-temporal features only based on the training data. After performing a suitable feature selection on the training data set, only the selected and most relevant features would then have to be calculated for the whole raster footprint.

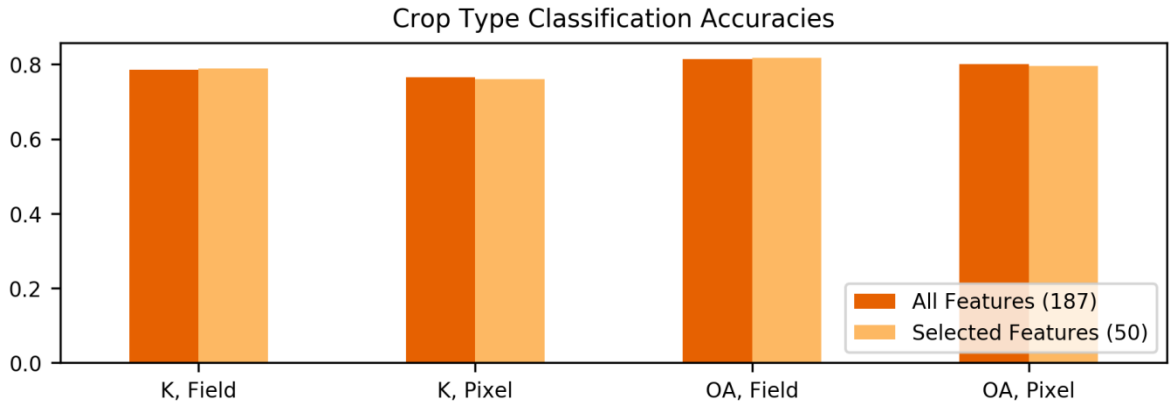


Figure 3-56. Barplot of Kappa (K) and Overall Accuracy (OA) for the classification based on all features, and the 50 selected features.

Q5 - Crop Types: Classification Performance During the Growing Cycle

In the mid-June scenario, where data is used from March 1, 2017 to June 19, 2017, the crop type classification accuracies are significantly lower than for the mid-July scenario, where data until July 19, 2017 was used. This is true for all the sensor scenarios: S1, S2 and S1&S2 (Figure 3-57). Instead, the improvement of the classification accuracies between the mid-July and the full period scenario is only marginal.

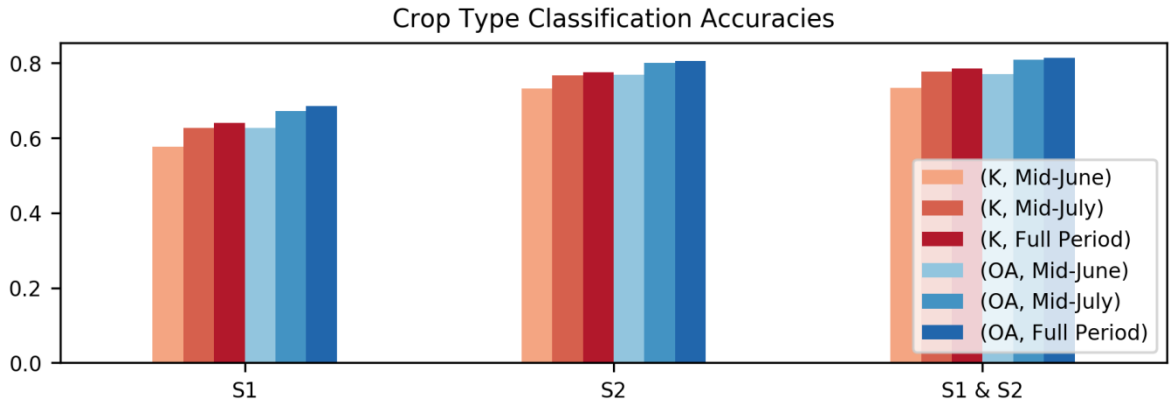


Figure 3-57. Kappa and Overall Accuracy on field level of the for the different experiment setups, particularly the three considered periods.

Looking at the class-wise classification accuracies, Figure 3-58 shows some expectable patterns. For example, rapeseed (311) blossoms before the end of the mid-June period. If the blossoms are captured in the data, this class is easily separable very early in the growing period. This is however only valid for optical data and in fact it can be seen that the accuracies for rapeseed are very high for all three periods in case S2 data is used. Instead, with S1 data the accuracies improve. Maize (411) improves when later data (e.g. from July and from the rest of the year) is included. This is also expectable since maize is sowed and harvested late with respect to the other crop types. Nevertheless it can already be classified relatively well in the mid-June period. The classes for agricultural grass (424), fallow (590) and potatoes (602) improve strongly especially when the data of the full period is used. This is particularly true for fallow which could

be the case because there is no harvesting event. The development of the different cereal accuracies is not clear. This is probably because there is a high confusion between these classes.

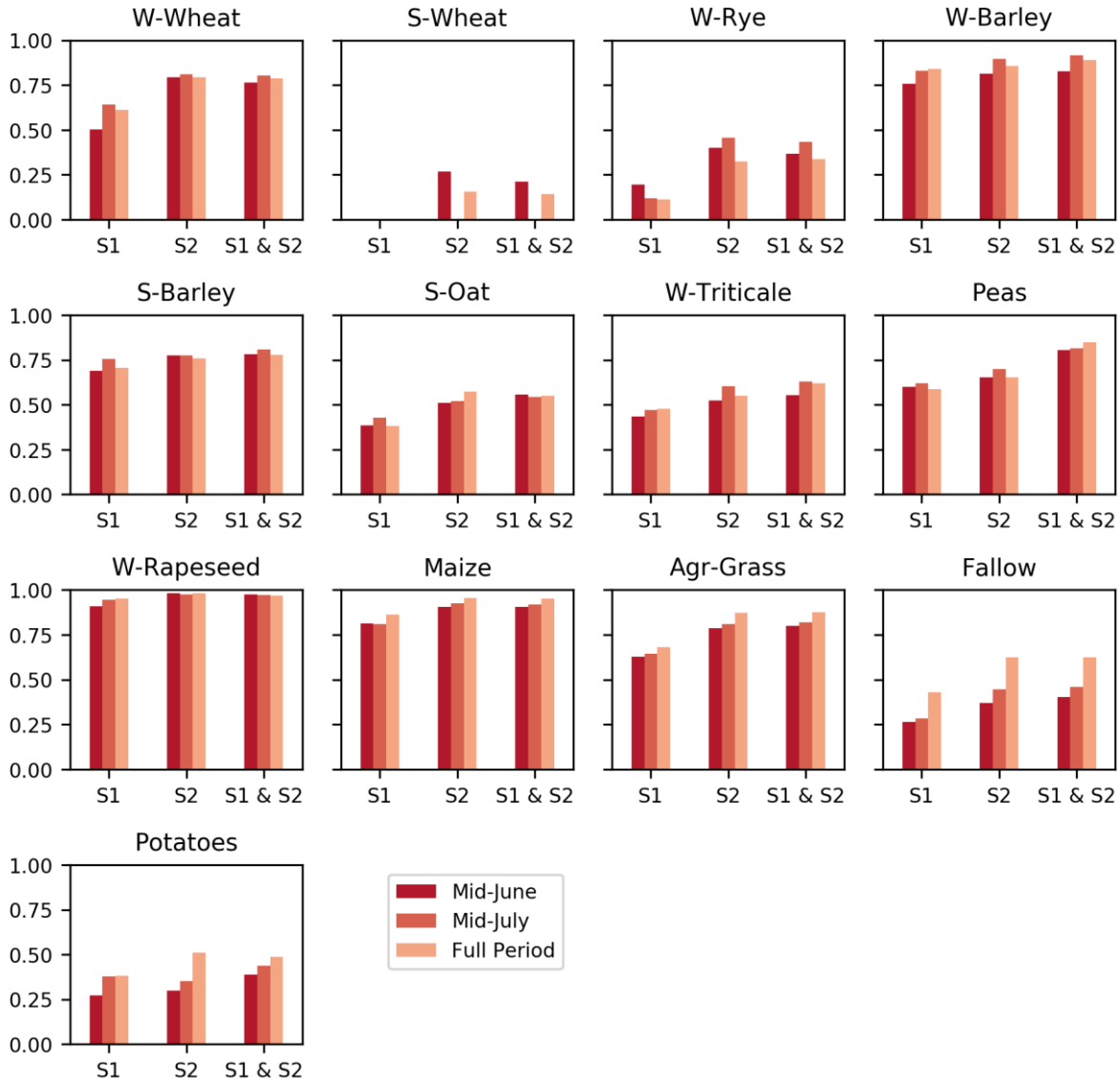


Figure 3-58. Class-wise field-level F1-Scores (mean of User's and Producer's Accuracy) for the different experiment setups, particularly the three considered periods.

Q6 - Crop Types: Reliability Layers

High classification reliabilities in the respective layers (3.4.2.3) usually correspond to correct predictions. This information can be used to prioritize the subsidy claims from farmers. Fields where the reported crop type label agrees with a high classification reliability can be directly approved. On the other hand, a high reliability for a specific class that does not agree with the reported crop type is an indicator for a likely incorrect claim and can then be investigated in more detail.

The validity of the reliability layers is crop dependent. Particularly for the well separable crop types, the reliability layers are informative with respect to the likelihood of a correct classification, e.g. for peas (210), rapeseed (311), and maize (411).

This can be observed in Figure 3-59 by the high separability of the reliability distributions, i.e. a high clustering of true predictions in the upper range of the reliability metric and of wrong predictions in the lower range of the reliability range. The more wrong predictions cluster in the lower reliability range, and correct predictions in the higher reliability range, the more informative is the reliability information since high reliabilities concur with high predictions. This is valid for both, the breakties and the entropy reliability. For crops, for which there is a high overlap between the two distributions (similar high values for correct and wrong predictions, the reliability is thus less informative.

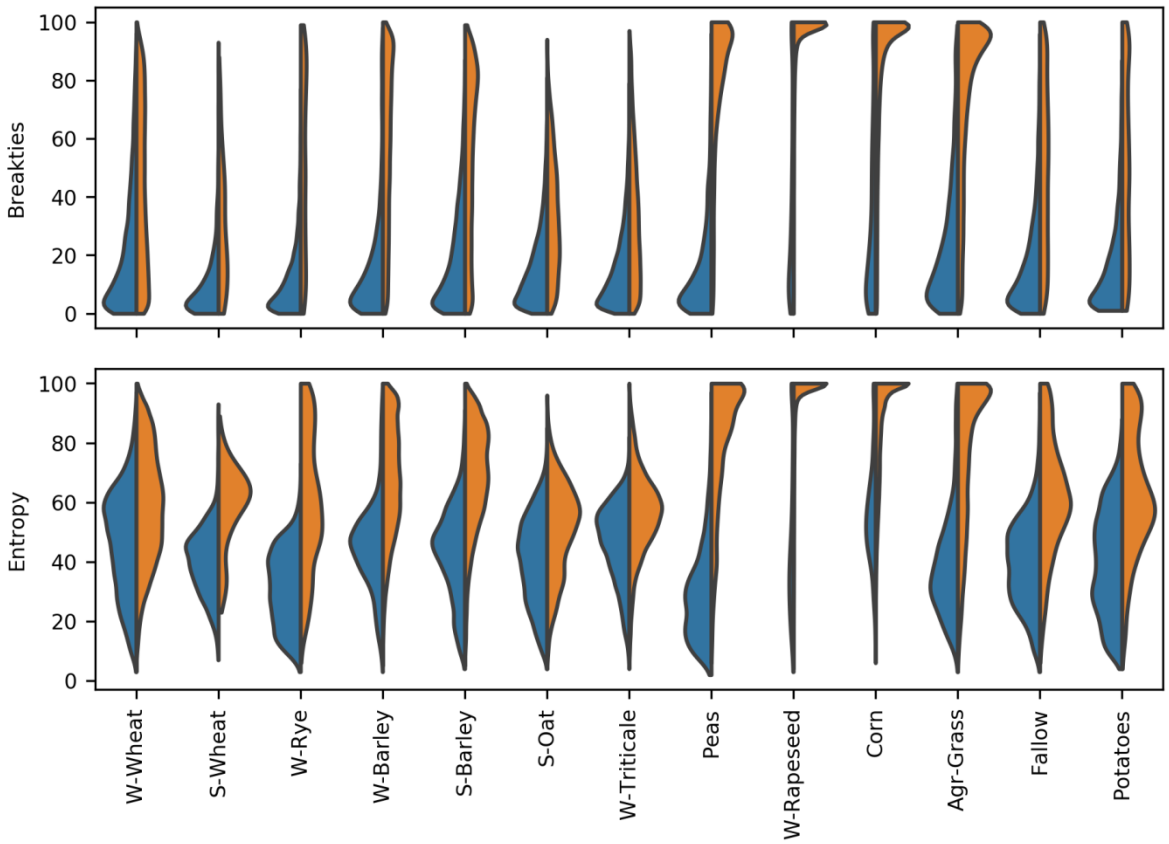


Figure 3-59. Distributions of the breaking ties and entropy reliabilities of wrong (blue, left) and correct (orange, right) predictions grouped by the predicted crop type.

ILLUSTRATIONS OF THE RESULTS

The following Figure 3-60 shows the final crop types map derived from the reference year 2017 data over the full test site. The four insets of an area close to Ulm (Merklingen) show (i) an RGB composite of the median NDVI layers of the three two-month periods as explained in section 3.2.4.3. (upper left), (ii) the crop mask (section 3.2.4.4.1) showing crop areas in white and masking out all other areas in black (upper right), (iii) the crop types with the all other areas masked out by the negative crop mask, and (iv) an RGB composite of the three reliability layers as explained in section 3.2.4.4.6: maximum probability, breaking ties and entropy.

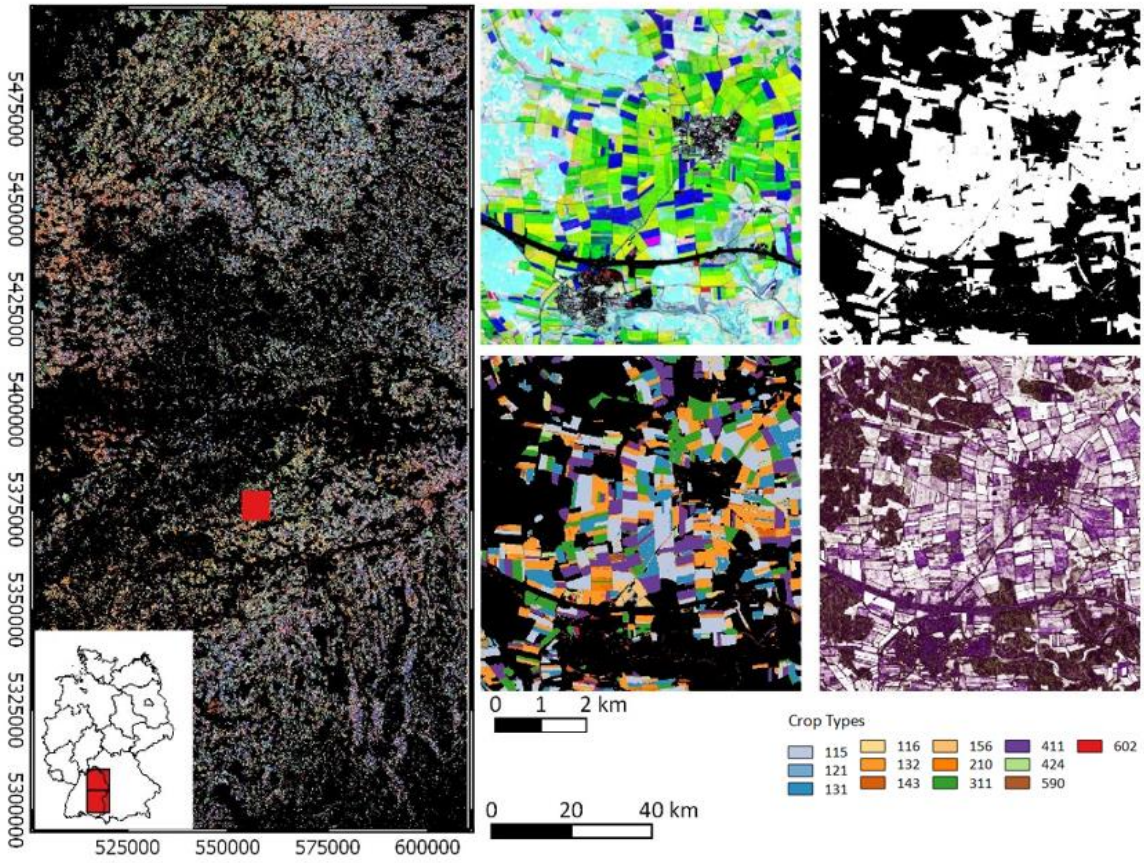


Figure 3-60. Final crop types map with the crop mask overlaid over the two processed Sentinel-2 tiles (left). The insets show an RGB composite of the median NDVI layers of the three two-month periods (upper left), the crop mask (upper right), the crop types with the crop mask overlaid (lower left) and an RGB composite of the three reliability layers maximum probability, breaking ties and entropy.

Figure 3-61 shows a selected area north of Ulm (Westerstetten) and presents a detailed view of the crop type classification for the 13 crops/groups of crops. In supplement to the crop mask, the HRL 2015 Grassland layer is displayed, showing that the crop and grassland layers are complementary and could be well distinguished. A probability layer for the class winter wheat, as described in section 3.2.4.3 is visualized, showing that most of the fields classified as winter wheat (light blue) were classified with a very high probability and therefore have a high reliability. An example of the reliability layer 'breaking ties' as described in chapter 3.2.4.2 is displayed and explains that pixels with a high probability also have a relative high reliability to be classified as the right class.

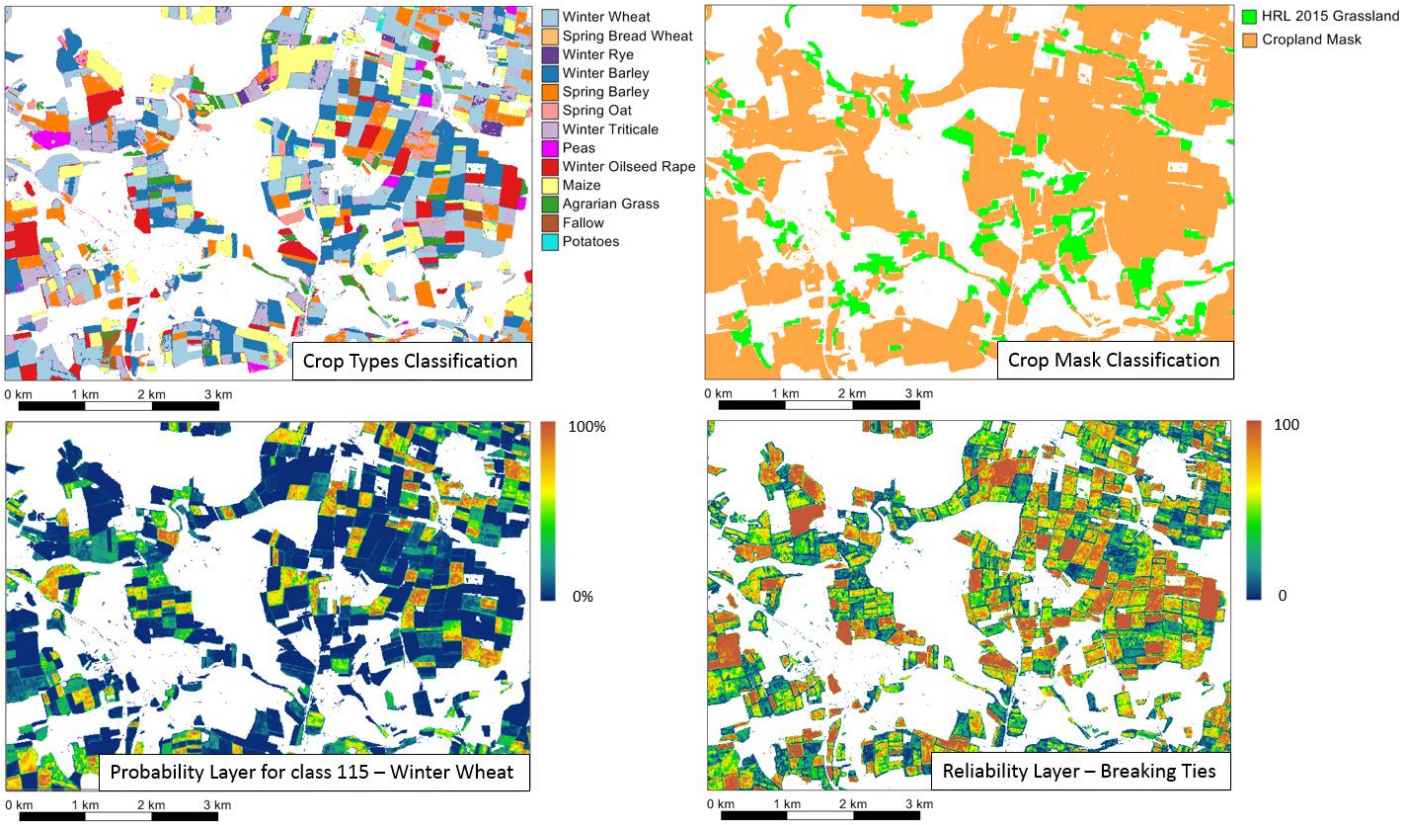


Figure 3-61. Top left: detailed view of the crop types classification for the 13 crops. Top right: crop mask classification together with the HRL 2015 Grassland layer. A good distinction between crops and grassland was achieved. Bottom right: Example of the reliability layer 'breaking ties' as described in chapter 3.2.4.2. Bottom left: probability layer for the class winter wheat.

3.2.4.1.5 Summary and conclusions

This study presents crop mask and type mapping using Sentinel-1 and -2 time series. This is performed using the Random Forest algorithm applied on several time features extracted from the time series, and Sentinel-1, -2 and combined approaches are compared. With the desired specifications for a potential future Copernicus HRL on agriculture not yet known, methods are developed and proposed. The tests presented in this study will be applied on the larger central prototyping site in the framework of WP 44.
For the crop mask as well as the crop types, the accuracies of the classifications based on S2 is significantly higher than those based on S1. Using both S1 and S2 increases the accuracies only marginally. As a consequence and in order to reduce the computational effort, the input data for the crop mask/types classification in the central site (and similar regions) could be restricted to Sentinel-2 data. Of course, this finding is not transferable to areas with higher cloud probability, where the integration of S1 data could be viable.

Similar classification accuracies can be achieved with data of the whole period (March to November 2017) and a limited period (March to mid-July). The third considered and shortest period (March to mid-June) shows significantly lower accuracies. It has also been analyzed if it is worth to include data of the late season of the previous year. However, the improvements are only marginal for all the considered scenarios (pixel/field sample unit for S1, S2 and S1&S2 respectively).

The accuracy gain of the evaluation based on field level is much higher for S1 compared to the modest gains for S2. The strong decrease in accuracy for the S1 based classifications is expectable due to the reduction of the speckle effect. Despite the applied multi-temporal filtering this is still present in the pixel

based result. Therefore, an object based classification could be particularly useful in case of S1 data, however it might be easier to derive the segments from the optical data.

Due to the eventually high computational cost of the feature calculation for the whole raster data footprint, it has been investigated if the number of features can be reduced significantly without a significant loss of classification accuracy. It has been shown that the optimal accuracy can be achieved with ca. 25% of the features. The feature selection could be further optimized by considering feature groups. These groups should comprise features such that the computational cost of calculating an additional feature of a specific group is relatively low compared to the calculation of a feature from another group. For example, it is computationally less expensive to calculate 10 features for one layer (e.g. NDVI) than 5 features for each of two different layers (e.g. NDVI and NDWI). This is simply because in the first case only half of the data (one layer) needs to be loaded, and in case of the percentile calculation sorted. Thus, a group-aware feature selection could further reduce the processing cost without loss of accuracy.

In addition to the primary class prediction result, the reliability layers can offer valuable information for secondary applications, e.g. the prioritization of likely incorrect field subsidy claims. It could be further investigated if the reliability layers can be further enhanced by improving the class probabilities they are derived from. For example, machine learning classifiers can be tuned to optimize the log-loss which is based on the class probabilities, and not an accuracy, which in contrast to the log-loss is only based on the binary information (correct or wrong). With log-loss, a false prediction that has a high probability is penalized much stronger than a false prediction with a lower probability. Instead of a typical accuracy score, the loss function only takes into account if a sample has been classified correct or false, but not the probabilities. Alternatively, it is also possible to calibrate the probabilities with a subset of the training samples in order to obtain improved probabilities with lower log-loss (Niculescu-Mizil and Caruana 2005). As a consequence of the improved probabilities the quality of the reliability layers increases.

The benchmarking results show promising accuracies and high potential of time series and derived time features for crop mask extraction and crop type monitoring. For a practical implementation of a future agricultural HRL, some more testing should be done when it comes to the differentiation of similar crop types, as well as regional diversity. But the high OA (> 90%) for the crop mask classification looks very promising for future applications, e.g., like a future HRL on Cropland.

3.2.4.2 Belgium site

3.2.4.2.1 Description of candidate methods

A random forest classifier is used for automatic crop type map production. This method has been selected based on the state-of-the-art review from (Inglada et al., 2015). This study showed overall accuracies above 80 % for most sites using this approach. Lower performances were attributed to fields smaller than the image resolution and fields with mix of trees and crops. These crop features are not expected in the Belgian test site from this benchmark. This method is fully automated but requires *in situ* data for the training.

3.2.4.2.2 Benchmarking criteria

Overall accuracy is reported for all benchmark scenarios to assess classification performances. A F-score for crop types is also provided in cases where low occurrences classes results were evaluated.

3.2.4.2.3 Implementation and results of benchmarking

Classifications were performed on the Belgium test site for the period 2017 with the preprocessed Sentinel-2 images as our optical data source. Object-based in situ data were obtained from the SIGEC

(Système intégré de gestion et de contrôles, Region Wallonia, Belgium). The area of interest is the Sentinel-2 tile 31UFR based on in situ data availability.

The method used linearly temporally gap-filled images as inputs for the classifier. To assess the performance of the random forest classifier we used several distinct inputs. We also tested for performance increase by cleaning border mixels in the in situ dataset and by oversampling the data prior to calibration of the model.

The classifier is performed on either Whittaker temporally gap-filled L2A images or L3A monthly composites. Mean composites and maximum NDVI composites for months with pixel coverage of 90 % or higher were used as inputs for the model. For 2017 only March, June, July, August, September and October were compliant. For each input, we extracted features for the model calibration: NDVI, NDWI and brightness in addition to the ten Sentinel-2 preprocessed bands.

The validation was operated independently from the calibration by splitting the dataset before operating the classifier. We used 25 % of the in situ data for the validation. From the remaining 75 %, 20 % were used for the model training. The data were selected randomly while keeping the proportion of each distinct crop type identical to the full dataset.

To improve the overall accuracy two methods were assessed. The former is a Synthetic Minority Over-sampling Technique (SMOTE) operated on the training in situ data to increase the sample size of minority crop types. The method was used to increase the sample sizes up to one thousand per crop type when in situ data were below this threshold. The latter is the removal of mixels located on field borders. We used a 15 meters buffer on the field objects to avoid border location errors as well as pixels values polluted by neighborhood land cover.

Every classification scenario achieved an overall accuracy higher than 80 % (Figure 3-62). The best results were obtained with the Whittaker temporally gap-filled L2A images as input and both SMOTE and mixel removal operations. Based on overall accuracy, results with SMOTE are not significantly different from random sample selection. Results from different inputs do not differ more than 1 % accuracy wise. Mixel removal improved accuracy by 4 to 5 % in every scenario.

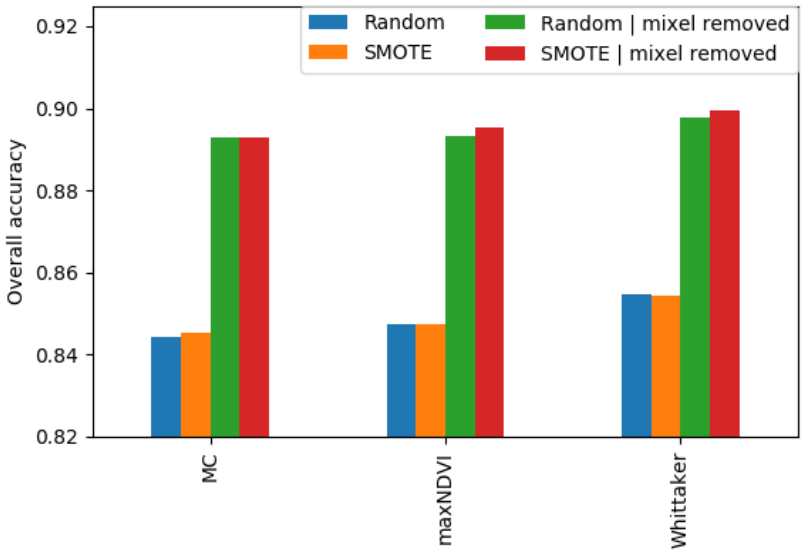


Figure 3-62. Overall accuracy for every classification scenario evaluated.

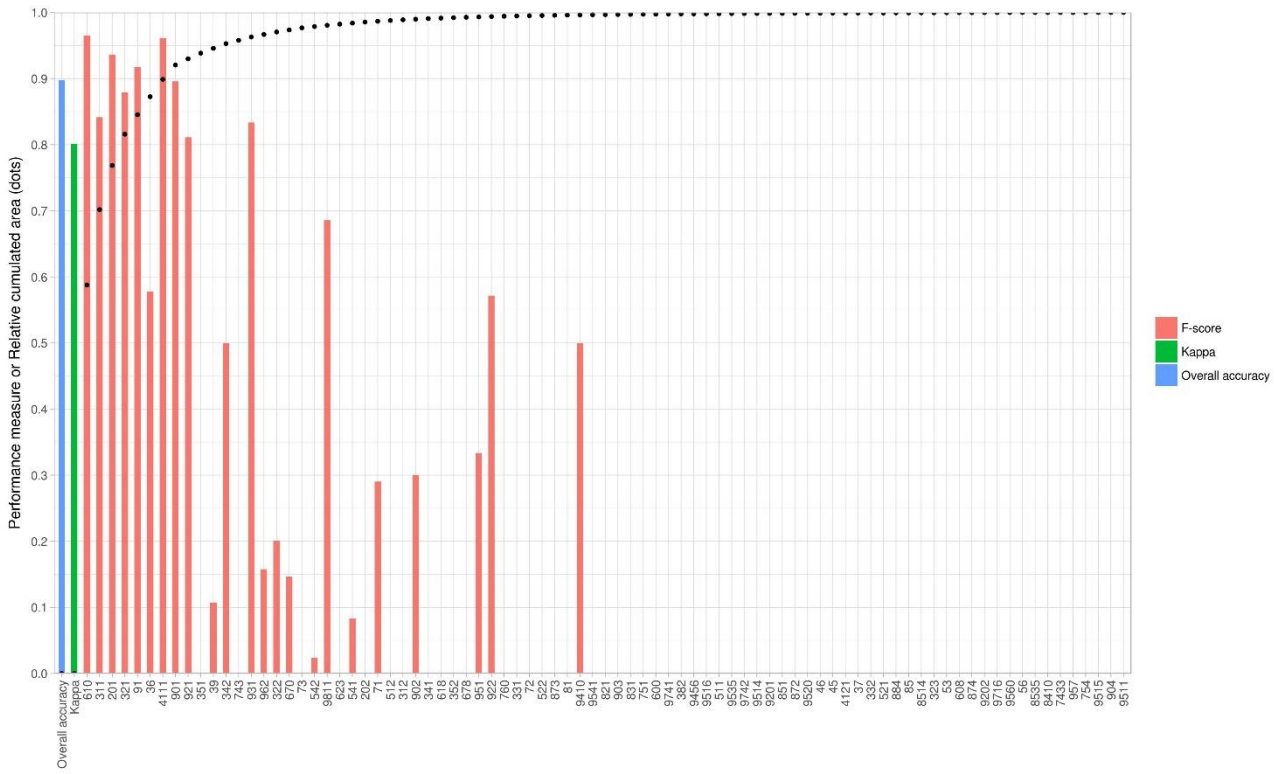


Figure 3-63. Classification F-score for each crop type ID for Whittaker inputs with random sampling and mixel removal (red). Overall accuracy (blue) for classification and Kappa (green). Relative cumulated area of crop types (black).

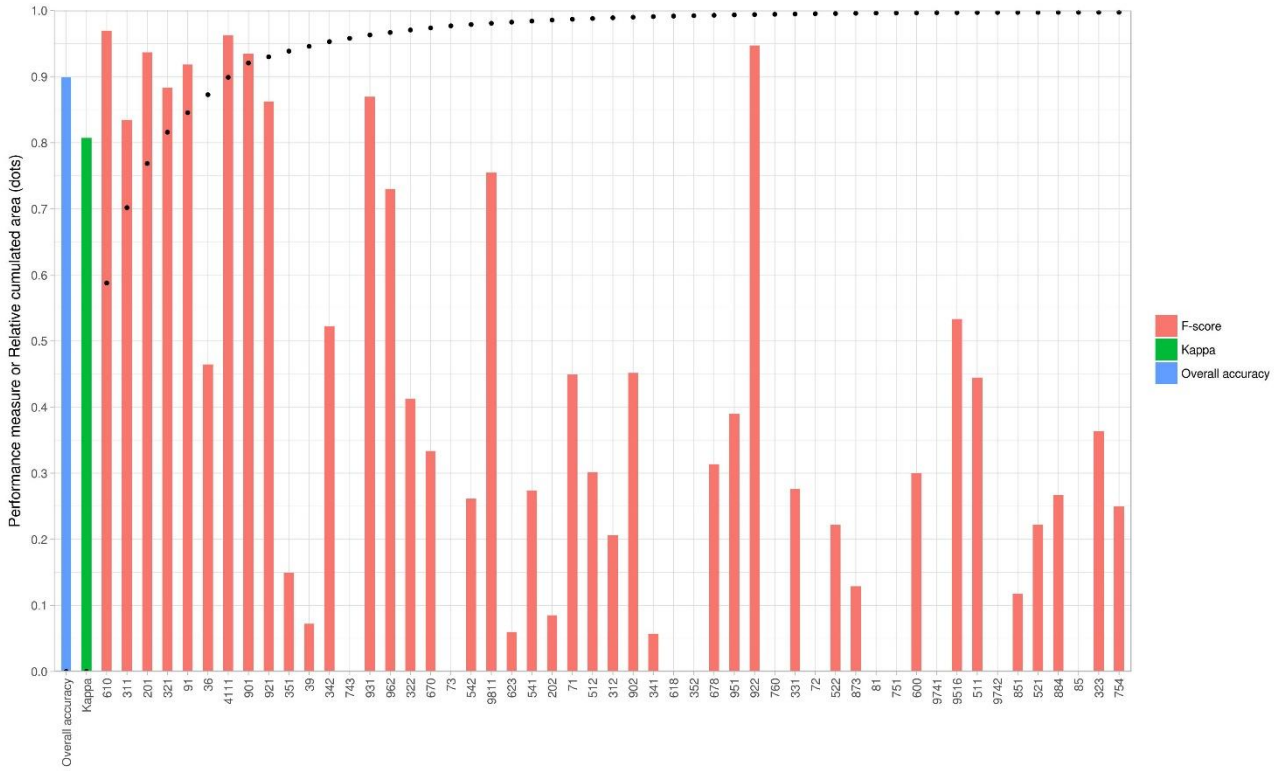


Figure 3-64. Classification F-score for each crop type ID for Whittaker inputs with SMOTE and mixel removal (red). Overall accuracy (blue) for classification and Kappa (green). Relative cumulated area of crop types (black).

To assess the performance improvement provided by the SMOTE method, classification errors on low occurrences classes must be evaluated. As seen in Figure 3-63, random sampling is not able to classify properly classes with small sample size in the *in situ* dataset. Figure 3-64 on the other hand shows that SMOTE method is useful to improve results for small occurrences classes.

3.2.4.2.4 Summary and conclusions

In conclusion, Whittaker temporally gap-filled L2A images provides better results than multiple monthly composites. The gap-filling process allow more features to be used in the classifier whereas composites still retains missing pixels from month to month. SMOTE method is necessary to be able to classify small occurrences classes. Mixel removal in the *in situ* dataset provides better results by allowing better features values for crop fields.

3.2.5 New land cover products

The aim of this part is to create a new land cover product that would be the fusion of all current HRLs, and new HRLs to fill the uncharacterized areas, such as cropland. Those four existing families of land covers (water bodies and wetlands, impervious areas, forests, grasslands) with an additional land cover for agricultural lands will aim at completely describing the pan-European territory.

3.2.5.1 Description of candidate methods

First of all, to produce a new land cover product, it has been decided to work on a simple set of cover classes. Only 6 classes have been selected in order to correctly separate the different pixels present in the tested AOI. The selected classes are then:

- grassland;
- summer croplands;
- winter croplands;
- forest;
- water;
- urban area.

Here, the following methodology, in the same spirit as the one exposed in the report of WP31, is explored:

- First, both testsite tiles are used, in the West-South of France (30TYP and 31TCJ). The year 2016 is used, and strongly cloudy images are set aside, leaving the study to focus on 14 exploitable images, which are still far from perfectly cloudless.
- Secondly, spectral indices, NDBI and NDVI, are computed for each image pre-processed by the JR, and statistical metrics, discussed in W31, are derived for each temporal period (spring-summer and autumn-winter) – which will be referred as ‘season’, in order to highlight different types of croplands.
- Thirdly, the map of persistent objects is generated, constituting a fixed skeleton. Databases are used: Open Street Maps (OSM), with a 5m buffer for the roads; the Permanent Water Bodies (HRL 2015), the Tree Cover Density (HRL 2015), with a MMU superior or equal to 5ha, and EU-Hydro, with a 2m buffer for “rivers” and a 5m buffer for “canals”. A priority order is set between those databases to ensure that each polygon has only one label.
- Fourthly, for each season, a selection of images, selected with a cloudfree surface represented at least 25% of the AOI, are layerstacked and a large-scale mean shift segmentation (LSMSS) is computed, at a MMU of 5ha. The generated shapefiles contain mean and variance information for polygons regarding the possible membership to each layer.
- Fifthly, the skeleton and the result of the previous segmentation are merged, at a MMU of 5ha.
- Sixthly, training data and validation samples are selected for each class, followed by a pixel-based classification, based on different statistics.
- Finally, temporal metrics are computed over the AOI.

The object-based classification may be tested in the second phase – the first attempts taken up on both S-2 tiles have been hugely resource-consuming without giving valuable results, and further steps need to be developed in order to tackle this restriction on the feasibility for production.

The various classification tested schemes, in the same spirit as those implemented for the imperviousness study, in section 2.3.1.

3.2.5.2 Benchmarking criteria

The benchmarking follows the same steps detailed in section 'benchmarking criteria' 2.3.1:

- A visual check – "look and feel" – is realized, excluding de facto methods presenting poor results;
- A thematic assessment measurement is produced, based on validation sampling extracted from the land cover map produced by the CESBIO over France (Inglada, et al., 2017a) – which has proven to be a reliable classification of land covers. Confusion matrices will be produced to quantify the divergence or the convergence between reference labels and produced labels.

For this part, the CESBIO map has been reclassified to match the 6 classes used in this study. Point samplings have been used, around 20 per classes for a full year, and down to 10 per classes for a season, but this choice of methodology could be challenged in the second phase – to investigate further the robustness of the land cover products, past the only thematic accuracy.

3.2.5.3 Implementation and results of benchmarking

As describe before, the benchmarking is only done on Sentinel-2 cloud-free images and they offer a high resolution (spectral and spatial). The implementation has been done on the test site in South-West site of France, over the tiles 30TYP and 31TCJ. Here, the following tests are proposed, after being sharpened by the section 2.3.1 on Imperviousness:

- one multitemporal classification (we use the layerstack of the stats)
- one classification per image and Dempster-Shafer precision method
- one classification per image and Dempster-Shafer recall method
- one classification per image and Dempster-Shafer accuracy method
- one classification per image and Dempster-Shafer kappa method

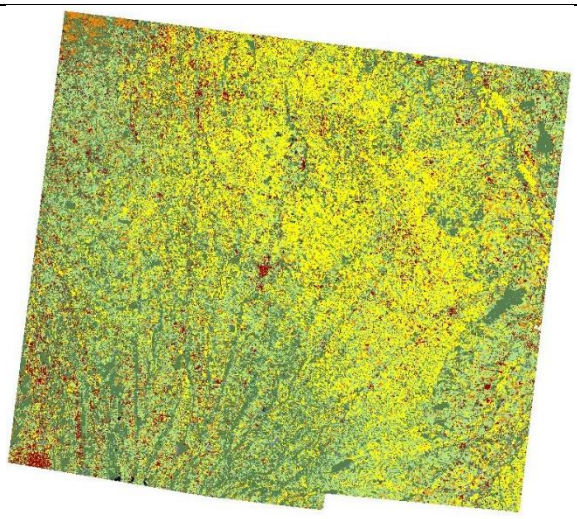
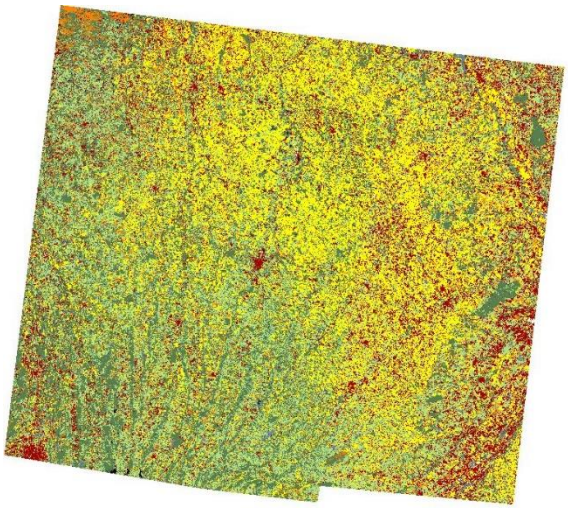
In order to differentiate crops, two temporal windows are tested – one without seasonality consideration, over the full span of the year 2016, and another with two distinct seasons, spring-summer and autumn-winter.

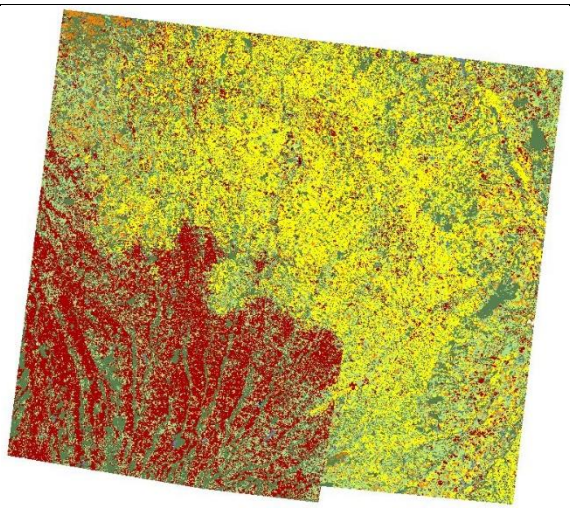
The results of the tests for the determination of the most relevant algorithm used for the Dempster-Shafer fusion of the classifications using the full Sentinel-2 pre-processed dataset divided into 2 seasons of 7 images each, composed of 10 bands are displayed in Table 3-52. Each matrix of confusion for the DS technics can be seen in Table 3-54 (based on the overall accuracy), in Table 3-5-5 (based on the kappa coefficient), in Table 3-56 (based on the precision rate) and in Table 3-57 (based on the recall rate).

Table 3-52. Summary of the tests to be implemented for the creation of the new LC products.

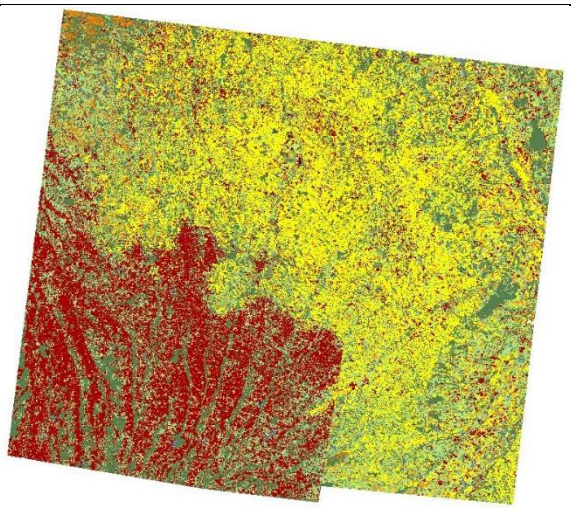
Test	Input Data	Fusion of classifications	Temporal window	Metrics used for the Dempster-Shafer fusion
1	Full dataset (14 useful images) – all 10 bands	Support Vector Machine	Full year	No Dempster-Shafer fusion needed
2	7 images per season – all 10 bands	Support Vector Machine	2 seasons (spring-summer and autumn-winter)	- Overall Accuracy - Kappa coefficient - Precision rate - Recall rate
3	7 images per season – only 5 significative bands, NDBI and NDVI	Support Vector Machine	2 seasons (spring-summer and autumn-winter)	- Overall Accuracy - Kappa coefficient - Precision rate - Recall rate
4	7 images per season – only 5 significative bands	Support Vector Machine	2 seasons (spring-summer and autumn-winter)	- Overall Accuracy - Kappa coefficient - Precision rate - Recall rate
5	7 images per season – only NDBI and NDVI with their statistical metrics over each season	Support Vector Machine	2 seasons (spring-summer and autumn-winter)	- Overall Accuracy - Kappa coefficient - Precision rate - Recall rate

Table 3-53. Visual check of the different algorithms for Dempster-Shafer fusion using the full dataset with all spectral bands as input.

	
Precision rate	Recall rate



Overall accuracy



Kappa

Table 3-54. Confusion matrix for the DS fusion of 2 seasons, 7 images each with all spectral bands, based on the overall accuracy.

All 10 bands Fusion DS – OA		Reference						Total	User Accuracy
		Grassland	Cropland 1	Cropland 2	Forest	Water	Urban Area		
Product	Grassland	6	1	2	0	0	0	9	67%
	Cropland 1	1	7	3	0	0	0	11	64%
	Cropland 2	0	1	9	0	0	0	10	90%
	Forest	0	0	0	10	0	0	10	100%
	Water	0	0	0	1	9	0	10	90%
	Urban Area	0	0	0	0	0	10	10	100%
Total		7	9	14	11	9	10	60	
Producer Accuracy		86%	78%	64%	91%	100%	100%		85%

Table 3-55. Confusion matrix for the DS fusion of 2 seasons, 7 images each with all spectral bands, based on the kappa coefficient.

All 10 bands Fusion DS – KC		Reference						Total	User Accuracy
		Grassland	Cropland 1	Cropland 2	Forest	Water	Urban Area		
Product	Grassland	7	1	2	0	0	0	10	70%
	Cropland 1	0	7	4	0	0	0	11	64%
	Cropland 2	0	1	8	0	0	0	9	89%
	Forest	0	0	0	10	0	0	10	100%
	Water	0	0	0	1	9	0	10	90%
	Urban Area	0	0	0	0	0	10	10	100%
Total		7	9	14	11	9	10	60	
Producer Accuracy		100%	78%	57%	91%	100%	100%		85%

Table 3-56. Confusion matrix for the DS fusion of 2 seasons, 7 images each with all spectral bands, based on the precision rate.

All 10 bands Fusion DS precision		Reference						Total	User Accuracy
		Grassland	Cropland 1	Cropland 2	Forest	Water	Urban Area		
Product	Grassland	7	1	2	0	0	0	10	70%
	Cropland 1	0	7	5	0	0	0	12	58%
	Cropland 2	0	1	7	0	0	0	8	88%
	Forest	0	0	0	10	0	0	10	100%
	Water	0	0	0	1	9	0	10	90%
	Urban Area	0	0	0	0	0	10	10	100%
Total		7	9	14	11	9	10	60	
Producer Accuracy		100%	78%	50%	91%	100%	100%		83%

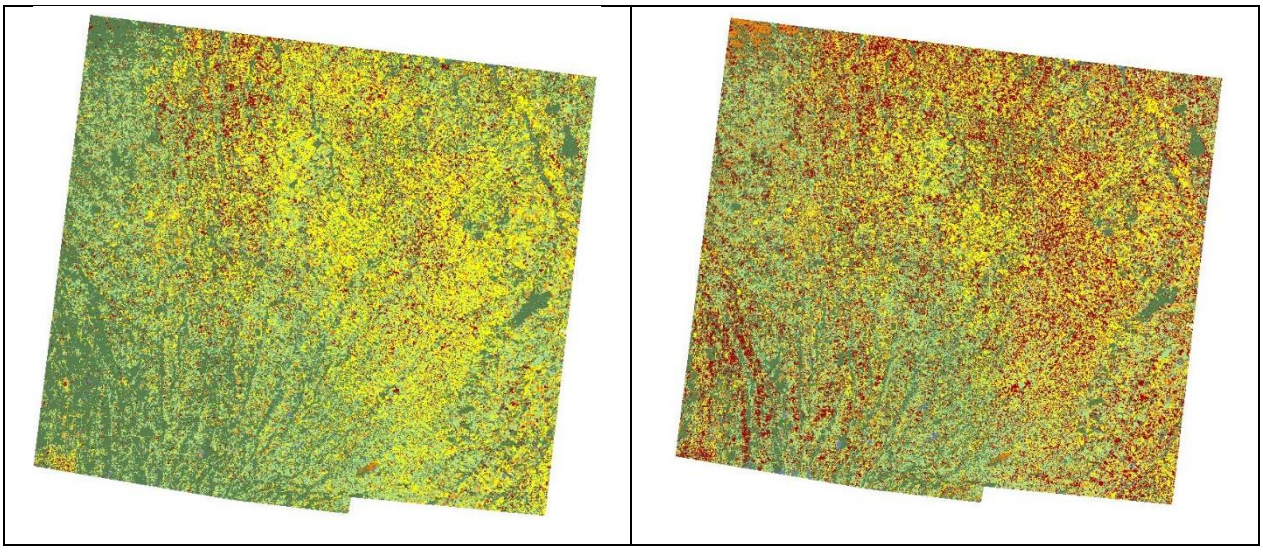
Table 3-57. Confusion matrix for the DS fusion of 2 seasons, 7 images each with all spectral bands, based on the recall rate.

All 10 bands Fusion DS recall		Reference						Total	User Accuracy
		Grassland	Cropland 1	Cropland 2	Forest	Water	Urban Area		
Product	Grassland	7	1	2	0	0	0	10	70%
	Cropland 1	0	7	3	0	0	0	10	70%
	Cropland 2	0	1	7	0	0	0	8	88%
	Forest	0	0	0	10	0	0	10	100%
	Water	0	0	1	1	9	0	11	82%
	Urban Area	0	0	1	0	0	10	11	91%
Total		7	9	14	11	9	10	60	
Producer Accuracy		100%	78%	50%	91%	100%	100%		83%

The look and feel is satisfying, even more regarding the Dempster-Shafer fusion using the precision and recall rate, which is confirmed by the associated confusion matrices. However, the overall accuracy and the kappa coefficient seem sensitive to the number of images used in the time series – yet this is not apparent in the matrices. During the next step on the demonstration sites, it will be mandatory to work on increasing the number of validation points.

The next set of tests uses the full Sentinel-2 pre-processed dataset divided into 2 seasons of 7 images each, this time composed of the 5 most significant spectral bands with NDVI and NDBI. The visual check is displayed in Table 3-58. Each matrix of confusion for the DS technics can be seen in Table 3-59 (based on the overall accuracy), in Table 3-60 (based on the kappa coefficient), in Table 3-61 (based on the precision rate) and in Table 3-62 (based on the recall rate).

Table 3-58. Visual check of the different algorithms for Dempster-Shafer fusion using the full dataset with the 5 most significant spectral bands and the NDBI and NDVI as input.



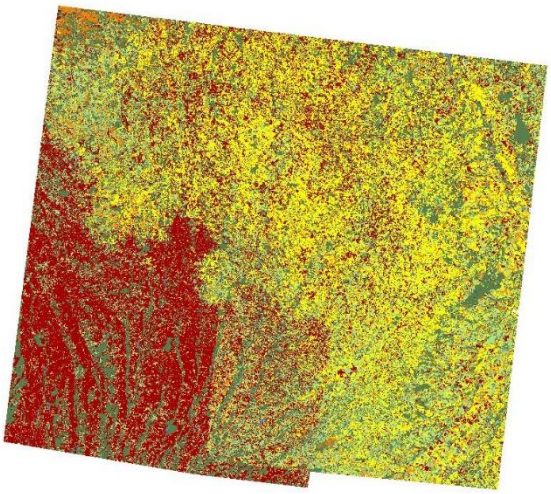
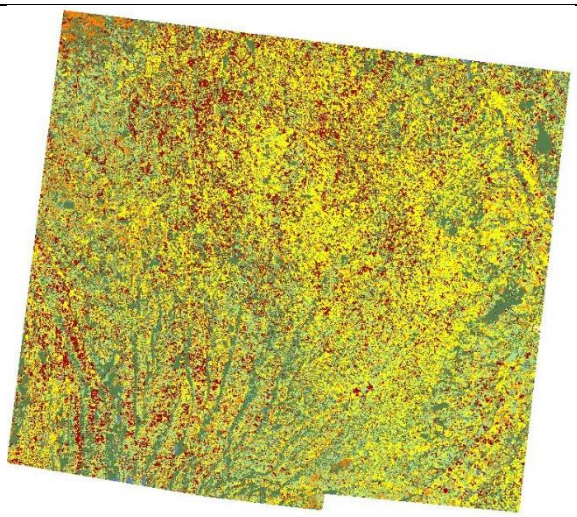
Precision rate	Recall rate
	
Overall accuracy	Kappa coefficient

Table 3-59. Confusion matrix for the DS fusion of 2 seasons, 7 images each with NDBI, NDVI and 5 spectral bands, based on the overall accuracy.

All 10 bands Fusion DS – OA		Reference						Total	User Accuracy
		Grassland	Cropland 1	Cropland 2	Forest	Water	Urban Area		
Product	Grassland	7	1	2	0	0	0	10	70%
	Cropland 1	0	6	2	0	0	0	8	75%
	Cropland 2	0	2	9	0	0	0	11	82%
	Forest	0	0	0	10	0	0	10	100%
	Water	0	0	0	1	9	0	10	90%
	Urban Area	0	0	1	0	0	10	11	91%
Total		7	9	14	11	9	10	60	
Producer Accuracy		100%	67%	64%	91%	100%	100%		85%

Table 3-60. Confusion matrix for the DS fusion of 2 seasons, 7 images each with NDBI, NDVI and 5 spectral bands, based on the kappa coefficient.

All 10 bands Fusion DS – KC		Reference						Total	User Accuracy
		Grassland	Cropland 1	Cropland 2	Forest	Water	Urban Area		
Product	Grassland	7	1	2	0	0	0	10	70%
	Cropland 1	0	6	2	0	0	0	8	75%
	Cropland 2	0	2	9	0	0	0	11	82%
	Forest	0	0	0	10	0	0	10	100%
	Water	0	0	0	1	9	0	10	90%
	Urban Area	0	0	1	0	0	10	11	91%
Total		7	9	14	11	9	10	60	
Producer Accuracy		100%	67%	64%	91%	100%	100%		85%

Table 3-61. Confusion matrix for the DS fusion of 2 seasons, 7 images each with NDBI, NDVI and 5 spectral bands, based on the precision rate.

All 10 bands Fusion DS precision		Reference						Total	User Accuracy
		Grassland	Cropland 1	Cropland 2	Forest	Water	Urban Area		
Product	Grassland	7	1	2	0	0	0	10	70%
	Cropland 1	0	6	2	0	0	0	8	75%
	Cropland 2	0	2	9	0	0	0	11	82%
	Forest	0	0	0	10	0	0	10	100%
	Water	0	0	0	1	9	0	10	90%
	Urban Area	0	0	1	0	0	10	11	91%
Total		7	9	14	11	9	10	60	
Producer Accuracy		100%	67%	64%	91%	100%	100%		85%

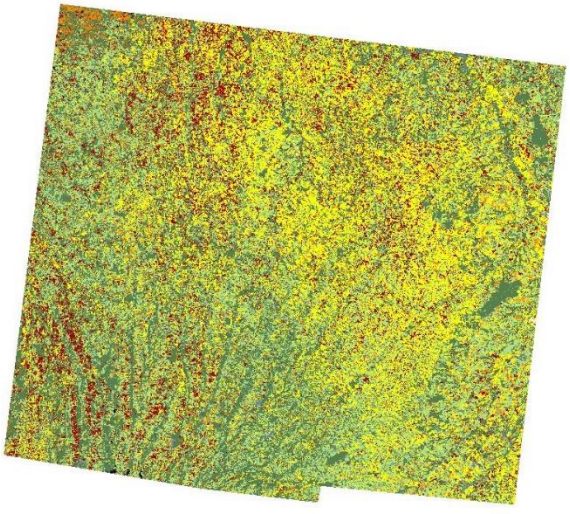
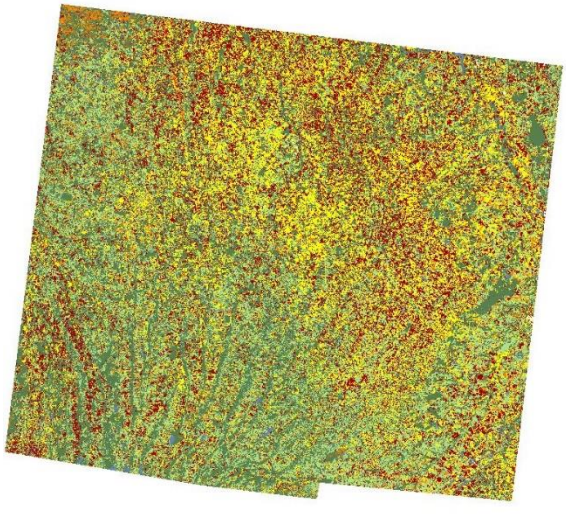
Table 3-62. Confusion matrix for the DS fusion of 2 seasons, 7 images each with NDBI, NDVI and 5 spectral bands, based on the recall rate.

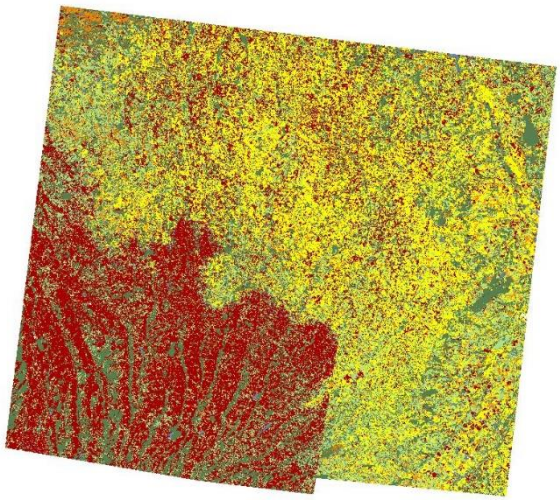
All 10 bands Fusion DS recall		Reference						Total	User Accuracy
		Grassland	Cropland 1	Cropland 2	Forest	Water	Urban Area		
Product	Grassland	7	1	2	0	0	0	10	70%
	Cropland 1	0	5	1	0	0	0	6	83%
	Cropland 2	0	1	7	0	0	0	8	88%
	Forest	0	0	2	10	0	0	12	83%
	Water	0	0	1	1	9	0	11	82%
	Urban Area	0	2	1	0	0	10	13	77%
Total		7	9	14	11	9	10	60	
Producer Accuracy		100%	56%	50%	91%	100%	100%		80%

The most significant spectral bands (2, 3, 4, 7 and 9) in combination with the NDVI and NDBI has strongly improved the results for the kappa coefficient, which can be seen in the look and feel as well as in the confusion matrix. The overall accuracy is still sensitive to the image mosaicking.

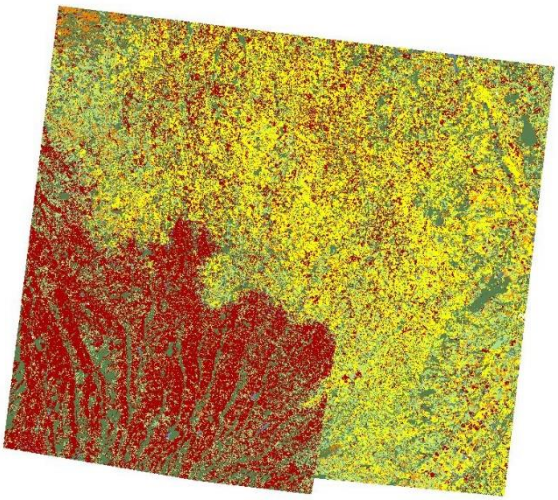
The final set of tests focuses on the full Sentinel-2 pre-processed dataset divided into 2 seasons of 7 images each with only their 5 most significant spectral bands. The visual checking of the results can be seen in Table 3-63. Each matrix of confusion for the DS technics can be seen in Table 3-64 (based on the overall accuracy), in (based on the kappa coefficient), in Table 3-66 (based on the precision rate) and in (based on the recall rate).

Table 3-63. Visual check of the different algorithms for Dempster-Shafer fusion using the full dataset with the 5 most significant spectral bands as input.

	
Precision rate	Recall rate



Overall accuracy



Kappa coefficient

Table 3-64. Confusion matrix for the DS fusion of 2 seasons, 7 images each with 5 spectral bands, based on the overall accuracy.

All 10 bands Fusion DS – OA		Reference						Total	User Accuracy
		Grassland	Cropland 1	Cropland 2	Forest	Water	Urban Area		
Product	Grassland	6	1	2	0	0	0	9	67%
	Cropland 1	0	6	2	0	0	0	8	75%
	Cropland 2	0	2	9	0	0	0	11	82%
	Forest	0	0	0	10	0	0	10	100%
	Water	0	0	0	1	9	0	10	90%
	Urban Area	1	0	1	0	0	10	12	83%
Total		7	9	14	11	9	10	60	
Producer Accuracy		86%	67%	64%	91%	100%	100%		83%

Table 3-65. Confusion matrix for the DS fusion of 2 seasons, 7 images each with 5 spectral bands, based on the kappa coefficient.

All 10 bands Fusion DS – KC		Reference						Total	User Accuracy
		Grassland	Cropland 1	Cropland 2	Forest	Water	Urban Area		
Product	Grassland	7	1	2	0	0	0	10	70%
	Cropland 1	0	6	2	0	0	0	8	75%
	Cropland 2	0	2	9	0	0	0	11	82%
	Forest	0	0	0	10	0	0	10	100%
	Water	0	0	0	1	9	0	10	90%
	Urban Area	0	0	1	0	0	10	11	91%
Total		7	9	14	11	9	10	60	
Producer Accuracy		100%	67%	64%	91%	100%	100%		85%

Table 3-66. Confusion matrix for the DS fusion of 2 seasons, 7 images each with 5 spectral bands, based on the precision rate.

All 10 bands Fusion DS precision		Reference						Total	User Accuracy
		Grassland	Cropland 1	Cropland 2	Forest	Water	Urban Area		
Product	Grassland	7	1	2	0	0	0	10	70%
	Cropland 1	0	7	3	0	0	0	10	70%
	Cropland 2	0	1	8	0	0	0	9	89%
	Forest	0	0	0	10	1	0	11	91%
	Water	0	0	0	1	8	3	12	67%
	Urban Area	0	0	1	0	0	7	8	88%
Total		7	9	14	11	9	10	60	
Producer Accuracy		100%	78%	57%	91%	89%	70%		78%

Table 3-67. Confusion matrix for the DS fusion of 2 seasons, 7 images each with 5 spectral bands, based on the recall rate.

All 10 bands Fusion DS recall		Reference						Total	User Accuracy
		Grassland	Cropland 1	Cropland 2	Forest	Water	Urban Area		
Product	Grassland	7	1	2	0	0	0	10	70%
	Cropland 1	0	5	1	0	0	0	6	83%
	Cropland 2	0	1	8	0	0	0	9	89%
	Forest	0	0	1	10	0	0	11	91%
	Water	0	1	1	1	9	0	12	75%
	Urban Area	0	1	1	0	0	10	12	83%
Total		7	9	14	11	9	10	60	
Producer Accuracy		100%	56%	57%	91%	100%	100%		82%

Without the combination of spectral indices, the NDVI and NDBI, the results, as seen in both the look and feel and the confusion matrices, are quite degraded – comforting us in the use of indices in the time series. The kappa coefficient still gives the best results.

3.2.5.4 Summary and conclusions

The use of object-based classifications is too time consuming (with the current process) so it seems that more work will be required to test it on time-series analysis.
There is no predominant fusion method for monodate pixel-based classifications – however, the best results are obtained when two temporal frames are used to separate the various type of crops into two families.

The best results are obtained for the full set of spectral bands, with the precision or recall rate methods – however, it is closely followed in term of performance by the spectral index metrics, using the kappa coefficient.

It should be highlighted that a deeper study is required to totally exclude some of the listed methodologies. For the demonstration on the various sites over Europe, several issues will need to be addressed:

- Enforcing a uniform set of validation sampling, extracted from viable reference sources;
- Resolving the current inability to run an object-based classification on the 2-tile testsite, in order to compare the results with the pixel-based classification;
- Realizing a denser time series, to obtain more than just two seasons – the current results look promising to create a new land cover product.

In the second phase, more fusion algorithms could also be tested, such as the majority filter.

4 Conclusions and outlook

This report presents a methods compendium for the WP33 - Time Series Analysis for Thematic Classification, which aims to develop a framework for time series analysis for thematic classification based on Sentinel multi-sensor constellation. With the others WP of ECoLaSS Task 3 (Automated High Data Volume Processing Lines), it constitutes a basis for the demonstration activities of Task 4 (Thematic Proof-of-Concept/Prototype on Continental/Global Scale), i.e. High Resolution Layers (HRLs), Grassland, Crop type and new LC/LU products.

The first part of the document describes the state-of-the-art methods and strategies for the selection of candidate methods for the benchmarking. It reviews the automated reference sampling methods and the image compositing methods needed for classification, and then provides state-of-the-art of time series classification methods for time series HRLs, agriculture and new land cover products.

The second part concerns the testing and benchmarking of input data for classification (automated reference sampling and image compositing methods) and of time series classification approaches selected. The latter are performed separately for different thematic fields: (i) Imperviousness, (ii) Forest, (iii) Grassland, (iv) Agriculture, and (v) new land cover products.

The benchmark of the automated reference sampling methods concluded that the iForest exhibits additional important properties valuable for an outlier detection method. It is therefore suitable to be used for such purposes in future applications. Several other approaches could be tested, for instance potential thresholding approaches to know the fraction of outliers, or the use of decision function values as instance weights when using the automatically sampled reference data. Further research is also required in order to better understand why the outlier detection of the non-forest class failed.

The compositing methods benchmark on S-2 images highlighted the importance of a performant cloud mask for such time series that is not as dense as medium resolution time series. With such a cloud mask that still present too many artefacts concerning delineation of cloud borders, the haze and cirrus detection and removal, the detection of cloud shadows and cloud commission for bright surfaces, the two feature-based algorithms are more appropriated as they achieve more spatial consistency and very few data gaps thanks to the use of the entire time series as input. On the contrary, the three time interval algorithms present many artefacts due to undetected clouds/cloud shadows and high confusion with bright surfaces in the cloud mask, and data gaps due to the short compositing period and a time series not dense enough. More specifically, other quantiles could be computed in phase two for the Quantile Compositing method.

The benchmark of the time series classification methods is performed on S-1 and S-2 data for Imperviousness, Forest, Grassland, Agriculture and New land cover products.

First, for Imperviousness, the analysis shows better results for a mono-temporal approach, the use of an active learning or SVM classifier and the input being all data available (or subset based on the best available cloud-free images) with both sensors S-1 and S-2. The active learning algorithm shows great classification performances whilst being very computer efficient, while The SVM classifier shows interesting results as an alternative method. The approach based on both sensors S-1 and S-2 shows the interest to use data fusion. The mono-source approach, based on one HR sensor, S-1/2, is not sufficient. The optical time series, in particular, is not dense enough to take advantage of the phenology of inter-yearly and intra-yearly seasonal dynamics. Further investigations need to be undertaken to enhance the efficiency of the classifiers and to explore the multi-sourcing approach with other sensors including S-3

Second, the potential of combining S-2 and S-1 data for the Forest delineation is assessed by applying a random forest classifier to five experiments, using different combinations of sensors and time periods. For this analysis, the gain of the combined use of S-2 and S-1 time features compared to only focusing on S-2 data is insignificant. Indeed, the use of S-2 data limited to the spring period provided the best ratio of high

accuracy and lowest benchmarking cost. Considering the limited availability of cloud free satellite scenes and heterogeneous character of the analysed forest types in the area of interest, the results are very promising for future application on larger areas. However, further research is required to validate the transferability to areas of different geographic conditions and seasonal patterns. Other further improvements concerns the detailed feature analysis of misclassified areas or pixels with low classifier probabilities.

Third, the Grassland classification benchmark highlights the potential of SAR data for the grassland classification and that the SAR threshold based grassland classification highly depends on dense time series. Largest misclassifications occur for water bodies, bare soil, and artificial surfaces. These areas can however easily be removed with optical data. The aggregated classification result with SAR and OPTICAL combined datasets are quite encouraging. More confusion between grasslands and cropland are present when using optical data only, whereas more misclassification between grassland and roads are present when using SAR data only. The combined approach shows more homogenous patches than using SAR data only. A further approach will be the combination of SAR features with vegetation indices derived from the optical data set. Recommendations for the demonstration sites in Task 4 is the application of the supervised random forest based approach, the precise pre-processing of the dense time series including a topographic normalisation for hilly to mountainous terrain, the application of multi-temporal filtering on gamma naught corrected imagery for SAR time series. Further research is specifically required to determine the optimal combination of features and indices derived from the optical as well as SAR dense time series.

Fourth, the Agriculture classification benchmark is performed on Central (Germany) and Belgium site. The benchmarking results show promising accuracies and high potential of time series and derived time features for crop mask extraction and crop type monitoring. As for the Forest benchmark, using both S-1 and S-2 increases the accuracies only marginally. The accuracies of the classifications based on S-2 is significantly higher than those based on S1. In order to reduce the computational effort, the input data for the crop mask/types classification similar regions than central site could be restricted to S-2 data. In order to reduce the processing cost without loss of accuracy, a group-aware feature could further selected. In addition, it could be further investigated if the reliability layers can be further enhanced by improving the class probabilities they are derived from. For a practical implementation of a future agricultural HRL, some more testing should be done when it comes to the differentiation of similar crop types, as well as regional diversity. For a practical implementation of a future agricultural HRL, some more testing should be done when it comes to the differentiation of similar crop types, as well as regional diversity.
Finally, the benchmark of classification methods for new land cover products concludes that the best results are obtained for the full set of spectral bands, closely followed in term of performance by the spectral index metrics. There is no predominant fusion method for mono-date pixel-based classifications. However, the best results are obtained when two temporal frames are used to separate the various type of crops into two families. Several issues need to be addressed such as enforcing a uniform set of validation sampling, resolving the current inability to run an object-based classification on the 2-tile testsite and realizing a denser time series, to obtain more than just two seasons. In the second phase, more fusion algorithms could also be tested, such as the majority filter.

The ECoLaSS project follows a two-phased approach of two times 18 months duration. This deliverable comprises the first issue. In the second 18-month project cycle, a second issue of this deliverable will be published, containing all relevant updates concerning the benchmarking of input data for classification as well as the time series classification methods.

References

Achard, F., Beuchle, R., Mayaux, P., Stibig, H.-J., Bodart, C., Brink, A., Carboni, S., Desclée, B., Donnay, F., Eva, H.G., Lupi, A., Raši, R., Seliger, R. and Simonetti, D. (2002). Determination of tropical deforestation rates and related carbon losses from 1990 to 2010. Clobal Change Biology, 2014.

Alajlan, N., Bazi, Y., AlHichri, H. S., Melgani, F., & Yager, R. R. (2013). Using OWA Fusion Operators for the Classification of Hyperspectral Images. *IEEE Journal of Selected Topics in Applied Earth Observations*, 602-614.

Banskota, A., Kayastha, N., Falkowski, M. J., Wulder, M. A., Froese, R. E. and White, J. C. (2014). Forest Monitoring Using Landsat Time Series Data: A Review. Canadian Journal of Remote Sensing, 2014.

Benediktsson, J., Swain, P., & Ersoy, O. (1990). Neural network approaches versus statistical methods in classification of multisource remote sensing data. *IEEE Transactions on Geoscience and remote Sensing, 28*(4), 540-552.

Bauer, M., B. Loffelholz & B. Wilson. 2007. Estimating and Mapping Impervious Surface Area by Regression Analysis of Landsat Imagery. In *Remote Sensing of Impervious Surfaces*.

Blaes, X. and Defourny, P. (2003). Retrieving crop parameters based on tandem ERS 1/2 interferometric coherence images. *Remote Sensing of Environment*, Vol. 88, No. 4, 374–385.

Blaes, X., Vanhalle, L. and Defourny, P. (2005). Efficiency of crop identification based on optical and SAR image time series, *Remote Sensing of Environment*, Vol. 96 No. 3-4, 352–365.

Betbeder, J., Rapinel, S., Corgne, S., Pottier, E., and Hubert-Moy, L. (2015). TerraSAR-X dual-pol time-series for mapping of wetland vegetation. *ISRSP Journal of Photogrammetry and Remote Sensing*, Vol. 107, 90-98.

Bingfang Wu, Jihua Meng, Qiangzi Li, Nana Yan, Xin Du & Miao Zhang (2014). Remote sensing-based global crop monitoring: experiences with China's CropWatch system, International Journal of Digital Earth, 7:2, 113-137

Bock, M. and Lessing, R. (2000). Remote sensing, formation of objects and determination of quality. In: Cremers, A.B. and Greve, K. (Eds.). *EnviroInfo 2000: Umweltinformatik '00 Umweltinformation für Planung, Politik und Öffentlichkeit, Bonn*, Metropolis Verlag, Marburg.

Bock, M., Rossner, G., Wissen, M., Remm, K., Langanke, T., Lang, S., Klug, H., Blaschke, T. and Vrščaj, B. (2005a). Spatial indicators for nature conservation from European to local scale. *Ecological Indicators*, Vol. 5, No. 4, 322–338.

Bock, M., Xofis, P., Mitchley, J., Rossner, G. and Wissen, M. (2005b). Object-oriented methods for habitat mapping at multiple scales – Case studies from Northern Germany and Wye Downs, UK. *Journal for Nature Conservation*, Vol. 13, No. 2–3, 75–89.

Braun, M. 2004. Mapping imperviousness using NDVI and linear spectral unmixing of ASTER data in the Cologne-Bonn region (Germany). In *Proceedings of SPIE*, 274-284.

Breiman, L. (1984). *Classification and Rgression Trees.* Chapman & Hall/CRC.

Breiman, L. (2001). Random Forests. *Machine Learning, 4*(1), 5-32.

Brisco, B. and Brown, R.J. (1995). Multidate SAR/TM Synergism for Crop Classification in Western Canada. *Photogrammetric Engineering & Remote Sensing*, Vol. 91, No. 8, 1009–1014.

Buck, O., Klink, A., Millán, V. E. G., Pakzad, K. and Müterthies, A. (2013). Image Analysis Methods to Monitor Natura 2000 Habitats at Regional Scales – the MS. MONINA State Service Example in Schleswig-Holstein, Germany. *Photogrammetrie - Fernerkundung - Geoinformation*, Vol. 2013, No. 5, 415–426.

Buck, O., Millán, V. E. G., Klink, A., & Pakzad, K. (2015). Using information layers for mapping grassland habitat distribution at local to regional scales. International Journal of Applied Earth Observation and Geoinformation, Vol. 37, 83-89.

Cabral, A., de Vasconcelos, M.J.P., Pereira, J.M.C., Bartholome, E. and Mayaux, P. (2003). Multitemporal compositing approaches for SPOT-4 VEGETATION data. *International Journal of Remote Sensing*, 24, 3343–3350.

Camp-Valls, G., & Bruzzone, L. (2009). *Kernek methods for remote sensing data analysis.* John Wiley & Sons.

Carlinet, E., & Géraud, T. (2014). A Comparative Review of Component TreeComputation Algorithms. *IEEE Transactions on Image Processing, 23*(9), 3885-3895.

Chawla, N., Bowyer, K., Hall, L. and Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, 16, 321-357.

Chen, J., Chen, J., Liao, A., Cao, X., Chen, L., Chen, X., He, C., Han, G., Peng, S., Lu, M. et al. (2015). Global land cover mapping at 30m resolution: A POK-based operational approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 103:7–27.

Cihlar, J., Li, Z., Chen, J., Pokrant, H. and Huang, F. (1997). Multitemporal, multichannel AVHRR data sets for land biosphere studies – Artefacts and corrections. *Remote Sensing of Environment*, 60, 35–57.

Cihlar, J., Manak, D. and Voisin, N. (1994a). AVHRR bi-directional reflectance effects and compositing. *Remote Sensing of Environment*, 48, 77–88.

Cohen, W. B., Spies, T. A., Alig, R. J., Oetter, D. R., Maiersperger, T. K. and M. Fiorella (2002). Characterizing 23 Years (1972–95) of Stand Replacement Disturbance in Western Oregon Forests with Landsat Imagery. Ecosystems, 2002.

Colditz, R., Lopez Saldana, G., Maeda, P., Argumedo Espinoza, J., Meneses Tovar, C., Victoria Hernandez, A., Zermeno Benitez, C., Cruz Lopez, I. and Ressl, R. (2012). Generation and analysis of the 2005 land cover map for Mexico using 250 m MODIS data. *Remote Sensing of Environment*, 123, 541–552.

Comber, A., Fisher, P., Wadsworth, R. (2005). What is land cover? Environment and Planning B: Planning and Design 2005, volume 32, pages 199-209. doi:10.1068/b31135.

Coppin, P. R. and Bauer, M. E. (1996). Change Detection in Forest Ecosystems with Remote Sensing Digital Imagery. Remote Sensing Reviews, 1996.

Corbane, C., Alleaume, S. and Deshayes, M. (2013). Mapping natural habitats using remote sensing and sparse partial least square discriminant analysis. *International Journal of Remote Sensing*, Vol. 34, No. 21, 7625–7647.

Corbane, C., Lang, S., Pipkins, K., Alleaume, S., Deshayes, M., Millán, V. E. G., Strasser, T., Vanden Borre, J., Toon, S. and Förster, M. (2015). Remote sensing for mapping natural habitats and their conservation status – New opportunities and challenges. *International Journal of Applied Earth Observation and Geoinformation*, Vol. 37, 7–16.

Crawford, M., Tuia, D., & Ynag, H. (2013). Active Learning: Any Value for Classification of Remotely Sensed Data? *Proceedings of the IEEE*, 593-608.

Cui, T., Martz, L., & Guo, X. (2017). Grassland Phenology Response to Drought in the Canadian Prairies. *Remote Sensing*, Vol. 9, No. 12, 1258.

D'Iorio, M.A., Cihlar, J. and Morasse, C.R. (1991). Effect of the calibration of AVHRR data on the normalised difference vegetation index and compositing. *Canadian Journal of Remote Sensing*, 17, 251–262.

Dalla Mura, M., Benediktsson, J., Waske, B., & Bruzzone, L. (2010, October). Morphological Attribute Profiles for the Analysisof Very High Resolution Images. *IEEE Transactions on Geoscience and Remote Sensing, 48*(10), 3747-3761.

Davidson, A.M. (2016). Review of satellite image classification methods. Internal document. Agriculture and Agri-Food Canada: Ottawa.

De Wasseige, C., Vancutsem, C. and Defourny, P., 2000, Sensitivity analysis of compositing strategies: Modelling and experimental investigations. *In VEGETATION 2000 conference*: Two years of operation to prepare the future, 21020 Ispra, Varese-Italy, G. Saint (Ed.), 267–274.

Díaz Varela, R., Ramil Rego, P., Calvo Iglesias, S. and Muñoz Sobrino, C. (2008). Automatic habitat classification methods based on satellite images: A practical assessment in the NW Iberia coastal mountains. *Environmental Monitoring and Assessment*, Vol. 144, No. 1, 229–250.

DiGregorio, A. (2013). A cropland nomenclature conform to the FAO Land Cover Meta-Language. *SIGMA Technical Report*.

Duchemin, B. and Maisongrande, P. (2002). Normalisation of directional effects in 10-day global syntheses derived from VEGETATION/SPOT: I. Investigation of concepts based on simulation. *Remote Sensing of Environment*, 81, 90–100.

Eklundh, L. and Jönsson, P (2015). TIMESAT: A Software Package for Time-Series Processing and Assessment of Vegetation Dynamics. In: C. Kuenzer, S. Dech and W. Wagner (Ed.): Remote Sensing Time Series - Revealing Land Surface Dynamics, Springer International Publishing, Vol. 22, 141-158.

Elvidge, C. D., B. T. Tuttle, P. C. Sutton, K. E. Baugh, A. T. Howard, C. Milesi, B. Bhaduri & R. Nemani (2007) Global Distribution and Density of Constructed Impervious Surfaces. *Sensors,* 7**,** 1962-1979.

Enßle, F., Haeusler, T., Gomez, S., Storch, C., Pape, M., Ott, H. and Ramminger, G. (2016). Bringing Earth Observation Servces for Monitoring Dynamic Forest Disturbances to the Users – EOMonDis Project. Proceedings Book 7th edition of the International Scientific Conference ForestSAT 2016.

Erasmi, S. (2013). Habitat Mapping from Optical and SAR Satellite Data: Implications of Synergy and Uncertainty for Landscape Analysis. *Photogrammetrie - Fernerkundung - Geoinformation*, Vol. 2013, No. 3, 139–148.

Esch, T., V. Himmler, G. Schorcht, M. Thiel, T. Wehrmann, F. Bachofer, C. Conrad, M. Schmidt & S. Dech (2009) Large-area assessment of impervious surface based on integrated analysis of single-date Landsat-7 images and geospatial vector data. *Remote Sensing of Environment,* 113**,** 1678-1690.

Esch, T., Metz, A., Marconcini, M. and Keil, M. (2014a). Combined use of multi-seasonal high and medium resolution satellite imagery for parcel-related mapping of cropland and grassland, *International Journal of Applied Earth Observation and Geoinformation*, Vol. 28, 230–237.

Esch, T., Metz, A., Marconcini, M. and Keil, M. (2014b). Differentiation of crop types and grassland by multi-scale analysis of seasonal satellite data. In: Manakos, I. and Braun, M. (Eds.). *Land Use and Land Cover Mapping in Europe: Practices & Trends*, *Remote Sensing and Digital Image Processing*, 1st ed., Springer, Dordrecht, 329–339.

Esch, T., W. Heldens, A. Hirner, M. Keil, M. Marconcini, A. Roth, J. Zeidler, S. Dech & E. Strano (2017) Breaking new ground in mapping human settlements from space – The Global Urban Footprint. *ISPRS Journal of Photogrammetry and Remote Sensing,* 134**,** 30-42.

Ferrazzoli, P., Paloscia, S., Pampaloni, S., Schiavon, G., Sigismondi, S. and Solimini, D. (1997). The Potential of Multifrequency Polarimetric SAR in Assessing Agricultural and Arboreous Biomass. *IEEE Transactions on Geoscience and Remote Sensing,* Vol. 35, No. 1, 5–17.

Florczyk, A., Ferri, S., Vasileios, S., Kemper, T., Halkia, M., Soille, P., & Pesaresi, M. (2015). A New European Settlement Map From Optical Remotely Sensed Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 1-15.

Foody, G. and Arora, M. (1997). An evaluation of some factors affecting the accuracy of classification by an artificial neural network. *International Journal of Remote Sensing*, 18, 799–810.

Foody, G. M. (2003). Remote sensing of tropical forest environments: towards the monitoring of environmental resources for sustainable development. International Journal of Remote Sensing, 2003.

Franke, J., Keuck, V. and Siegert, F. (2012). Assessment of grassland use intensity by remote sensing to support conservation schemes. *Journal for Nature Conservation*, Vol. 20, No. 3, 125–134.

Fuller, D. O. (2006). Tropical forest monitoring and remote sensing: A new era of transparency in forest governance? Singapore Journal of Tropical Geography, 2006.

Gallego, J. (1995). *Sampling Frames of Square Segments.* Ispra: Office for Publications of the E.C. Luxembourg.

Gallego, J. (2004). Area Frames for Land Cover Estimation: Improving the European LUCAS Survey. *3rd International Conference on Agricultural Statistics.* Mexico.

Gao, B.-C. (1996). NDWI – A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, 58, 257–266.

Gilsason, P., Benediktsson, J., & Sveinsson, J. (2006). Random Forests for land cover classification. *Pattern Recognition Letters, 27*(4), 294-300.

Gong, P., Wang, J., Yu, L., Zhao, Y., Liang, L., Niu, Z., . . . Yu, L. e. (2013). Finer resolution observation and monitoring of global land cover: first mapping results with Landsat TM and ETM+ data. *International Journal of Remote Sensing, 34*(7), 2607-2654.

Gross, J.E., Goetz, S.J. and Cihlar, J. (2009). Application of remote sensing to parks and protected area monitoring: Introduction to the special issue. *Monitoring Protected Areas*, Vol. 113, No. 7, 1343–1345.

Gu, Y., Brown, J.F., Verdin, J.P. and Wardlow, B. (2007). A five-year analysis of MODIS NDVI and NDWI for grassland drought assessment over the central Great Plains of the United States. *Geophysical Research Letters*, Vol. 34, No. 6.

Guo, W., D. Lu & W. Kuang (2017) Improving Fractional Impervious Surface Mapping Performance through Combination of DMSP-OLS and MODIS NDVI Data. *Remote Sensing,* 9.

Hagolle, O., Lobo, A., Maisongrande, P., Cabot, F., Duchemin, B. and Pereyra, A.D. (2004). Quality assessment and improvement of temporally composited products of remotely sensed imagery by combination of VEGETATION 1 and 2. *International Journal of Remote Sensing*, 94, 172-186.

Hagolle, O. and Morin, D. (2015). Design Justification File: benchmarking for L3 monthly composite product. Sen2Agri project, ESA

Hansen, M., Dubayah, R., & DeFries, R. (1996). Classification trees: an alternative to traditional land cover classifiers. *International journal of remote sensing, 17*(5), 1075-1081.

Hansen, M. C., DeFries, R. S., Townshend, J. R. G. and Sohlberg, R. (2000). Global land cover classification at 1 km spatial resolution using a classification tree approach. International Journal of Remote Sensing, 2000.

Hansen, M. C., DeFries, R. S., Townshend, J. R. G., Carroll, M., Dimiceli, C. and Sohlberg, R. A. (2003). Global Percent Tree Cover at a Spatial Resolution of 500 Meters: First Results of the MODIS Vegetation Continuous Fields Algorithm. Earth Interactions, 2003.

Hansen, M. C., Stehman, S. V., Potapov, P. V., Loveland, T. R., Townshend, J. R. G., DeFries, R. S., Pittman, K. W., Arunarwati, B., Stolle, F., Steininger, M. K., Carroll, M. and DiMiceli, C. (2008). Humid tropical forest clearing from 2000 to 2005 quantified by using multitemporal and multiresolution remotely sensed data. Proceedings of the National Academy of Sciences of the United States of America, 2008.

Hansen, M., Roy, D., Lindquist, E., Adusei, B., Justice, C. and Altstatt, Al. (2008). A method for integrating MODIS and Landsat data for systematic monitoring of forest cover and change in the Congo Basin. *Remote Sensing of Environment,* 112, 2495-2513.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kammareddy, A., Egorov, A., Chini, L., Justice, C.O. and Townshend, J. R. G. (2013). High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science, 342*, 850.

Hansen, M., Krylov, A., Tyukavina, A., Potapov, P., Turubanova, S., Zutta, B., Ifo, S., Margono, B., Stolle, F. and Moore, R. (2016). Humid tropical forest disturbance alerts using Landsat data. *Environmental Research Letters*, 11, 034008.

Haralick, R., Shanmugam, K., & Dinstein, I. (1973). Textural features for image classification. *Proceedings of the IEEE, 5*(41), 786-804. Retrieved from http://haralick.org/journals/TexturalFeatures.pdf

Healey, S. P., Cohen, W. B., Zhiqiang, Y. and Krankina, O. N. (2005). Comparison of Tasseled Cap-based Landsat data structures for use in forest disturbance detection. Remote Sensing of Environment, 2005.

Hill, M.J., Vickery, P.J., Furnival, E.P. and Donald, G.E. (1999). Pasture Land Cover in Eastern Australia from NOAA-AVHRR NDVI and Classified Landsat TM. *Remote Sensing of Environment*, Vol. 67, No. 1, 32–50.

Hill, M.J., Smith, A.M. and Foster, T.C. (2000). Remote Sensing of Grassland with RADARSAT; Case Studies from Australia and Canada. *Canadian Journal of Remote Sensing*, Vol. 26, No. 4, 285–296.

Hill, M.J., Ticehurst, C.J., Lee, J.-S., Grunes, M.R., Donald, G.E. and Henry, D. (2005). Integration of Optical and Radar Classifications for Mapping Pasture Type in Western Australia. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 43, No. 7, 1665–1681.

Hirschmugl, M, Gallaun, H., Dees, M., Datta, P., Deutscher, J., Koutsias, N. and Schardt, M. (2017).

Review of methods for mapping forest disturbance and degradation from optical earth observation data. Current Forestry Reports, 2017.

Holben, B. N. (1986). Characteristics of maximum-value composite images from temporal AVHRR data. *International Journal of Remote Sensing*, 7, 1417–1434.

Hong G., Zhang A., Zhou F. and Brisco B. (2014). Integration of optical and synthetic aperture radar (SAR) images to differentiate grassland and alfalfa in Prairie area. In: International Journal of Applied Earth Observation and Geoinformation, Volume 28, Pages 12-19

Huang, X., Zhang, L., & Li, P. (2007, May). Classification and Extraction of Spatial Features inUrban Areas Using High-ResolutionMultispectral Imagery. *IEEE Geoscience and Remote Sensing Letters, 4*(2), 260-264.

Imhoff, M. L., P. Zhang, R. E. Wolfe & L. Bounoua (2010) Remote sensing of the urban heat island effect across biomes in the continental USA. *Remote Sensing of Environment,* 114**,** 504-513.

Immitzer, M.; Vuolo, F.; Atzberger, C. First Experience with Sentinel-2 Data for Crop and Tree Species Classifications in Central Europe. Remote Sens. 2016, 8, 166.

Inglada, J.; Vincent, A.; Arias, M.; Tardy, B.; Morin, D.; Rodes, I. (2017). Operational High Resolution Land Cover Map Production at the Country Scale Using Satellite Image Time Series. Remote Sensing, 9, 95.

Jensen, M.E., Dibenedetto, J.P., Barber, J.A., Montagne, C. and Bourgeron, P.S. (2001). Spatial Modeling of Rangeland Potential Vegetation Environments. *Journal of Range Management*, Vol. 54, No. 5, 528-536.

Jensen, J. (2005). *Introductory digital image processing: A remote sensing perspective* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Kaspersen, P., R. Fensholt & M. Drews (2015) Using Landsat Vegetation Indices to Estimate Impervious Surface Fractions for European Cities. *Remote Sensing,* 7**,** 8224-8249.

Keil, M., Metz, A. and Nieland, S. (2013). Begleitende Arbeiten zur Aktualisierung von CORINE Land Cover 2006 Abschlussbericht. UBA Auftrag Z6-00335 4218, DLR-DFD Oberpfaffenhofen (Internal Report to the German Federal Environment Agency).

Kempeneers, P., Sedano, F., Seebach, L., Strobl, P. and San-Miguel-Ayanz, J. (2011). Data Fusion of Different Spatial Resolution Remote Sensing Images Applied to Forest-Type Mapping. IEEE Transactions on Geoscience and Remote Sensing (49), 2011.

Kemper, T., Mudau, N., Mangara, P., & Pesaresi, M. (2015). *Towards a country-wide mapping & monitoring of formal and informal settlements in South Africa.* Ispra: Publications Office of the European Union.

Kulkarni, A. D. and Lowe, B. (2016). Random Forest Algorithm for Land Cover Classification. Computer Science Faculty Publications and Presentations, 2016.

Lambert, M.-J., Waldner, F. and Defourny, P. (2016). Cropland mapping over Sahelian and Sudanian agrosystems: a knowledge-based approach using PROBA-V time series at 100m. *Remote Sensing*, 8, 232.

Lary, D., Alavi, A., Gandomi, A., & Walker, A. (2015). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, 1-9.

Lefebvre, A., Corpetti, T., & Hubert-Moy, L. (2011a). Estimation of the orientation of textured patterns via wavelet analysis. *Pattern Recognition Letters, 32*(2), 190-196.

Lefebvre, A., Corpetti, T., & Hubert-Moy, L. (2011b). Wavelet and evidence theory for object-oriented classification: Application to change detection in Rennes metropolitan area. *Revue Internationale de Géomatique, 21*(3), 297-325.

Lefebvre, A., Sannier, C., & Corpetti, T. (2016, July). Monitoring Urban Areas with Sentinel-2A Data: Application to the Update of the Copernicus High Resolution Layer Imperviousness Degree. *Remote Sensing, 8*(606), 1-21.

Li, Z., Huffman, T., McConkey, B. and Townley-Smith, L. (2013). Monitoring and modeling spatial and temporal patterns of grassland dynamics using time-series MODIS NDVI with climate and stocking data. *Remote Sensing of Environment*, Vol. 138, 232–244.

Liu, Y., Zha, Y., Gao, J. and Ni, S. (2004), Assessment of grassland degradation near Lake Qinghai, West China, using Landsat TM and in situ reflectance spectra data. *International Journal of Remote Sensing*, Vol. 25, No. 20, 4177–4189.

Liu, Fei Tony, Ting, Kai Ming and Zhou, Zhi-Hua. "Isolation forest." Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on.

Liu, K., H. Su, L. Zhang, H. Yang, R. Zhang & X. Li (2015a) Analysis of the Urban Heat Island Effect in Shijiazhuang, China Using Satellite and Airborne Data. *Remote Sensing,* 7**,** 4804-4833.

Liu, X., G. Hu, B. Ai, X. Li & Q. Shi (2015b) A Normalized Urban Areas Composite Index (NUACI) Based on Combination of DMSP-OLS and MODIS for Mapping Impervious Surface Area. *Remote Sensing,* 7**,** 17168-17189.

Liu, Y., Hill, M.J., Zhang, X., Wang, Z., Richardson, A.D., Hufkens, K., Filippa, G., Baldocchi, D.D., Ma, S., Verfaillie, J., Schaaf, C.B., (2017). Agricultural and Forest Meteorology Using data from Landsat , MODIS , VIIRS and PhenoCams to monitor the phenology of California oak / grass savanna and open grassland across spatial scales. *Agricultural and Forest Meteorology*, 237-238, 311-32.

Liu, C., H. Luo & Y. Yao (2017) Optimizing Subpixel Impervious Surface Area Mapping Through Adaptive Integration of Spectral, Phenological, and Spatial Features. *IEEE Geoscience and Remote Sensing Letters,* 14**,** 1017-1021.

Lopes, M., Fauvel, M., Ouin, A., & Girard, S. (2017). Spectro-Temporal Heterogeneity Measures from Dense High Spatial Resolution Satellite Image Time Series: Application to Grassland Species Diversity Estimation. *Remote Sensing*, Vol. 9, No. 10, 993.

Lopes, M., Fauvel, M., Girard, S., & Sheeren, D. (2017a). Object-based classification of grasslands from high resolution satellite image time series using Gaussian mean map kernels. *Remote Sensing*, 1-25.

Lopez-Sanchez, J.M., Ballester-Berman, J.D. and Hajnsek, I. (2011). First Results of Rice Monitoring Practices in Spain by Means of Time Series of TerraSAR-X Dual-Pol Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* Vol. 4, No. 2, 412–422.

Lu, Y., & Trinder, J. K. (2006). Automatic Building Detection Using the Dempster-Shafer Algorithm. *Photogrammetric Engineering & Remote Sensing, 72*(4), 395-403.

Lu, D., G. Li, W. Kuang & E. Moran (2013) Methods to extract impervious surface areas from satellite images. *International Journal of Digital Earth,* 7**,** 93-112.

Lucas, R., Rowlands, A., Brown, A., Keyworth, S. and Bunting, P. (2007). Rule-based classification of multi-temporal satellite imagery for habitat and agricultural land cover mapping. *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 62, No. 3, 165–185.

Mander, Ü., Mitchley, J., Keramitsoglou, I., Bock, M. and Xofis, P. (2005). Earth observation methods for habitat mapping and spatial indicators for nature conservation in Europe. *Journal for Nature Conservation*, Vol. 13, No. 2-3, 69–73.

Mas, J., & Flores, J. (2008). The application of artificial neural networks to the analysis of remotely sensed data. *Internation Journal of Remote Sensing*, 617-663.

Matton, N., Sepulcre Canto, G., Waldner, F., Valero, S., Morin, D., Inglada, J., Arias, M., Bontemps, S., Koetz, B. and Defourny, P. (2015). An automated method for annual cropland mapping along the season for various globally-distributed agrosystems using high spatial and temporal resolution time series. *Remote Sensing*, 7, 13208-13232.

McInnes, W. S., Smith, B., and McDermid, G. J. (2015). Discriminating Native and Nonnative Grasses in the Dry Mixedgrass Prairie With MODIS NDVI Time Series. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* Vol. 8, No. 4, 1395-1403.

McNairn, H. and Brisco, B. (2004). The application of C-band polarimetric SAR for agriculture: a review. *Canadian Journal of Remote Sensing*, Vol. 30, No. 3, 525–542.

McNairn, H., Champagne, C., Shang, J., Holmstrom, D. and Reichert, G. (2009). Integration of optical and Synthetic Aperture Radar (SAR) imagery for delivering operational annual crop inventories. *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 64, No. 5, 434–449.

Metz, A., (2016). An advanced system for the targeted classification of grassland types with multi-temporal SAR imagery. Doctoral thesis. University of Osnabrueck, Germany.

Miettinen, J., Stibig, H.-J. and Achard, F. (2014). Remote sensing of forest degradation in Southeast Asia—Aiming for a regional view through 5–30 m satellite data. Global Ecology and Conservation, 2014.

Mitchell, A. L., Rosenqvist, A. and Mora, B. (2017). Current remote sensing approaches to monitoring forest degradation in support of countries measurement, reporting and verification (MRV) systems for REDD+. Carbon Balance Manage, 2017.

Mountrakis, G., Im, J., & Ogola, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and remote Sensing, 66*(3), 247-259.

Müller, H., Rufin, P., Griffiths, P., Barros Siqueira, A. J., & Hostert, P. (2015). Mining dense Landsat time series for separating cropland and pasture in a heterogeneous Brazilian savanna landscape. *Remote Sensing of Environment*, Vol. 156, 490-499.

Numata, I., Roberts, D.A., Sawada, Y., Chadwick, O.A., Schimel, J.P. and Soares, J.V. (2007). Regional Characterization of Pasture Changes through Time and Space in Rondônia, Brazil. *Earth Interact*, Vol. 11, No. 14, 1–25.

Pal, M. and Mather, P. (2006). Some issues in the classification of DAIS hyperspectral data. *International Journal of Remote Sensing*, 27, 2895–2916.

Pesaresi, M., & Benediktsson, J. A. (2001, March). A new approach for the morphological segmentation of high-resolution satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing, 39*(2), 309-319.

Pesaresi, M., Gerhardinger, A., & Kayitakire, F. (2008, September). A Robust Built-Up Area Presence Index by Anisotropic Rotation-Invariant Textural Measure. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 1*(3), 180-192.

Pesaresi, M., G. Huadong, X. Blaes, D. Ehrlich, S. Ferri, L. Gueguen, M. Halkia, M. Kauffmann, T. Kemper, L. Lu, M. A. Marin-Herrera, G. K. Ouzounis, M. Scavazzon, P. Soille, V. Syrris & L. Zanchetta (2013) A Global Human Settlement Layer From Optical HR/VHR RS Data: Concept and First Results. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* 6**,** 2102-2131.

Petrou, Z.I., Kosmidou, V., Manakos, I., Stathaki, T., Adamo, M., Tarantino, C., Tomaselli, V., Blonda, P. and Petrou, M. (2014). A rule-based classification methodology to handle uncertainty in habitat mapping employing evidential reasoning and fuzzy logic. *Pattern Recognition Letters*, Vol. 48, 24–33.

Phillips, S.J., Dudík, M. and Schapire, R.E. (2004). A Maximum Entropy Approach to Species Distribution Modelling. Proceedings of the twenty-first international conference on Machine learning, Banff, Alberta, Canada.

Potapov, P. V., Turubanova, S. A., Tyukavina, A., Krylov, A. M., McCarty, J. L., Radeloff, V. C. and Hansen, M. C. (2015). Eastern Europe's forest cover dynamics from 1985 to 2012 quantified from the full Landsat archive. Remote Sensing of Environment, 2015.

Potapov, P. V., Yaroshenko, A., Turubanova, S., Dubinin, M., Laestadius, L., Thies, C., Aksenov, D., Egorov, A., Yesipova, Y., Glushkov, I., Karpachevskiy, M., Kostikova, A., Manisha, A., Tsybikova, E. and Zhuraleva, I. (2008). Mapping the World's Intact Forest Landscapes by Remote Sensing. Ecology and Society, 2008.

Price, K.P., Guo, X. and Stiles, J.M. (2002a). Comparison of Landsat TM and ERS-2 SAR data for discriminating among grassland types and treatments in eastern Kansas. *Computers and Electronics in Agriculture*, Vol. 37, No. 1-3, 157–171.

Price, K.P., Guo, X. and Stiles, J.M. (2002b). Optimal Landsat TM band combinations and vegetation indices for discrimination of six grassland types in eastern Kansas. *International Journal of Remote Sensing*, Vol. 23, No. 23, 5031–5042.

Qi, J. and Kerr, Y. (1995). On current compositing algorithms. *Remote Sensing Reviews*, 15, 235–256.

Radoux, J. & Defourny, P. (2008). Quality assessment of segmentation results devoted to object-based classification. In Blaschke, T., Lang, S. & Hay, G.J. (eds), Object-Based Image Analysis : Spatial concepts for knowledge driven remote sensing applications (pp. 257–271), Springer-Verlag: Berlin – Heidelberg.

Radoux, J. and Defourny, P. (2010). Automated image-to-map discrepancy detection using iterative trimming. *Photogramm. Eng. Remote Sens*., 76, 173–181.

Radoux, J., Lamarche, C., Van Bogaert, E., Bontemps, S., Brockmann, C. and Defourny, P. (2014). Automated training sample extraction for global land cover mapping. *Remote Sensing*, 6, 3965-3987.

Ridd, M. K. (2007) Exploring a V-I-S (vegetation-impervious surface-soil) model for urban ecosystem analysis through remote sensing: comparative anatomy for cities†. *International Journal of Remote Sensing,* 16**,** 2165-2185.

Rodriguez, F., H. Andrieu & F. Morena (2008) A distributed hydrological model for urbanized areas – Model development and application to case studies. *Journal of Hydrology,* 351**,** 268-287.

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

Rodríguez-Maturino, A., Martínez-Guerrero, J., Chairez-Hernández, I., Pereda-Solis, M., Villarreal-Guerrero, F., Renteria-Villalobos, M., & Pinedo-Alvarez, A. (2017). Mapping Land Cover and Estimating the Grassland Structure in a Priority Area of the Chihuahuan Desert. *Land*, Vol. 6, No. 4, 70.

Roy, P. S., Dutt, C. B. S. and Joshi, P. K. (2002). Tropical forest resource assessment and monitoring. Tropical Ecology, 2002.

Rufin, P., Müller, H., Pflugmacher, D. and Hostert, P. (2015). Land use intensity trajectories on Amazonian pastures derived from Landsat time series. *International Journal of Applied Earth Observation and Geoinformation*, Vol. 41, 1–10.

Sanchez-Hernandez, C., Boyd, D.S. and Foody, G.M. (2007). Mapping specific habitats from remotely sensed imagery: Support vector machine and support vector data description based classification of coastal saltmarsh habitats. *Ecological Informatics*, Vol. 2, No. 2, 83–88.

Sarkar Chaudhuri, A., P. Singh & S. C. Rai (2017) Assessment of impervious surface growth in urban environment through remote sensing estimates. *Environmental Earth Sciences,* 76.

Schlager, P., Krismann, A., Wiedmann, K., Hiltscher, H., Hochschild, V. and Schmieder, K. (2013). Multisensoral, object- and GIS-based classification of grassland habitats in the Biosphere Reserve Schwäbische Alb. *Photogrammetrie - Fernerkundung - Geoinformation*, Vol. 2013, No. 3, 163–172.

Schmidt, T., Schuster, C., Kleinschmit, B. and Förster, M. (2014). Evaluating an Intra-Annual Time Series for Grassland Classification - How Many Acquisitions and What Seasonal Origin Are Optimal?. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* Vol. 7, No. 8, 3428–3439.

Schneider, A., M. A. Friedl & D. Potere (2010) Mapping global urban areas using MODIS 500-m data: New methods and datasets based on 'urban ecoregions'. *Remote Sensing of Environment,* 114**,** 1733-1746.

Schölkopf, B., Platt, J., Shawe-Taylor, J., A., S., & Williamson, R. (1999). Estimating the support of a high-dimensional distribution. Neural Computation, 13:7.

Schuster, C., Ali, I., Lohmann, P., Frick, A., Förster, M. and Kleinschmit, B. (2011). Towards Detecting Swath Events in TerraSAR-X Time Series to Establish NATURA 2000 Grassland Habitat Swath Management as Monitoring Parameter. *Remote Sensing*, Vol. 3, No. 7, 1308–1322.

Schuster, C., Schmidt, T., Conrad, C., Kleinschmit, B. and Förster, M. (2015). Grassland habitat mapping by intra-annual time series analysis - Comparison of RapidEye and TerraSAR-X satellite data. *International Journal of Applied Earth Observation and Geoinformation*, Vol. 34, 25–34.

Sezgin, M. and B. Sankur (2004). Survey over image thresholding techniques and quantitative performance evaluation. Journal of Electronic Imaging 13(1), 146–165

Shang, N., & Breiman, L. (1996). Distribution based trees are more accurate. *Ionosphere, 33*(2), 351.

Shimada, M., Itoh, T., Matooka, T., Watanabe, M., Tomohiro, S., Thapa, R. and R. Lucas (2014). New Global Forest/Non-forest Maps from ALOS PALSAR Data (2007-2010). Remote Sensing of Environment, 2014.

Smith, A.M., Major, D.J., McNeil, R.L., Willms, W.D., Brisco, B. and Brown, R.J. (1995). Complementarity of radar and visible-infrared sensors in assessing rangeland condition. *Remote Sensing of Environment*, Vol. 52, No. 3, 173–180.

Smith, A.M. and Buckley, J.R. (2011). Investigating RADARSAT-2 as a tool for monitoring grassland in western Canada. *Canadian Journal of Remote Sensing*, Vol. 37, No. 1, 93–102.

Stehman, S., & Czaplewski, R. (1998). Design and Analysis for Thematic Map Accuracy Assessment: Fundamental Principles. *Remore Sensing of Environment, 62*, 331-334.

Stenzel, S., Feilhauer, H., Mack, B., Metz, A. and Schmidtlein, S. (2014). Remote sensing of scattered Natura 2000 habitats using a one-class classifier. *International Journal of Applied Earth Observation and Geoinformation*, Vol. 33, 211–217.

Stibig, H.-J., Malingreau, J.P. and Beuchle, R. (2001). New possibilities of regional assessment of tropical forest cover in insular Southeast Asia using SPOTVEGETATION satellite image mosaics. *International Journal of Remote Sensing*, 22, 503–505.

Svirejeva-Hopkins, A., H. J. Schellnhuber & V. L. Pomaz (2004) Urbanised territories as a specific component of the Global Carbon Cycle. *Ecological Modelling,* 173**,** 295-312.

Svoray, T. and Shoshany, M. (2003). Herbaceous biomass retrieval in habitats of complex composition: a model merging SAR images with unmixed Landsat TM data. *IEEE Transactions on Geoscience and Remote Sensing,* Vol. 41, No. 7, 1592–1601.

Tamm, T., Zalite, K., Voormansik, K., & Talgre, L. (2016). Relating Sentinel-1 Interferometric Coherence to Mowing Events on Grasslands. *Remote Sensing*, Vol. 8, No. 10, 802

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining.* Pearson.

Tasumi, M., Hirakawa, K., Hasegawa, N., Nishiwaki, A. and Kimura, R. (2014). Application of MODIS Land Products to Assessment of Land Degradation of Alpine Rangeland in Northern India with Limited Ground-Based Information. *Remote Sensing*, Vol. 6, No. 10, 9260-9276.

Tax, M.J.D. and P. W. Duin (2004). Support Vector Data Description, Machine Learning, 54, 45–66.

Tax, M.J.D. (2001): One-class classification. PhD thesis, Delft University of Technology.

Thoonen, G., Spanhove, T., Haest, B., Borre, J.V. and Scheunders, P. (2010). Habitat mapping and quality assessment of heathlands using a modified kernel-based reclassification technique. 2010 IEEE International Geoscience and Remote Sensing Symposium, 25-30 July 2010, Honolulu, HI, USA, 2707-2710.

Tsutsumida, N., A. Comber, K. Barrett, I. Saizen & E. Rustiadi (2016) Sub-Pixel Classification of MODIS EVI for Annual Mappings of Impervious Surface Areas. *Remote Sensing,* 8.

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

Tucker, C.J. (1980). Remote sensing of leaf water content in the near infrared. *Remote sensing of Environment*, 10(1), 2-32.

Tuia, D., Ratle, F., Pacifici, F., Kanevski, M., & Emery, W. (2009, July). Active Learning Methods for Remote Sensing Image Classification. *IEEE Transactions on geoscience and Remote Sensing, 47*(7), 2218-2232.

Tuia, D., Volpi, M., Copa, L., Kanesvski, M., & J., M.-M. (2011, July). A survey of active learning algorithms for supervised remote sensing image classification. *IEEE Journal of Selected Topics in Signal Processing, 5*(3), 606-617.

Turner, W., Spector, S., Gardiner, N., Fladeland, M., Sterling, E. and Steininger, M. (2003). Remote sensing for biodiversity science and conservation. *Trends in Ecology & Evolution*, Vol. 18, No. 6, 306–314.

Vanden Borre, J., Paelinckx, D., Mücher, C.A., Kooistra, L., Haest, B., Blust, G. de and Schmidt, A.M. (2011). Integrating remote sensing in Natura 2000 habitat monitoring: Prospects on the way forward. *Journal for Nature Conservation*, Vol. 19 No. 2, 116–125.

Van de Voorde, T., W. Jacquet & F. Canters (2011) Mapping form and function in urban areas: An approach based on urban metrics and continuous impervious surface data. *Landscape and Urban Planning,* 102**,** 143-155.

Vancutsem, C., Pekel, J;-F., Bogaert, P. and Defourny, P. (2007a). Mean compositing, an alternative strategy for producing temporal syntheses. Concepts and performances assessment for SPOT-VEGETATION time series. *International Journal of Remote Sensing*, 28, 22, 5123-5141.

Vancutsem, C., Bicheron, P. Cayrol, P. and Defourny, P. (2007b). An assessment of three candidate compositing methods for global MERIS time series. *Canadian Journal of Remote Sensing*, 33, 6, 492-502.

Vancutsem, C. and Defourny, P. (2009). A decision support tool for the optimization of compositing parameters. *International Journal of Remote Sensing*, 1, 41-56.

Viovy, N., Arino, O. and Belward, A.S. (1992). The Best Index Slope Extraction (BISE): A method for reducing noise in NDVI time series. *International Journal of Remote Sensing*, 13, 1585–1590.

Waldner, F., Sepulcre Canto, G. and Defourny, P. (2015). Automated annual cropland mapping using knowledge-based temporal features. *ISPRS Journal of Photogrammetry and Remote Sensing*, 110, 1-13.

Waldner, F., De Abelleyra, D., Veron, S.R., Zhang, M., Wu, B., Plotnikov, D., Bartalev, S., Lavreniuk, M., Skakun, S., Kussul, N., Le Maire, G., Dupuy, S., Jarvis, I, & Defourny, P. (2016). Towards a set of agrosystems specific cropland mapping methods to address the global cropland diversity. *International Journal of Remote Sensing*, 37(14): 3196–3231.

Waldner, F., Hansen, M., Potapov, P. V., Low, F., Newby, T., Ferreira, S. and Defourny, P. (2017). National-scale cropland mapping based on spectral-temporal features and outdated land cover information. *PLoS ONE*, 12(8), e181911.

Wan, Z., Wang, P. and Li, X. (2004). Using MODIS Land Surface Temperature and Normalized Difference Vegetation Index products for monitoring drought in the southern Great Plains, USA. *International Journal of Remote Sensing*, Vol. 25, No. 1, 61–72.

Wang, C., Hunt, E. R., Zhang, L., & Guo, H. (2013). Phenology-assisted classification of C3 and C4 grasses in the U.S. Great Plains and their climate dependency with MODIS time series. *Remote Sensing of Environment*, Vol. 138, 90-101.

Wang, J., Xiao, X., Qin, Y., Dong, J., Geissler, G., Zhang, G., Cejda, N., Alikhani, B., & Doughty, R. B. (2017). Mapping the dynamics of eastern redcedar encroachment into grasslands during 1984–2010 through PALSAR and time series Landsat images. *Remote Sensing of Environment*, Vol. 190, 233-246.

Waske, B. and Benediktsson, J.A. (2007). Fusion of Support Vector Machines for Classification of Multisensor Data. *IEEE Transactions on Geoscience and Remote Sensing,* Vol. 45, No. 12, 3858–3866.

Waske, B. and van der Linden, S. (2008). Classifying Multilevel Imagery from SAR and Optical Sensors by Decision Fusion, *IEEE Transactions on Geoscience and Remote Sensing,* Vol. 46, No. 5, 1457–1466.

Wegmüller, U. and Werner, C. (1997). Retrieval of Vegetation Parameters with SAR Interferometry. *IEEE Transactions on Geoscience and Remote Sensing,* Vol. 35, No. 1, 18–24.

Weng, Q., P. Gamba, G. Mountrakis, M. Pesaresi, L. Lu, T. Kemper, J. Heinzel, G. Xian, H. Jin, H. Miyazaki, B. Xu, S. Quresh, I. Keramitsoglou, Y. Ban, T. Esch, A. Roth & C. Elvidge. 2014. Urban Observing Sensors. In *Global Urban Monitoring and Assessment through Earth Observation*, 49-80.

Wood, E.M., Pidgeon, A.M., Radeloff, V.C. and Keuler, N.S. (2012). Image texture as a remotely sensed measure of vegetation structure. *Remote Sensing of Environment*, Vol. 121, 516–526.

Yang, X., Smith, A. M. and Hill, M.J. (2017). Updating the Grassland Vegetation Inventory Using Change Vector Analysis and Functionally-Based Vegetation Indices. *Canadian Journal of Remote Sensing*, Vol. 43, 62-78.

Yu, L., Zhou, L., Liu, W. and Zhou, H.-K. (2010). Using Remote Sensing and GIS Technologies to Estimate Grass Yield and Livestock Carrying Capacity of Alpine Grasslands in Golog Prefecture, China. *Pedosphere*, Vol. 20, No. 3, 342–351.

Yu, L., Wang, J., & Gong, P. (2013). Improving 30m global land-over map FROM-GLC with time series MODIS and auxiliary datasets: a segmentation based approach. *International Journal of Remote Sensing, 34*, 5851-5867.

Yu, L., Wang, J., Li, X., Li, C., Zhao, Y., & Gong, P. (2014). A multi-resolution global land cover dataset through multisource data aggregation. *Science China Earth Sciences, 57*, 2317-2329.

Yuan, F. & M. E. Bauer (2007) Comparison of impervious surface area and normalized difference vegetation index as indicators of surface urban heat island effects in Landsat imagery. *Remote Sensing of Environment,* 106**,** 375-386.

Zha, Y. and Gao, J. (2011). Quantitative detection of change in grass cover from multi-temporal TM satellite data. *International Journal of Remote Sensing*, Vol. 32, No. 5, 1289–1302.

Zhang, L., Zhu, X., Zhang, L., & Du, B. (2016). Multidomain Subspace Classification for Hyperspectral Images. *IEEE Transactions on Geoscience and Remote Sensing, 54*(10), 1-13.

Zhao, F., Xu, B., Yang, X., Jin, Y., Li, J., Xia, L., Chen, S. and Ma, H. (2014). Remote Sensing Estimates of Grassland Aboveground Biomass Based on MODIS Net Primary Productivity (NPP): A Case Study in the Xilingol Grassland of Northern China. *Remote Sensing*, Vol. 6, No. 6, 5368-5386.

Zheng B, Campbell J.B. and de Beurs K. M. (2011). Remote sensing of crop residue cover using multi-temporal Landsat imagery. – In: Remote Sensing of Environment,Volume 117,2012,Pages 177-183

Zhou, Y., Y. Wang, A. J. Gold & P. V. August (2010) Modeling watershed rainfall–runoff relations using impervious surface-area data with high spatial resolution. *Hydrogeology Journal,* 18**,** 1413-1423.

Zhu, Z. and Woodcock, C. E. (2014). Continuous change detection and classification of land cover using all available Landsat data. Remote Sensing of Environment, 2014.

Zhu, Z. (2017). Change detection using landsat time series: A review of frequencies, preprocessing, algorithms, and applications. ISPRS Journal of Photogrammetry and Remote Sensing, 130, 2017.

Zillmann, E., Gonzalez, A., Montero Herrero, Enrique J., van Wolvelaer, J., Esch, T., Keil, M., Weichelt, H. and Garzon, A.M. (2014). Pan-European Grassland Mapping Using Seasonal Statistics From Multisensor Image Time Series. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* Vol. 7, No. 8, 3461–3472.