

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



D8.2

“D33.1b – Time Series Analysis for Thematic Classification (Issue 2)”

Issue/Rev.: 2.0

Date Issued: 18.12.2019

submitted by:



in collaboration with the consortium partners:



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:

The contents of this document are the copyright of GAF AG and Partners. It is released by GAF AG on the condition that it will not be copied in whole, in section or otherwise reproduced (whether by photographic, reprographic or any other method) and that the contents thereof shall not be divulged to any other person other than of the addressed (save to the other authorised officers of their organisation having a need to know such contents, for the purpose of which disclosure is made by GAF AG) without prior consent of GAF AG. Nevertheless, single or common copyrights of the contributing partners remain unconditionally unaffected.

DOCUMENT RELEASE SHEET

	NAME, FUNCTION	DATE	SIGNATURE
Author(s):	Pierre Defourny (UCL) Ines Moreau (UCL) Jolan Wolter (UCL) Élie Khalil (UCL) Heinz Gallaun (JR) Petra Miletich (JR) Martin Puhm (JR) Sophie Villerot (SIRS) Alexandre Pennec (SIRS) Alice Lhernould (SIRS) Yoann Courmont (SIRS) Tatiana De-Wyse (SIRS) David Herrmann (GAF) Kathrin Schweitz (GAF) Benjamin Leutner (DLR)	20.10.2019	
Review:	David Herrmann (GAF) Alice Lhernould (SIRS) Antoine Masse (SIRS) Sophie Villerot (SIRS)	10.11.2019	
Approval:	Eva Sevillano Marco (GAF)	18.12.2019	
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
2.0	18.12.2019	317	Second issue of the Methods Compendium: Time series Analysis for Thematic Classification. Major edits in all sections with final outcomes from phase 2 Implementation. New section 2.4 Accuracy Assessment Principles, 3.2 Indicators and Variables

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	D21.1b - Service Evolution Requirements Report, Issue 2.0, Issued: December 2019
AD06	D31.1 - Methods Compendium: Sentinel-1/2/3 Integration Strategies, Issue 1.0, Issued: December 2019
AD07	D32.1b- Methods Compendium: Time Series Preparation, Issue 2.0, Issued: 15.05.2019
AD08	D34.1b - Methods Compendium: Time Series Analysis for Change Detection, Issue 2.0, Issued: December 2019
AD09	D35.1b – Methods Compendium: HRL Time Series Consistency for HRL Product (incremental) Updates, Issue 2.0, Issued: December 2019
AD10	D41.1b – Prototype Report: Time Series-derived Indicators and Variables, Issue 2.0, Issued: December 2019
AD11	D42.1b – Prototype Report: Consistent HR Layer Time Series/Incremental Updates, Issue 2.0, Issued: December 2019
AD12	D43.1b – Prototype Report: Improved Permanent Grasslands, Issue 2.0, Issued: December 2019
AD13	D44.1b – Prototype Report: Crop Area and Crop Status/Parameters Products, Issue 2.0, Issued: December 2019
AD14	D45.1b – Prototype Report: New LC/LU Products, Issue 2.0, Issued: December 2019
AD15	Technical Note LUCAS 2018. Issued: 05.07.2019

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 is being conducted from 2017–2019 and aims at developing and prototypically demonstrating selected innovative products and methods as candidates for future next-generation operational Copernicus Land Monitoring Service (CLMS) products of the pan-European and Global Components. ECoLaSS assesses the operational readiness of such candidate products and eventually suggests some of these for implementation. This shall enable the key CLMS stakeholders (i.e. mainly the Entrusted European Entities (EEE) EEA and JRC) to take informed decisions on potential procurement as (part of) the next generation of Copernicus Land services from 2020 onwards.

To achieve this goal, ECoLaSS makes full use of dense time series of High-Resolution (HR) Sentinel-2 optical and Sentinel-1 Synthetic Aperture Radar (SAR) data, complemented by Medium-Resolution (MR) Sentinel-3 optical data if needed and feasible. Rapidly evolving scientific developments as well as user requirements are continuously analysed in a close stakeholder interaction process, targeting a future pan-European roll-out of new/improved CLMS products, and assessing the potential transferability 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 is 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, Imperviousness and Forest, Crop Mask and Crop type and new LC/LU products.

Section 1 of the document presents 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 these 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. The first issue of this deliverable presented preliminary results. In the second 18-month project cycle, the second issue of this deliverable is published, containing all relevant updates and final results after completion of all technical developments activities within Task 3 and Task 4 WPs.

TABLE OF CONTENTS

1	INTRODUCTION	1
1.1	PURPOSE AND OBJECTIVES OF THE WP	1
1.2	DOCUMENT STRUCTURE	3
2	REVIEW (THEORY/STATE OF THE ART)	4
2.1	AUTOMATED REFERENCE SAMPLING	4
2.2	OPTICAL IMAGE COMPOSITING	5
2.2.1	Time interval algorithms.....	6
2.2.2	Feature-based algorithms	8
2.3	TIME SERIES CLASSIFICATION METHODS.....	10
2.3.1	HRL Imperviousness	10
2.3.2	HRL Forest.....	12
2.3.2.1	Forest state of the art.....	12
2.3.2.2	HRL Forest production.....	15
2.3.3	HRL Grassland.....	21
2.3.3.1	Grassland state of the art	21
2.3.3.2	HRL Grassland production	26
2.3.3.3	Mapping Mediterranean Grassland with Multi-temporal Earth Observation Data desk study	31
2.3.4	Agriculture	44
2.3.4.1	The concept of land cover for cropland mapping	45
2.3.4.2	Image processing and cropland map production	47
2.3.5	New land cover products.....	49
2.3.5.1	Previous attempts	49
2.3.5.2	CORINE Land Cover.....	50
2.3.5.3	Current state-of-the-art.....	51
2.3.5.4	Toward CLC+	53
2.4	ACCURACY ASSESSMENT PRINCIPLES	56
3	METHODS.....	60
3.1	INPUT DATA	60
3.1.1	Automated reference sampling.....	61
3.1.1.1	Description of candidate methods	61
3.1.1.2	Benchmarking criteria	62
3.1.1.3	Implementation of benchmarking.....	62
3.1.1.4	Results of benchmarking	63
3.1.1.5	Summary and conclusions.....	68
3.1.2	Compositing methods on S-2 time series and PROBA-V compositing	68
3.1.2.1	Description of candidate methods	69
3.1.2.2	Benchmarking criteria	70
3.1.2.3	Implementation and results of benchmarking.....	72
3.1.2.4	PROBA-V Compositing.....	85
3.1.2.5	Summary and conclusions	86
3.1.3	Indices.....	87
3.1.4	Time Features	88
3.1.4.1	(Preliminary) Set of Implemented Features	88
3.1.4.2	Feature Selection.....	90
3.2	INDICATORS AND VARIABLES	93
3.2.1	Method for generic LC metrics.....	93
3.2.2	Method for crop growth condition	94

3.2.3	Method for multiannual trends and potential changes based on SAR data from S-1	94
3.2.3.1	Pre-processing on the input data	95
3.2.3.2	Statistical analysis of seasonal and annual metrics within the HRL classes to identify potential change.....	97
3.2.4	Methods for emergence date detection	99
3.2.4.1	Emergence date as phenological parameter.....	99
3.2.4.2	Vis and hue time series as candidate data sources.....	99
3.2.4.3	Candidate detection methods.....	100
3.2.4.4	Performance indicators	104
3.3	TIME SERIES CLASSIFICATION METHODS.....	104
3.3.1	Imperviousness.....	104
3.3.1.1	Description of candidate methods	104
3.3.1.2	Benchmarking criteria	112
3.3.1.3	Implementation and results of benchmarking.....	114
3.3.1.4	Experimental Setup for phase 2	124
3.3.1.5	Classification Results and Validation	137
3.3.1.6	Summary and conclusions	141
3.3.2	Forest.....	143
3.3.2.1	Description of candidate methods	143
3.3.2.2	Benchmarking criteria	144
3.3.2.3	Implementation and results of benchmarking.....	144
3.3.2.4	Summary and conclusions	165
3.3.3	Grassland	167
3.3.3.1	Description of candidate methods	170
3.3.3.2	Benchmarking criteria	177
3.3.3.3	Implementation and Results of Benchmarking	179
3.3.3.4	Summary and conclusions	212
3.3.4	Agriculture	213
3.3.4.1	Central test site – Germany.....	214
3.3.4.2	Belgium site	260
3.3.4.3	African site	263
3.3.4.4	Summary and conclusions	265
3.3.5	New land cover products.....	268
3.3.5.1	Description of candidate methods	272
3.3.5.2	Benchmarking criteria	284
3.3.5.3	Implementation and results of benchmarking.....	284
3.3.5.4	Summary and conclusions	300
4	CONCLUSIONS AND OUTLOOK.....	301
	REFERENCES.....	304

List of Figures

Figure 1-1: ECoLaSS Test- and Demonstration- Sites in Europe.....	2
Figure 1-2: ECoLaSS test sites in Africa.....	2
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)	8
Figure 2-2: Ambiguity of forest classification systems: Canopy cover and tree height as minimum physical requirements of a forest. Source: Comber et al. (2005)	14
Figure 2-3: Number of Scenes for HRL Forest Tree Cover and Dominant Leaf Type Mapping 2015.	19
Figure 2-4: 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.....	20
Figure 2-5: Summary of the total amount of images used for the production of the GRA 2015 mask, per year and satellite.	29
Figure 2-6: Final Grassland layer in Central Europe (green) and PLOUGH, indicating the number of years since the last ploughing activity in orange/red shades.	30
Figure 2-7: 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.	30
Figure 2-8: Mediterranean climatic map after Köppen & Geiger (Peel et al. 2007).	33
Figure 2-9: Biogeographic regions in Europe 2011 (EEA 2012).	35
Figure 2-10: Dominant land cover in Europe and the Mediterranean region (red boundaries) (EU 2017 and EUROSTAT 2013)	39
Figure 2-11: Share of utilized agricultural areas (UUA) in different land uses at NUTS 2 level, 2010 (EU2017).	41
Figure 2-12: Workflow for cropland mapping from satellite observation time series. (Dashed lines correspond to alternative pathways).	44
Figure 2-13: Overview if the 2017 cover map produced by the CESBio (Source: http://osr-cesbio.upstlse.fr/~oso/).	52
Figure 2-14: - Figure extracted from the ITT EEA/DIS/R0/19/012 - caption can be read as: " CLC with a 1 km raster/grid superimposed (top) illustrating the difference between encoding a particular unit as raster pixel (centre) or a grid cell (bottom). "daa" is a Norwegian unit: 10 daa = 1 ha."	54
Figure 2-15: Simple random (left) and random systematic (right) sampling designs	56
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.	63
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.	64
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.	66
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.	67
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.	73
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.	74

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.	75
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 (e) Minimum NDVI.	76
Figure 3-9: False colour (b8, b3, b2) quantile compositing features over the Belgium site: (a) Quantile 10 and (b) Quantile 90.	77
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 (e) Minimum NDVI.	78
Figure 3-11: False colour (b8, b3, b2) quantile compositing features over the Mali site: (a) Quantile 10 and (b) Quantile 90.	79
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).	80
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.	81
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.	82
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.	83
Figure 3-16: Average percentage of data gaps remaining in the composites for the Belgium site.	84
Figure 3-17: Artefacts in the Red and NIR bands for the five selected algorithms for (a) the Belgium site and (b)	85
Figure 3-18: PROBA-V 100 Composite from a time series acquired the 1 st to the 15 th July 2018. The mean compositing was applied on the Collection 1 cloud flag recently reprocessed. The white pixels in the zoom to Antwerpen (Belgium) on the right corresponds to features permanently flagged as cloud.	86
Figure 3-19: PROBA-V cloud free image acquired on the 2 nd July 2018 over Belgium and The Netherlands (left image). The corresponding cloud and cloud shadow flags as detected by the Collection 1 algorithm.	86
Figure 3-20: 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.	89
Figure 3-21: Concept of the calculation of a complex time feature shown for the dif_max time feature.	90
Figure 3-22: Classification workflow.	92
Figure 3-23: Sentinel-2 tiles for Belgium test site and harmonized grassland HRL of 2015	95
Figure 3-24: Examples of temporal S-1 metrics: Backscattering temporal statistics of 2015 of the S-1 relative orbit 161 in ascending pass over the BELGIUM demo site: (a) backscattering temporal mean, (b) backscattering temporal standard deviation, (c) backscattering temporal minimum, (d) backscattering temporal maximum (outlines of S-2 granules in yellow)	97
Figure 3-25: Workflow for identification of potential pixels for update within a certain HRL	98
Figure 3-26: Four methods based on the NDVI to detect start and end of the season. a) fixed threshold, b) derivative, c) Fourier transform, d) quadratic fitting based on AGDD (de Beurs and Henebry, 2010)	101
Figure 3-27: Overview of the six emergence estimation methods. Methods without parameterization: (a) inflection point, (b) base logistic (c) maximum value. Methods with parameterization: (d) highest slope, (e) absolute threshold, (f) relative threshold. The threshold methods are both tested on the linear and logistic interpolated observations. Presented vegetation profiles are examples from the maize calibration sample	103
Figure 3-28: Validation samples overlaid on the HRL IMD 2015, reference map.	113
Figure 3-29: Subset of Imperviousness Layer compared with Sentinel-2 imagery.	123
Figure 3-30: 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.	124

Figure 3-31: 2018 NDVI Sentinel-2 based maximum feature for the year 2018.....	128
Figure 3-32: SAR statistical features (NB: the 4 features have different value ranges and scaling)	129
Figure 3-33: Sentinel-2 based initial sealed and non-sealed mask for 2018 for the test sites	131
Figure 3-34: Final HRL Imperviousness 2018 layers for the test sites.....	133
Figure 3-35: 2018 PanTex Sentinel-2 based feature for the Central test site	134
Figure 3-36: Sentinel-2 based initial built-up and non-built-up masks for 2018 for the test sites	135
Figure 3-37: Final HRL Built-up 2018 layers for the test sites	137
Figure 3-38: Scatterplots for the continuous IMD layers 2018 validation	138
Figure 3-39: Comparison of Built-up Layers based on VHR vs HR data compared with VHR imagery (Toulouse, France).....	142
Figure 3-40: Cloud coverage of Sentinel-2 tile VVF (left) and VWF (right) of the test site North in Sweden. Blue: Scenes with < 50% cloud cover.	145
Figure 3-41: 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.....	145
Figure 3-42: Example for the Sample Layer 2015 derived from the Copernicus High-Resolution Layers 2015. (<i>© EuroGeographics for the administrative boundaries</i>).....	147
Figure 3-43: Forest class separability box plots for selected Sentinel-2 time features.	149
Figure 3-44: Forest class separability box plots for selected Sentinel-1 time features.	149
Figure 3-45: NDVI profiles 2017/2018 for broadleaf and coniferous tree stands in the North test site. .	149
Figure 3-46: Sentinel-2 SWIR Band (B12) reflectance characteristics for broadleaf and coniferous tree stands.	150
Figure 3-47: NDVI profiles 2017/2018 for broadleaf and coniferous tree stands in the Central test site.	150
Figure 3-48: NDVI profiles 2017/2018 for broadleaf and coniferous tree stands in the South-East test site.	151
Figure 3-49: Kappa and overall accuracy for the five DLT input data configurations.	151
Figure 3-50: 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). <i>Modified Copernicus Sentinel data [2017]</i>	153
Figure 3-51: Top 20 time feature ranking for combined S1/S2 TCM 2018 (left) and DLT 2018 classification (right).....	154
Figure 3-52: Number of time features versus TCM 2018 accuracy performance for test site Central.	155
Figure 3-53: Parameter set evaluation for Tree Cover Mapping in the Central test site.....	155
Figure 3-54: Parameter set evaluation for Dominant Leaf Type Mapping in the Central test site.....	156
Figure 3-55: Sentinel-2 median time feature stacks for Tree Cover Density classification. Blue areas represent remaining clouds, nodata gaps and/or snow and ice cover. <i>Modified Copernicus Sentinel data [2018]</i>	160
Figure 3-56: Comparison of Tree Cover Density classification results based on different input data. <i>Modified Copernicus Sentinel data [2018]</i>	162
Figure 3-57: Sentinel-2 median time feature stack (01.06.-31.08.2018) and unmasked Tree Cover Density. Arrows point to issues caused by the terrain correction and cloud masking. <i>Modified Copernicus Sentinel data [2018]</i>	163
Figure 3-58: Comparison of the TCD 2015 (20 m) and the improved TCD 2018 at 10 m spatial resolution. <i>Produced using modified Copernicus Sentinel data [2016/2018]</i>	163
Figure 3-59: Grassland mapping workflow overview. Sentinel-2 (S2); Sentinel-1 (S1); Land Use and Land Cover Survey (LUCAS).....	169
Figure 3-60: Multispectral time series of a single grassland pixel, one mowing event according to INVEKOS.....	172
Figure 3-61: Variables estimated by the Kalman filter from the observations plotted in Figure 3-60.	173
Figure 3-62: Multispectral time series of a single grassland pixel, two mowing events according to INVEKOS.....	173
Figure 3-63: Variables estimated by the Kalman filter from the observations plotted in Figure 3-62.	174
Figure 3-64: Multispectral time series of a single grassland pixel, three mowing events according to INVEKOS.....	175

Figure 3-65: Variables estimated by the Kalman filter from the observations plotted in Figure 3-64.	175
Figure 3-66: Intensive/Extensive grassland use layer workflow.	176
Figure 3-67: Grassland mowing intensity in Central 2018.	177
Figure 3-68: Detailed view of the grassland mowing intensity layer compared to Bing Maps Aerial of the same region, showing more intense management is concentrated in valleys around human settlements.	177
Figure 3-69: InSar Coherence 6-days (March-October).	181
Figure 3-70: Feature importance S1 coherence, S1 backscatter time features and S2 time features (reflectance + indices)	182
Figure 3-71: Example of a Tasseled Cap Brightness time series of a grassland pixel and fitted regression models using OLS and IRLS.	184
Figure 3-72: Computed observation weights of the IRLS fit.	184
Figure 3-73: Example of Greenness time series of a grassland pixel and fitted regression models using OLS.	185
Figure 3-74: Comparison of an agricultural pixel to a grassland pixel.	185
Figure 3-75: Greenness trend and amplitudes of an agriculture and grassland pixel.	186
Figure 3-76: Regression function second order (R: A1 G: A2 B: P1).	187
Figure 3-77: Regression function third order (R: A1 G: A2 B: A3).	187
Figure 3-78: SAR grassland threshold-based classification for 2017 (grassland in yellow).	188
Figure 3-79: LGP grassland areas in red. Basis layer: ArcGIS Basemap.	188
Figure 3-80: OBB error in relation the number of S1 input features.	190
Figure 3-81: Top 10 S1 features for the grassland discrimination.	190
Figure 3-82: SAR grassland classification (grassland in yellow) with random forest and selected S1 features for 2017 (p>60%).	192
Figure 3-83: LGP grassland areas 2016 in green mapped on the World View 1 image from the 27. 06. 2018.	192
Figure 3-84: OBB error in relation the number of S2 input features.	192
Figure 3-85: Top 40 S2 features for the grassland discrimination.	193
Figure 3-86: Optical grassland classification with random forest and selected features for 2017 (p>60%). (grassland in yellow)	194
Figure 3-87: LGP grassland areas 2016 in green mapped on the World View 1 image from the 27. 06. 2018.	194
Figure 3-88: OBB error in relation the number of S1 and S2 input features.	195
Figure 3-89: Top 40 S1 and S2 features for the grassland discrimination.	195
Figure 3-90: SAR + OPT grassland classification with random forest and selected features for 2017 (p>50%). (grassland in yellow)	197
Figure 3-91: LGP grassland areas 2016 in green mapped on the World View 1 image from the 27. 06. 2018.	197
Figure 3-92: SAR + OPT classification 2017 10m aggregated: Omission errors (green). Classification (yellow) vs. LGP polygons 2016 (green) mapped on the World View 1 image from the 27. 06. 2018.	198
Figure 3-93: SAR + OPT classification 2017 10m aggregated: Commission errors (in yellow). Classification (yellow) vs. LGP polygons 2016 (green) mapped on the World View 1 image from the 27. 06. 2018.	198
Figure 3-94: World View 1 image from the 05.10.2018.	198
Figure 3-95: S2 based classification mapped on the World View 1 image from the 05.10.2018.	198
Figure 3-96: S1/S2 based classification mapped on the World View 1 image from the 05.10.2018.	198
Figure 3-97: Comparison of multisensor grassland classifications: S1, S2 and combined in Central test site	200
Figure 3-98: Omission and commission errors in the 2018 Grassland mask	201
Figure 3-99: Probability layer for the grassland classification 2018 in the Central test site	201
Figure 3-100: Grassland use intensity in Central 2018.	203
Figure 3-101: Location of the test site in western Austria (Background © basemap.at)	204
Figure 3-102: Mowing intensity map based on Tasseled Cap tracking (within INVEKOS grassland mask)	205
Figure 3-103: Agreement between the INVEKOS reference layer and the mowing intensity map.	206

Figure 3-104: Detail analysis of an area with poor agreement between reference and map. Inspection of the available observations for sample pixels A and B (see Figure 3-105) indicates problems caused by data gaps.	207
Figure 3-105: NIR time series of sample pixels A and B (see Figure 3-104). Two key observations required to detect mowing events are missing in series A, but not in B.	207
Figure 3-106: Analysis of an area with moderate agreement between reference and map. Inspection of the estimated state variables for sample pixels C and D (see Figure 3-107 and Figure 3-108) suggests that the reference could be wrong in this case.	208
Figure 3-107: Estimated state variables of sample pixel C (see Figure 3-106). Greenness and Wetness patterns indicate three mowing events, which is in agreement with the reference.	208
Figure 3-108: Estimated state variables of sample pixel D (see Figure 3-106). Greenness and Wetness patterns indicate three mowing events similar to sample pixel C. However, the reference states extensive usage.	209
Figure 3-109: Development of out-of-bag grassland F1 score during Gini-based backward feature selection for the predictor sets: Sentinel-1 (S1), Sentinel-2 (S2) and combined features (S1/S2).	210
Figure 3-110: Phase 1 - Sentinel-2 data score (inverted cloud value count) for ECoLaSS central test site (T32UNU+32UNV tiles) for the time period March-Nov 2017.	216
Figure 3-111: Phase 2 - Sentinel-2 data score (inverted cloud value count) for ECoLaSS central test site (left: whole demo site, right: test site T32UNU/TNT tiles) for the time period Mid-March – Mid-Oct 2018.	216
Figure 3-112: Phase 1 - Imagery used in phase 1: monthly data availability for the two test tiles of Sentinel-1 (left) and Sentinel-2 (right) with cloud cover <50%.	217
Figure 3-113: Phase 2 - Available imagery for S-1 and S-2: number of scenes per time window for the test site (32UNU + 32TNT) with cloud cover up to 90% (S-2 data) and minimum tile coverage of 20% (S-1).	218
Figure 3-114: Phase 1 - 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.	219
Figure 3-115: Phase 2- Exemplary selected time features from the Mid-March - Mid-Oct 2018 period (brightness 90th percentile, NDVI mean, NDWI 75th percentile) and an RGB composite of 3 different periods.	222
Figure 3-116: Phase 1- Frequency (left) and mean parcel size (right) of the reference samples used for crop type classification.	224
Figure 3-117: Phase 2 - available LPIS data for the 2 test tiles 32UNU and 32TNT for 2018.	228
Figure 3-118: Number of parcels per crop Type and total area per Crop Type in ha, indicating the average parcel size as well as the degree of representation within the test site.	232
Figure 3-119: Phase 2 - Maximum parcel size per Crop Type in ha; classes marked in red have been left out due to small parcel sizes, the green ones have been kept – they offer distinct spectral information and can be well classified.	233
Figure 3-120: Phase 1 - Barplot of Kappa (K) and Overall Accuracy (OA) for the different experiment setups (Sentinel-1, Sentinel-2, Sentinel-1 & Sentinel-2 on field and pixel level).	234
Figure 3-121: Phase 2 – Crop Mask: Overall Accuracy (OA) and Kappa Coefficient (K) for the different experiment setups for Sentinel-1, Sentinel-2, Sentinel-1 & Sentinel-2 (pixel level).	235
Figure 3-122: Phase 1 - Barplot of Kappa (K) and Overall Accuracy (OA) for the different experiment setups.	237
Figure 3-123. Phase 1 - Class-wise F1-Score (mean of User's and Producer's Accuracy) for field vs. pixel-based classifications, reference year 2017.	238
Figure 3-124: Confusion Matrix of the crop type classification on field level based on the combination of Sentinel-1 and Sentinel-2 time features.	239
Figure 3-125: Barplot of Kappa (K) and Overall Accuracy (OA) for the different experiment setups.	240
Figure 3-126: Overall accuracy (OA) based on the cross-validated training samples dependent on the number of selected features.	240
Figure 3-127: Barplot of Kappa (K) and Overall Accuracy (OA) for the classification based on all features, and the 50 selected features.	241

Figure 3-128: Kappa and Overall Accuracy on field level of the for the different experiment setups, particularly the three considered periods.....	241
Figure 3-129: Phase 1 - Class-wise field-level F1-Scores (mean of User's and Producer's Accuracy) for the different experiment setups, particularly the three considered periods.	242
Figure 3-130: Phase 1 - Distributions of the breaking ties and entropy reliabilities of wrong (blue, left) and correct (orange, right) predictions grouped by the predicted crop type.....	243
Figure 3-131: 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 and RGB composite of th three reliability layers maximum, breaking ties and entropy.	244
Figure 3-132: Phase 1 Details - Top left: detailed view of the crop type 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.....	245
Figure 3-133: Phase 2 - F1 Score for all Crop Type classes: S1, S2 and combined approach S1 & S2 and the improvement of accuracies by using both sensors.....	246
Figure 3-134: Phase 2 - Maximum size of parcels and F1-Score per Crop Type	247
Figure 3-135: Phase 2 - Connection between percentage of area covered by each Crop Type and accuracy within the classification.....	248
Figure 3-136: Phase 2 - Confusion matrix for Crop Type classification at Pixel level (PA indicated down left, UA down right).....	249
Figure 3-137: Phase 2 - Overall accuracy OA based on the cross-validated training samples dependent on the number of selected features for the Crop Type classification. The curve describes the saturation process where the slope is less steep than usual indicating that a higher number of time features is necessary for getting sufficient accuracies.	250
Figure 3-138: Phase 2 - Overall accuracy OA based on the cross-validated training samples dependent on the number of selected features for the Crop Mask classification.....	250
Figure 3-139: Phase 2 - Crop Mask for test site tiles 32TNT and 32UNU (left) and location of the test site within the border region of Germany, Switzerland and Austria (right)	252
Figure 3-140: Phase 2 - Detail of crop mask 2018 complemented by the grassland mask 2018.....	253
Figure 3-141: Phase 2 - surroundings by google Earth imagery in 2018: Laupheim, SW of Stuttgart, Baden-Wurttemberg.....	253
Figure 3-142: Phase 2 - Part of the Crop Mask and Details for test site Central.....	254
Figure 3-143: Phase 2 - Crop Type Mask for test site in tiles 32TNT and 32UNU (left) and location of the test site within the border region of Germany, Switzerland and Austria (right)	256
Figure 3-144: Phase 2: Part of the Crop Type Mask 2018 and Details for test site Central 32UNU and 32TNT	257
Figure 3-145: Phase 2 - Detail of Crop Type Mask with RGB of NDVI median TF for 3 time windows (March-May, may-July, July, Oct)	258
Figure 3-146: : Phase 2 - accuracies for the Crop Mask 2018 for the test site per experimental setup, S1 only, S2 only and the combination of S1 & S2	259
Figure 3-147: accuracies for the Crop Type Mask 2018 for the test site per experimental setup, Sentinel-1 only, Sentinel-2 only, and the combination of Sentinel-1 & Sentinel-2.....	260
Figure 3-148: Overall accuracy for every classification scenario evaluated.	262
Figure 3-149: 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).	262
Figure 3-150: 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).	263
Figure 3-151: Location of the three benchmarking tiles for the South-African sites in the Western Cape province.....	264
Figure 3-152: Accuracy assessment of the crop type mapping in the South African site (3 tiles).	264

Figure 3-153: Number of parcels per crop Type and parcel area per Crop Type.....	265
Figure 3-154: Proposed automated approach in the ITT for CLC+ Backbone (EEA, 2019).....	269
Figure 3-155: In red: OSM roads and railways; in dark blue: EU-Hydro; in light blue: OSM water classes.	273
Figure 3-156 – Hard bones superposed over the S-2 tiles delineation (in blue) for the Central test site	285
Figure 3-157 - Hard bones superposed over the S-2 tiles for the South-West tests site.....	285
Figure 3-158: Maximum values of all images over the year 2018 (tile T32TNT). On the lower right, lower values in dark grey are created by the gaps in swath trajectories. Dark grey squares come from the cloud detection algorithm of MAJA, while the light grey trace in the left corner is produced by undetected atmospheric veil.	286
Figure 3-159: Mean values of a selection of the best scenes over the same area from the previous figure. Cloudy and snowy images have been removed from this restrained time series.	287
Figure 3-160: LSMSS as a vector layer with the best selected parameters drapped over the RGB bands (S- 2 bands 2, 3 and 4) of the mean value raster of the selection of best scenes over the T32TNT.....	288
Figure 3-161: For comparison, the LSMSS as a vector layer drapped over an ESRI Imagery VHR.....	288
Figure 3-162: Watershed segmentation as a vector layer with the best selected parameters drapped over the RGB bands (S-2 bands 2, 3 and 4) of the mean value raster of the selection of best scenes over the T32TNT.	290
Figure 3-163: For comparison, the watershed segmentation as a vector layer drapped over the same ESRI Imagery VHR.....	290
Figure 3-164: The separation between the two tiles generated by the watershed algorithm is in the middle of the image. On the left, over-segmentation of agricultural fields can be seen, while on the right, individual fields are quite well separated from other LC.	291
Figure 3-165: - SLIC segmentation as a vector layer with the best selected parameters drapped over the RGB bands (S-2 bands 2, 3 and 4) of the mean value raster of the selection of best scenes over the T32TNT.	292
Figure 3-166: For comparison, the SLIC segmentation as a vector layer drapped over the same ESRI Imagery VHR.....	292
Figure 3-167: “Phenological segmentation” as a vector layer with the best selected parameters drapped over the RGB bands (S-2 bands 2, 3 and 4) of the mean value raster of the selection of best scenes over the T32TNT.	293
Figure 3-168: For comparison, the “phenological segmentation” as a vector layer drapped over the same ESRI Imagery VHR.	294
Figure 3-169: Random Forest classification over T31TCJ and T30TYP	296
Figure 3-170: Raster classification from random forest algorithm over T32TNT and T32UNU	298

List of Tables

Table 2-1: LC/LU features to be included/excluded from the tree cover mask.....	17
Table 2-2: Definition of Grassland according to the HRL Grassland 2015	27
Table 2-3: Strengths and weaknesses of algorithms used for large-area classification of satellite image data (based on Gómez <i>et al.</i> , 2016).	48
Table 2-4: Key elements regarding the evolution of CLC through the years	50
Table 3-1. Length of time series per site.	72
Table 3-2. Tests and compositing periods for the composite benchmarking achieved on the five compositing methods.....	72
Table 3-3: Time features calculated for various bands and indices.	89
Table 3-4: Tests related to the Dempster-Shafer fusion algorithm choice.	115
Table 3-5: Tests related to the classification algorithm selection.....	115
Table 3-6. Selection of the best input dataset based on the results given by various classifications.	116
Table 3-7: Selection of the best sensor dataset based on the results given by SVM.....	116
Table 3-8: Visual check for the Detspster-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.	117
Table 3-9: User and producer accuracy for the diverse Dempster-Shafer algorithms.	118
Table 3-10: 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.....	119
Table 3-11: Full dataset of images for the yearly time series with all spectral bands results	120
Table 3-12: DLR Settlement Extent and Growth Classifier	120
Table 3-13: Subset dataset (36 best images) with all spectral bands results.....	120
Table 3-14: Visual check for different input datasets – the full dataset input gives the best result compared to the HRL IMD layer for 2015.	121
Table 3-15: Overall results for the selection of the proper input data	122
Table 3-16: Impact of the sensor used for the SVM classification	123
Table 3-17.....	124
Table 3-18: Used Sentinel-2 reflectance bands.....	125
Table 3-19: SAR annual statistical features.....	129
Table 3-20: Confusion matrix of the internal validation of the IMD 2018 in test site South-West (area-weighted).....	139
Table 3-21: Confusion matrix of the internal validation of the IMD 2018 in test site Central (area-weighted).....	139
Table 3-22: Confusion matrix of the internal validation of the IMD 2018 in test site South-East (area-weighted).....	140
Table 3-23: Confusion matrix of the internal validation of the BU 2018 in test site South-West (area-weighted).....	140
Table 3-24: Confusion matrix of the internal validation of the BU 2018 in test site Central (area-weighted)	141
Table 3-25: Confusion matrix of the internal validation of the BU 2018 in test site South-East (area-weighted).....	141
Table 3-26: Comparison of Built-up Layers based on VHR vs HR data compared	142
Table 3-27: Sentinel data scenarios and time periods for Forest classification.....	143
Table 3-28: Validation dataset specifications.....	146
Table 3-29: Sample distribution of training and validation dataset.....	147
Table 3-30: Accuracy metrics for the five DLT input data configurations.....	152
Table 3-31: User and producer accuracy for the five DLT input data configurations.	152
Table 3-32: Error matrix for the improved TCM 2018 of the test site Sweden.....	156
Table 3-33: Error matrix for the improved TCM 2018 of the test site Austria/Germany	157
Table 3-34: Error matrix for the improved TCM 2018 of the test site Bulgaria/Greece	157
Table 3-35: Error matrix for the improved DLT 2018 status layer of the test site Sweden	157

Table 3-36: Error matrix for the improved DLT 2018 status layer of the test site Austria/Germany	158
Table 3-37: Error matrix for the improved DLT 2018 status layer of the test site Bulgaria/Greece	158
Table 3-38: Parameter testing for time series TCD classification.....	159
Table 3-39: Benchmarking criteria, chances, and issues of the different input data scenarios	165
Table 3-40: Reference data comparison LPG2016 vs VIRP2016.	180
Table 3-41: LUCAS 2018 inclusion rules.	180
Table 3-42: Confusion matrix using VIRP-LUCAS points and the threshold based S1 classification for 2017.	188
Table 3-43: Confusion matrix using LPG2016 points and the threshold based S1 classification for 2017.	188
Table 3-44: SAR threshold based grassland classification confusions.	189
Table 3-45: Count based accuracy metrics (in %) for random forest based classification for 2017 using S1 features.	191
Table 3-46: Count based accuracy metrics (in %) for random forest based classification for 2018 using S1 features.	191
Table 3-47: Count based accuracy metrics (in %) for random forest based classification for 2017 using S2 features.	193
Table 3-48: Count based accuracy metrics (in %) for random forest based classification for 2018 using S2 features.	194
Table 3-49: Count based accuracy metrics (in %) for random forest based classification for 2017 using S1 and S2 features.....	196
Table 3-50: Count based accuracy metrics (in %) for random forest based classification for 2018 using S1 and S2 features.....	196
Table 3-51: Count based accuracy metrics (in %) for random forest based classification for 2017 using only S1 and 10m S2 features.	197
Table 3-52: Count based accuracy metrics (in %) for random forest based classification for 2018 using S1 and 10m S2 features.....	197
Table 3-53: Thematic accuracy (in %) comparison of different features.	199
Table 3-54: INVEKOS grassland classes and associated area within the test site	205
Table 3-55: Confusion matrix for the mowing intensity map	206
Table 3-56: Important variables identified for Grassland status layer production in demonstration site South-East in both years, 2017 and 2018 for the combination of S1+S2 features. Temporal statistics: q10 = 10% percentile, q50 = 50% percentile (median), q90 = 90% percentile, sd = standard deviation, cv = coefficient of variation. TC green = tasselled cap greenness component, TC wet = tasselled cap wetness component.	211
Table 3-57: Test-set count based accuracy metrics (in %) for random forest based classification for 2017 using S1, S2, and S1+S2 features.	212
Table 3-58: Test-set count based accuracy metrics (in %) for random forest based classification for 2018 using S1, S2, and S1+S2 features.	212
Table 3-59: Phase 1 - Number of Sentinel-2 (< 50% Cloud cover) and Sentinel-1 scenes for the period October 2016 - December 2017.	215
Table 3-60: Phase 2 - Number of Sentinel-2 (<90% Cloud cover) and Sentinel_1 scenes for the growing period Mid-March 2018 to Mid-October 2019	215
Table 3-61: Phase 1 - Overview of the number of features for the different sensor and period combinations in phase 1.....	219
Table 3-62: Phase 1 - Comparison of the number of features when excluding and including the October and November 2016 data.....	220
Table 3-63: Phase 1 - Number of features available for specific time period data scenarios.....	220
Table 3-64: Phase 2 - Overview over calculated features per band and index for Sentinel-1	221
Table 3-65: Phase 2 - Overview over calculated features per band and index for Sentinel-2 in phase 2.	221
Table 3-66: Phase 2 - Total number of time features per sensor and per time period	222
Table 3-67: Phase 1 - Overview of the type and number of reference parcels used for crop type classification. Crop code and crop name derived from LPIS	224

Table 3-68: Phase 1 - Overview of the reference samples used for the crop mask classification.	225
Table 3-69: Phase 2 – Overview of the LUCAS classes that were used as sample base for classes grassland, forest, imperviousness and water bodies.	226
Table 3-70: Phase 2 - Overview of the reference samples used for the crop mask classification	227
Table 3-71: Phase 2 - ECoLaSS crop type nomenclature in phase 2 revealing the hierarchical structure to be customized to regional or local characteristics.	230
Table 3-72: Phase 1 - Kappa Coefficient (K) and Overall Accuracy (OA) for the different crop mask experiment setups (Sentinel-1, Sentinel-2, and Sentinel-1 & Sentinel-2 on pixel and field level).	234
Table 3-73: Phase 2 – Crop Mask: Kappa Coefficient (K) and Overall Accuracy (OA) for the different experiment setups (Sentinel-1, Sentinel-2, and Sentinel-1 & Sentinel-2 (pixel level).	235
Table 3-74: Phase 1 - Benchmarking criteria and specific problems of the different crop mask experiment setups.	236
Table 3-75: Phase 2 - Benchmarking criteria and specific problems of the different experiment setups.	237
Table 3-76: Phase 1 - Kappa Coefficient (K) and Overall Accuracy (OA) for the different experiment setups (Sentinel-1, Sentinel-2, and Sentinel-1 & Sentinel-2 on pixel and field level).	237
Table 3-77: Phase 2 - accuracies for Crop Mask compared to Crop Types at Pixel level and with different sensor experiments; strengths and weaknesses of all three experimental set ups	245
Table 3-78: Number of Time Features selected by the grouped Forward Feature Selection per sensor and per time window	251
Table 3-79: Phase 2 - accuracies for the Crop Mask 2018 for the test site per experimental setup, S1 only, S2 only and the combination of S1 & S2	259
Table 3-80:- Technical details for intermediate and final products created for the CLC+ Backbone.	270
Table 3-81: - Characteristics of classical segmentation algorithms, available in open-source libraries. ..	275
Table 3-82: - Targeted typology, used typology over the test sites and matching with LUCAS, CLC and other ancillary datasets	278
Table 3-83: - List of rules to populate the vector layer created by the fusion of hard and soft bones, using the raster classification.	283
Table 3-84: - Parameters tested for the LSMS segmentations.	287
Table 3-85: Parameters tested for the classical watershed segmentation.....	289
Table 3-86: Segmentation testing parameters.....	291
Table 3-87: - Benchmarking for the segmentation algorithms.	294
Table 3-88: - Available points in the LUCAS dataset from 2018, made available for ECoLaSS over the 31TCJ and 30TYP Sentinel-2 tiles.....	295
Table 3-89: Automatically generated confusion matrix for the test site in the South-West.....	297
Table 3-90: - Available points in the LUCAS dataset from 2018, made available for ECoLaSS over the T32TNT and T32UNU tiles.	298
Table 3-91: Automatically generated confusion matrix for the test site in the Central test site.	299

Abbreviations

AGDD	Accumulated Growing Degree Days
AL	Active Learning
ALOS	Advanced Land Observing Satellite
ANNs	Artificial Neural Networks
ANOVA	Analysis of Variance
AOI	Area Of Interest
AUC	Area Under the ROC Curve
AVG	Average
AVHRR	Advanced Very High Resolution Radiometer
AWiFS	Advanced Wide Field Sensor
BISE	Best Index Slope Extraction
BDC	Bi-Directional Compositing
BOA	Bottom Of the Atmosphere
BPA	Best Available Pixel
BRDF	Bidirectional Reflectance Distribution Function
BU	Built Up
IBU	Imperviousness Built Up
CAP	Common Agricultural Policy
CART	Classification and Regression Tree
CESBIO	Centre d'Études Spatiales de la BIOSphère
CI	Confidence Interval
CIGreen	Chlorophyll Index Green
Clrededge	Chlorophyll Index Red Edge
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
DG	Dormancy of Green Vegetation
DIAS	Data and Information Access Services
DLT	Dominant Leaf Type
DLR	Deutschen Zentrums für Luft- und Raumfahrt
DMP	Differential Morphological Profile
DMSP-OLS	Defense Meteorological Satellite Program's Operational Line-scan System
DST	Dempster-Shafer Theory
DWH	Data Warehouse
EAGLE	EIONET Action Group on Land Monitoring in Europe
EC	European Commission
ECoLaSS	Evolution of Copernicus Land Services based on Sentinel data
EEA	European Environment Agency
EEE	Entrusted European Entities
EG	End of Greenness
EIONET	Environment Information and Observation Network
EO	Earth Observation
ERS	European Remote-Sensing Satellite

ESA	European Space Agency
ESM	European Settlement Map
ESRI	Environmental Systems Research Institute
ETM+	Enhanced Thematic Mapper Plus
EU	European Union
EUROSTAT	European Statistics
EVI	Enhanced Vegetation Index
FAO	Food and Agriculture Organization
fAPAR	Fraction of Absorbed Photosynthetically Active Radiation
FN	False Negative
FOR	Forest
FROM-GLC	Finer Resolution Observation and Monitoring of Global Land Cover
GAI	Green Area Index
GC	Ground Cover
GHSL	Global Human Settlement Layer
GLC	Global Land Cover
GLCM	Grey Level Co-occurrence Matrix
GIO	GMES Initial Operations
GMES	Global Monitoring for Environment and Security
GRA	Grassland
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
HSI	Human Settlement Index
HSV	Hue Saturation Value
IACS	Integrated Administration and Control System
ID	Identifier
iForest	Isolation Forest
IMD	Imperviousness Density
IMP	Imperviousness
INSPIRE	Infrastructure for Spatial Information in Europe
IRECI	Inverted Red Edge Chlorophyll Index
IRS	Indian Remote-Sensing Satellite
IRSL	Iteratively Weighted Least Squares
ISA	Impervious Surface Area
ISO	International Organization for Standardization
ITT	Invitation To Tender
iTree	Isolation Trees
IW	Interferometric Wide Swath Mode
JAXA	Japan Aerospace Exploration Agency
JECAM	Joint Experiment for Crop Assessment and Monitoring network
JRC	Joint Research Centre
KC	Knowledge-Based Compositing
LAI	Leaf Area Index
LC	Land cover
LCCS	Land Cover Classification System
LCML	Land Cover Meta-Language
LGP	Landbouwgebruiksperscelen ALV

LiDAR	Light Detection And Ranging
LISS	Linear Imaging Self-Scanning Sensor
LPIS	Land-Parcel Identification System
LSMS	Large-Scale Mean Shift
LU	Land Use
LUCAS	Land Use/Cover Area frame statistical Survey
MACCS	Multi-Sensor Atmospheric Correction and Cloud Screening
MAD	Median Absolute Deviation from the median
MAJA	Maccs-Atcor Joint Algorithm
MAP	Mapped Value for the Product
MASD	Mean Absolute Spectral Dynamic
MaxEnt	Maximum Entropy
MC	Mean Compositing
MERIS	Medium Resolution Imaging Spectrometer
MG	Maturity of Green canopy
MGRS	Military Grid Reference System
MIR	Medium InfraRed
MMA	Maximal Monthly Activity
MMU	Minimum Mapping Unit
MMW	Minimum Mapping Width
MNDWI	Modified Normalized Difference Water Index
MODIS	Moderate Resolution Imaging Spectroradiometer
MR	Medium Resolution
MSAVI	Modified Soil-Adjusted Vegetation Index
MTV2	Modified Triangular Vegetation Index II
MVC	Maximum Value Composite
NBR	Normalized Burn Ratio
NDBI	Normalized Difference Built-Up Index
NDII	Normalized Difference Infrared Index
NDMI	Normalized Difference Moisture Index
NDSVI	Normalized Difference Senescent Vegetation Index
NDVI	Normalized Difference Vegetation Index
NDRE	Normalized Difference Red Edge Index
NDVVHH	Normalized Difference VV/VH Ratio
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
OCSVM	One-Class Support Vector Machine
OG	Onset of Greenup
OHM	Object Height Model
OLI	Operational Land Imager
OLS	Ordinary Least Squares
OOB	Out Of Bag
OSM	Open Street Map
PA	Producer Accuracy
PALSAR	Phased Array type L-band Synthetic Aperture Radar
PCA	Principal Components Analysis

PIS	Percentage of impervious surface
PLS	Phenological Length of Season
PPS	Phenological Peak of Season
PROBA-V	Project for On-Board Autonomy - Vegetation
PSRI	Plant Senescence Reflectance Index
PSS	Phenological Start of Season
PSU	Primary Sampling Units
QC	Quantile Compositing
QS	Quick Shift
RBF	Radial Basis Function
REDD	Reducing Emissions from Deforestation and forest Degradation
RF	Random Forest
RFE	Rainfall Estimates
RGB	Red Green Blue
RGR	Red-Green Ratio
RMSE	Root Mean Square Error
ROC	Receiver Operation Characteristic
ROI	Region Of Interest
RSG	Remote Sensing Software Graz
RTM	Radiative Transfer Models
S-1	Sentinel-1
S-2	Sentinel-2
S2GLC	Sentinel-2 Global Land Cover
S-3	Sentinel-3
SAR	Synthetic Aperture Radar
SAVI	Soil Adjusted Vegetation Index
SD	Standard Deviation
SE	Shannon Entropy
SEN2COR	Sentinel-2 Atmospheric Correction
SEN4CAP	Sentinels for Common Agriculture Policy
SENSAGRI	Sentinels Synergy for Agriculture
SEOM	Scientific Exploitation of Operational Missions
SFS	Structural Features Set
SIGEC	Système intégré de gestion et de contrôles
SLC	Single Look Complex
SLIC	Simple Linear Iterative Clustering
SMOTE	Synthetic Minority Over-Sampling Technique
SPOT	Satellite Pour l'Observation de la Terre/Satellite for observation of Earth
SRTM	Shuttle Radar Topography Mission
SSU	Secondary Sampling Unit
SVDD	Support Vector Data Description
SVM	Support Vector Machine
SWIR	Short Wavelength Infrared
TCB	Tasseled Cap Brightness
TCG	Tasseled Cap Greenness
TCD	Tree Cover Density
TCM	Tree Cover Mask
TM	Thematic Mapper
TN	True Negatives
TOA	Top Of the Atmosphere
TP	True Positives
UA	User Accuracy
UK	United Kingdom

USA	United States of America
USGS	United States Geological Survey
UTM	Universal Transverse Mercator
V-I-S	Vegetation-Impervious-Soil
VANUI	Vegetation Adjusted Nighttime Light Urban Index
VHR	Very High Resolution
VH	Vertical transmit/Horizontal receive (polarization)
VIIRS	Visible Infrared Imaging Radiometer Suite
VIRP	Visual Interoperation Reference Plots
VV	Vertical transmit/Vertical receive (polarization)
VZA	View Zenith Angle
WAC	Weighted Average Compositing
WGS	World Geodetic System
WI	Wetness Index
WP	Work Package
WU	Working Unit
WV	World View

- EC approval pending -

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 is being 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 makes full use of dense Sentinel time series of High-Resolution (HR) Sentinel-2 optical and Sentinel-1 Synthetic Aperture Radar (SAR) data, complemented by Medium-Resolution (MR) Sentinel-3 optical data if needed and feasible. Rapidly evolving scientific developments as well as user requirements are continuously analyzed in a close stakeholder interaction process, targeting a future pan-European roll-out of new/improved CLMS products, and assessing the potential transferability 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 supports 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 figures below show respectively the distribution of the tests sites within the larger demo sites in Europe (Figure 1-1) and Africa (Figure 1-2).

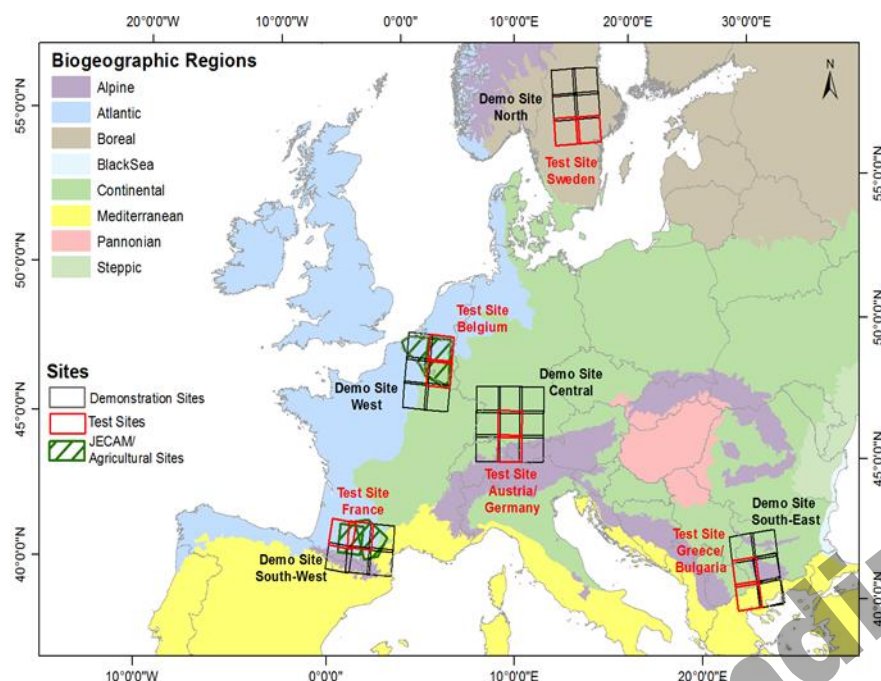


Figure 1-1: ECoLaSS Test- and Demonstration- Sites in Europe

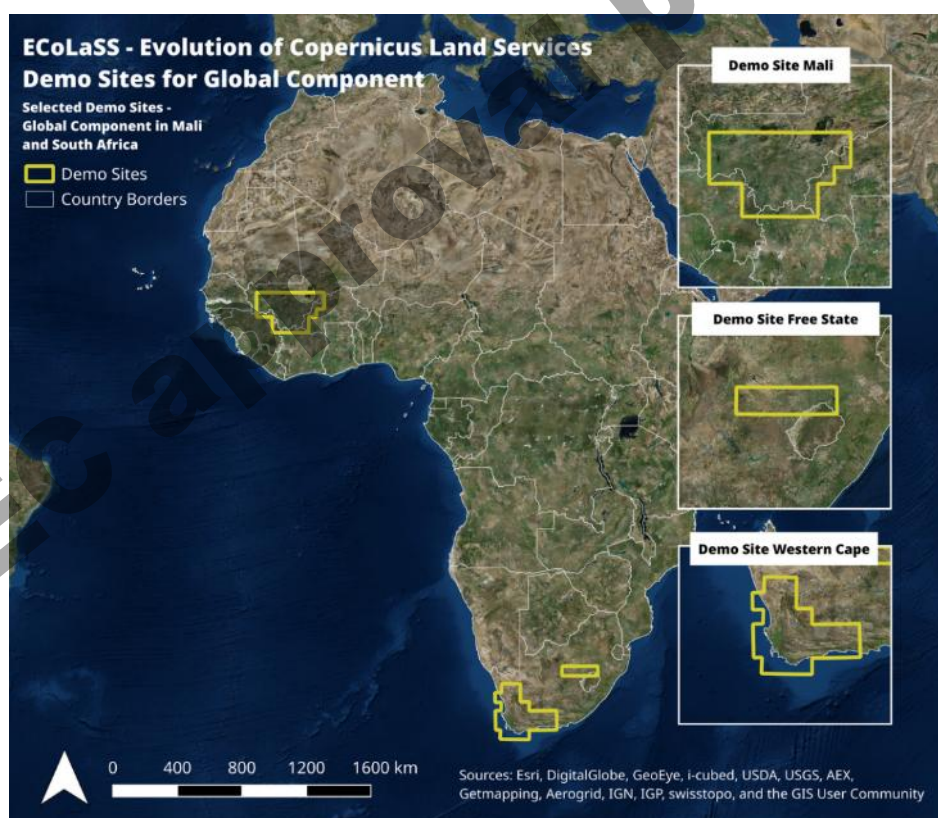


Figure 1-2: ECoLaSS test sites in Africa

The ECoLaSS project follows a two-phased approach of two times 18 months duration. The first issue of this deliverable presented preliminary results. In the second 18-month project cycle, a second issue of this deliverable is 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.

- EC approval pending -

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 propose 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 mislabelled pixels, well-labelled 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 ν 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 and 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'lorio 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 repetitiveness 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) (Viomy 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 Vancutsem 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 to 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 lead 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.

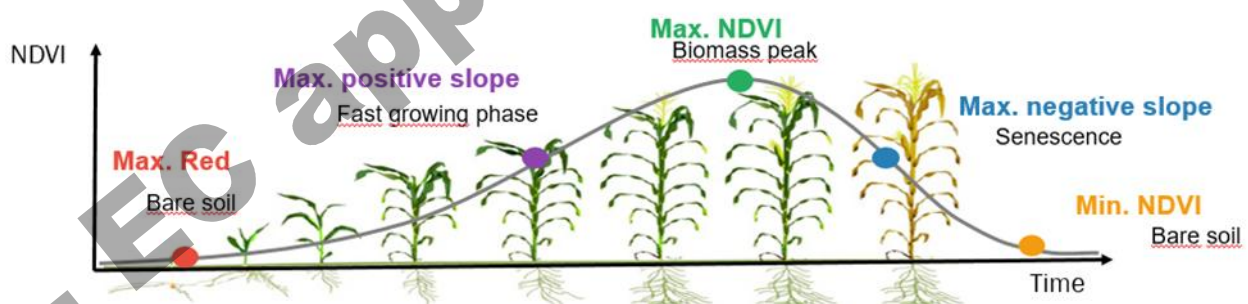


Figure 2-1: Representation of five temporal features of the Knowledge-based Compositing for cropland mapping (minimum NDVI, maximum NDVI, increasing slope, decreasing slope and maximum RED)

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 interquartile 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, Rodríguez, 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 night-time 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 night-time 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 Night-time 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 night-time 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 night-time light data along with MODIS Enhanced Vegetation Index (EVI) as well as MODIS NDWI data and is designed to overcome the limitations of night-time 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, Löffelholz 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 state of the art, gives a general overview of the currently available forest maps on regional to global scales. Afterwards, the production of the HRL Forest is described, including the product definitions and methodology.

2.3.2.1 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., 2013; 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. As explained in WP 34 [AD08], 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, followed by integrating various classifications applying a rule-based approach.

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). The described approach is also applied in the ECOLaSS prototypic mapping and the HRL 2018 FOR production.

Until recently, 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 was included, except for Antarctica and some Arctic islands. 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 analyzed 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 at a spatial resolution of 25 m, based on SAR data and 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 May 2019, DLR released a Global Forest/Non-Forest Map at a spatial resolution of 50 m derived from TanDEM-X bistatic in-terferometric synthetic aperture radar (InSAR) data, optimizing algorithms for different types of forest based on tree height, density and structure (Martone et al., 2018). For the derivation of the final forest mask, a number of predictors are combined, such as the calibrated amplitude, height information and bistatic coherence, i.e. the degree of decorrelation due to multiple scattering within a canopy volume. This product provides for the first time a homogeneous overview of cloud-prone rainforest in South America, South East Asia and Africa.

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).

A land cover map on pan-European scale at 10m spatial resolution and based on Sentinel-2 has been recently produced in framework of Sentinel-2 Global Land Cover (S2GLC) project (<http://s2glc.cbk.waw.pl/>) as part of ESA's Scientific Exploitation of Operational Missions (SEOM) element. The land cover map with reference year 2017 provides up to 13 land cover classes (including broadleaf tree cover and coniferous tree cover) and uses CORINE Land Cover 2012 and the High Resolution Layers 2015 as input training data. The pixel-based map has been produced fully automatic in a C-DIAS Cloud infrastructure (CREODIAS) with multi-temporal features derived from more than 16,000 Sentinel-2 scenes using a Random Forest classifier (Lewiński S. et al, 2019).

On a regional scale Potapov et al. (2015) focus on the former Eastern bloc countries and analyzed 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 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 analysed 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 conceptualization 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 (see Figure 2-2). The legend definition is in truth a key element to be taken into account when comparing different land cover products. Even when addressing the same theme (e.g. forest), nomenclature, specifications and data models are associated to the specific purpose of the corresponding products and may vary significantly. This makes a direct comparison of specific forest/land cover maps challenging.

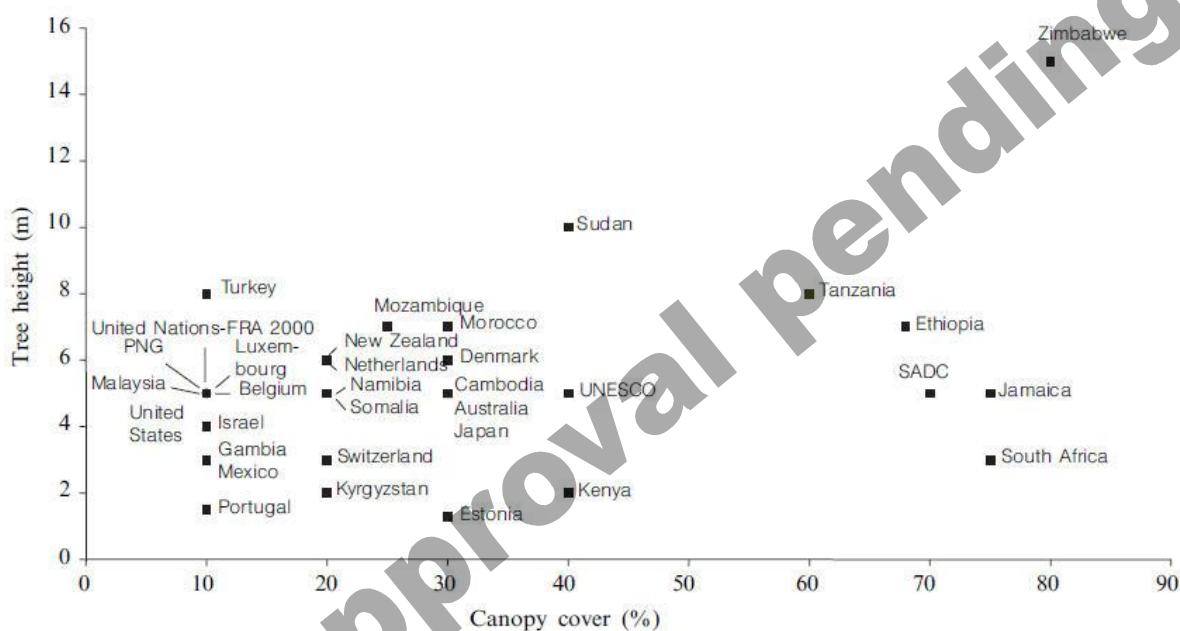


Figure 2-2: Ambiguity of forest classification systems: Canopy cover and tree height as minimum physical requirements of a forest. Source: Comber et al. (2005)

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 25 m maximum. The HRL Forest 2015 (based on Sentinel-2) has been produced with 20m resolution, but the upcoming HRL Forest 2018 is to be produced in 10 m 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 exploit 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 has now reached a quite good standard. 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. The challenge now is to set-up suitable processing chains and infrastructure environments to handle the ever-growing volume of EO mass data in an efficient manner. In this context, processing costs (storage costs & processing costs) will play a major role within operational land cover mapping activities from continental to global scale.

Actually, in ECoLaSS the Forest prototypes production has proved the feasibility of accurate tree cover mapping based on Sentinel time series. The three Forest demonstration sites cover different biogeographic regions (i.e., North, Central and South-East), where the diversity of data availability and forest types is addressed, as will be explained in chapter 3. The production of accurate tree cover masks is key to approach in turn change detection and the incremental updates of High Resolution Layers in forest monitoring assessments [AD08, AD09].

2.3.2.2 HRL Forest production

This subsection is dedicated to HRL Forest production specifically because of the relevance of the topic in the context of ECoLaSS, aimed at improving and producing new products for the Copernicus portfolio. 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 20 m spatial resolution using optical Earth Observation (EO) data for certain reference years in a 3-years update cycle. From 2018 onwards, the HRL Forest will provide status information in 10 m spatial resolution using a combined optical and SAR classification approach

The HRL Forest 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 exceeded the contractually specified 3 years considerably.

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

- 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. The HRL Forest 2015 has been almost produced in the specified timeframe and with an overall high quality.

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.

The ongoing implementation of the HRL Forest with reference year 2018 at 10 m spatial resolution strongly benefits from a further increased data situation thanks to the availability of Sentinel-2B and the efforts of the Copernicus In-situ component. In addition, and thanks to the findings from ECoLaSS, the integration of Sentinel-1 SAR data is foreseen in cloud-prone regions. Thus, production will fully rely on Sentinel time series data.

2.3.2.2.1 HRL Forest Product Definitions

In the following, a brief overview on the HRL Forest product definitions will be given. A detailed HRL Forest 2015 Product Specifications Document is available for download at the CLMS website under <https://land.copernicus.eu/user-corner/technical-library/hrl-forest>. Specifications of the upcoming HRL Forest 2018 products at 10 m spatial resolution will be added to the technical library as soon as the HRL products are published. However, specifications of the two 10m primary status layers Dominant Leaf Type and Tree Cover Density are identical to the ones used in ECoLaSS.

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 20/10 m input satellite data), resulting in a binary Tree Cover Mask (TCM). The derived TCM represents the baseline for the two 20/10 m primary status layers Tree Cover Density (TCD) and Dominant Leaf Type (DLT). The 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. Both layers are sharing the same spatial extent and provide information on the leaf type (broadleaved /coniferous) and the proportional tree cover at pixel level. This allows users to apply a (national) forest definition, taking any canopy crown cover into account, which fits best to their specific needs.

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

Included Features (if detectable from the 20/10 m imagery)	Excluded Features (if detectable from the 20/10 m imagery)
<ul style="list-style-type: none"> • 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, thinning, 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 	<ul style="list-style-type: none"> • 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 and has the following main specifications:

- 20/10 m spatial resolution
- Tree Cover Density range of 0-100%
- No Minimum Mapping Unit (MMU); pixel-based
- Minimum Mapping Width (MMW) of 20/10m

DOMINANT LEAF TYPE

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

- 20/10 m 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 20/10 m

FOREST TYPE

The 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:

- 20/10 m spatial resolution
- Tree Cover Density range of ≥ 10 -100%
- Minimum Mapping Unit (MMU) of 0.52/0.5 ha (13/50 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 20/10 m

2.3.2.2.2 Methodology

HRL FOREST 2012

In 2012, HRL Forest products have been produced in five geographically split lots by four different consortia, using different methodologies in terms of pre-processing, classification and post-processing. Since the methodologies applied are partially unknown to the ECoLaSS consortium, this cannot not be discussed in any further detail here. However, data basis was a mono-temporal pan-European coverage from the HR_IMAGE_2012 dataset, which has shown a series of shortcomings (5 different sensors, acquisitions outside the vegetation period, high cloud/haze cover, data gaps) related to the HRL production. Even though a streamlining phase has been conducted in order to harmonize the output results between the lots, the combination of the points mentioned above led to different results in quality across Europe. In consequence, this led to a partial correction of 2012 status layers in frame of the HRL Forest 2015 production.

HRL FOREST 2015

In 2015, a fundamental change in the overall methodology (see above) as well as in the overall data availability (and quality) has been taken place. Besides a consistent and standardized pre-processing (geometric correction, Top-of-Atmosphere correction, topographic normalization) 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 2-3) 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 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).

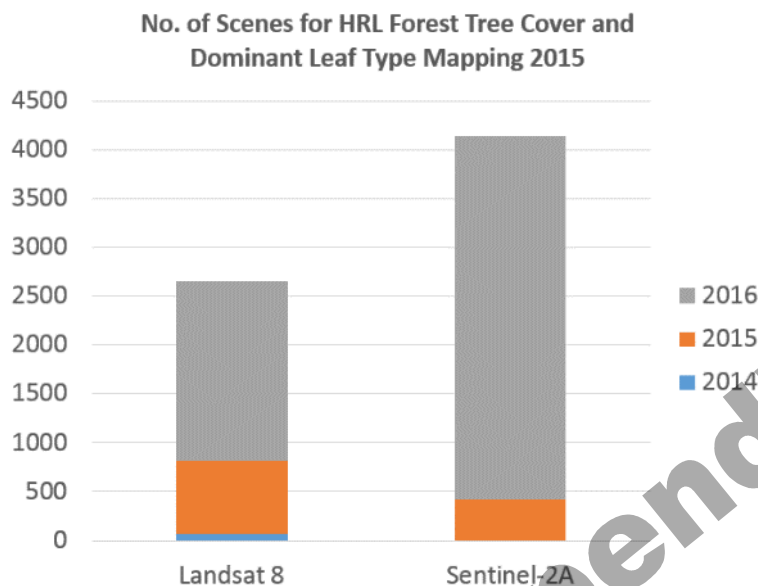


Figure 2-3: Number of Scenes for HRL Forest Tree Cover and Dominant Leaf Type Mapping 2015.

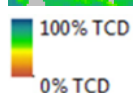
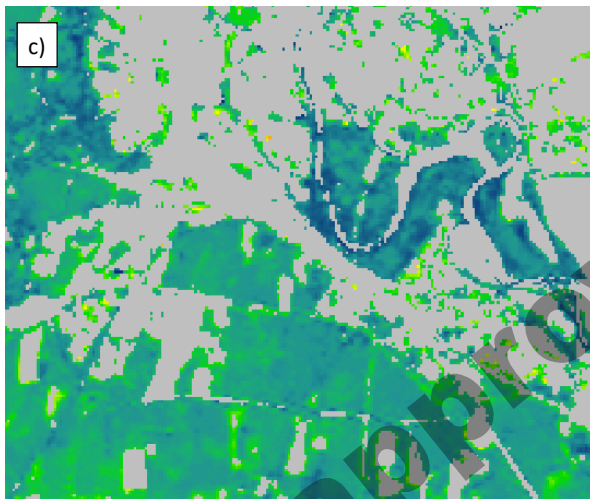
Contrary to the DLT 2015, the status of the TCD 2015 has been derived by classifying nearly 1,000 single satellite images (Sentinel-2, Landsat 8, ResourceSat-2, SPOT-5) from the 2015 reference year (± 1 year) on a mono-temporal basis, but within the confines of the multi-temporally derived Tree Cover 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, which 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.

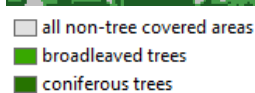
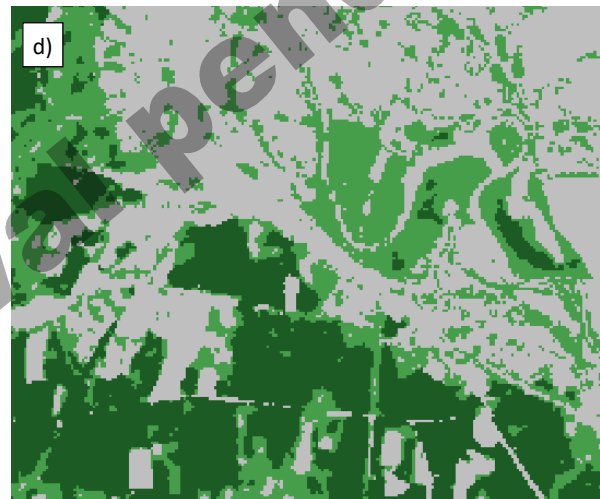


Sentinel-2A acquired on 27.08.2016
modified Copernicus Sentinel data [2016]



20m Tree Cover Density 2015

Courtesy of the European Environment Agency



20m Dominant Leaf Type 2015

Courtesy of the European Environment Agency

Figure 2-4: 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 2-4 shows the outcome of both, the TCD 2015 (Figure 2-4c) and the DLT 2015 (Figure 2-4d) classification based on modified Copernicus Sentinel-2 data (Figure 2-4b). Compared to the VHR_IMAGE_2015 (Figure 2-4a) the DLT distinguishes well between broadleaved and coniferous trees. The parts of lower and higher density visible in Figure 2-4a are well represented by the TCD in Figure 2-4c. 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%.

HRL FOREST 2018

The HRL Forest 2018 will provide improved status layers at 10m resolution. The whole production is being performed by one Service Provider and utilises some of the findings made in ECoLaSS. This relates in particular to the use of time features derived from optical Sentinel-2 and Sentinel-1 SAR data for the vegetation period. The production scheme is largely considering the distribution of biogeographic regions and environmental zones according to Metzger et al. (2005). Processing takes place on the DIAS platform [Mundi](#) to take advantage of the storage and processing capacities of a powerful cloud infrastructure. Test and training data are automatically selected and feed a Random Forest classifier for generation of the Tree Cover Mask and the primary status layer Dominant Leaf Type. The primary status layer Tree Cover Density benefits from a revision and extension of the already existing Forest Reference Database. Change products (Tree Cover Change Mask 2015-2018, Dominant Leaf Type Change 2012-2015-2018) are being created using a Reference Database for Change Calibration. Validation of the HRL Forest 2018 products will be performed with a systematic random sampling grid, consisting of more than 12,000 PSUs. Results will be published in 2019.

2.3.3 HRL Grassland

The HRL Grassland chapter contains information about the state of the art, concerning different technical approaches for grassland mapping and then focuses specifically on the HRL Grassland production. Afterwards, a desk study about the mapping of Mediterranean grassland, elucidates the particular challenges in this context.

2.3.3.1 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).

2.3.3.1.1 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+) (Sánchez-Hernández 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; Sánchez-Hernández 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 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 practices (Zillmann et al., 2014). In ECoLaSS prototypes implementation, and tests carried out in the different biogeographic regions, it is clear that the time windows selection for a correct grassland identification must be within the periods when the crops (being the class that generates more frequent confusion with grasslands) show a clearly distinctive phenological status with respect to grasslands.

Object based approaches based on optical data

While the above mentioned methods focused on pixel-based approaches, others focused 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 Authoritative 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.

2.3.3.1.2 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 σ_0 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. Erasmí (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.

2.3.3.1.3 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).

2.3.3.2 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 20 m spatial resolution and a minimum mapping unit of 1 ha 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 was fundamentally reconsidered and comes now with a revolutionized approach concerning definition, workflow and technical aspects, as well as an improved data base. At present, the HRL 2018 production has been recently started. The concept of the present HRL 2018 Grassland and Grassland Change constitutes a further development step in terms of grassland products and requirements. Besides a change of the HRL Grassland product specifications concerning improved spatial resolution (HRL 2015: 20m, HRL 2018: 10m) and minimum mapping unit (MMU) (HRL 2015: 1 ha,

HRL 2018: 10m) a new product has been defined, the grassland change layer 2015-2018. Further, a possible grassland use-intensity product is under discussion. Class definitions of the products are unchanged in order keep product traceability and comparability. HRL production targets the HRL 2018 Grassland Status Map at 10 m, the HRL Grassland Change 2015-2018 Map at 20 m, and additional products, in consistency with the HRL 2015 additional products: Grassland Vegetation Probability Index GRAVPI, Ploughing Indicator PLOUGH and Confidence Layers. In addition, guaranteeing traceability, the following products are envisaged: time feature identification layer, production unit layer, time features, parent scene identification layer, data score layer, data density layer and time series completeness layer. These layers contribute to the quality assessment. The processing environment is Mundi DIAS. In this regard, it must be noted that the grassland prototypes in ECoLaSS have well ahead addressed the spatial resolution specifications and improved the preexisting grasslands, defining innovative workflows towards operational roll-out.

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
<ul style="list-style-type: none"> ▪ 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 	<ul style="list-style-type: none"> ▪ 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 $\geq 30\%$ 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. The definition of the HRL 2015 has proved to be a very practicable one during the production. It allows continuity, consistency and comparability regarding

a time and geographic perspective on pan-European level and it is considered to give valuable results. It is continued in HRL2018 production.

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 2-5, the total amount of images used for the production of the GRA 2015 mask, is displayed per year and satellite. The available multi-temporal Sentinel-2A data coverage of the project area was quite satisfying in southern Europe, but the number of cloud free acquisitions decreased steadily to the North (i.e., hardly any cloud-less Sentinel-2A observation could be observed). For that reasons additional HR data sets from other sensors and with different specifications,

like e.g. Landsat-8 OLI data, were used as additional data source to complement as much as possible seasonal time-series. The launch of Sentinel-2B in March 2017, now provides European coverage every 5 days instead of every 10 days with Sentinel-2A only, which guarantees a much higher rate of optical data and a larger amount of cloud-free acquisitions. These dense acquisitions highly increase the possibility of covering the relevant time windows for grassland mapping with cloud-free data. Nonetheless, the implementation in ECoLaSS suggests that data situation is highly yearly and seasonal dependent. In this regard, the contribution of Sentinel-1 data should be exploited for classifying grasslands and to complement the time-series in those areas, where optical data are not of suitable quality due to haze, clouds or cloud shadows. Moreover, Sentinel-1 significantly enhances the discrimination between grasslands and other land cover features if appropriate training samples are available.

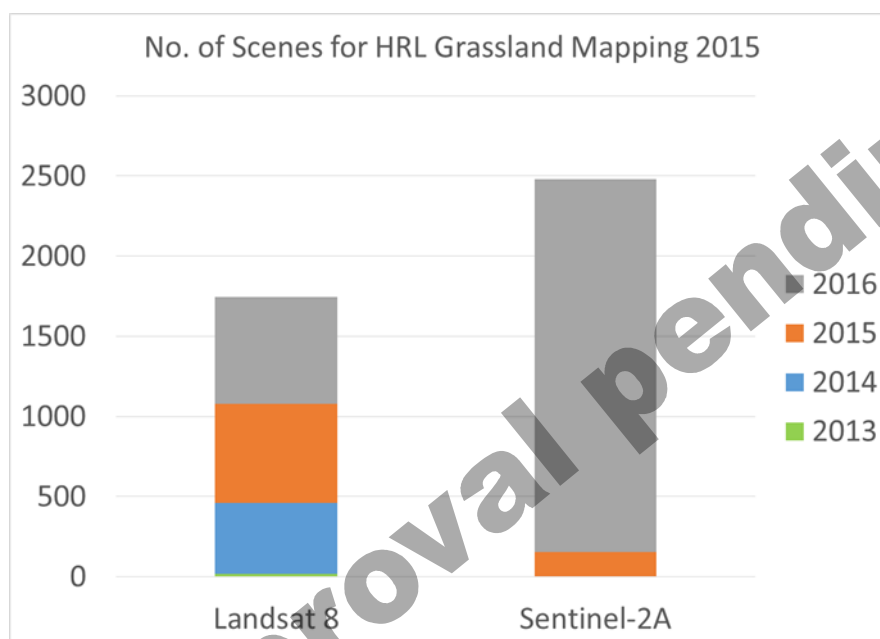


Figure 2-5: Summary of the total amount of images used for the production of the GRA 2015 mask, per year and satellite.

Additional products

The *Ploughing Indicator PLOUGH* indicates the time period (in years) since the last ploughing activity has taken place (Figure 2-6), 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 2-6: 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* (Figure 2-7) 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.

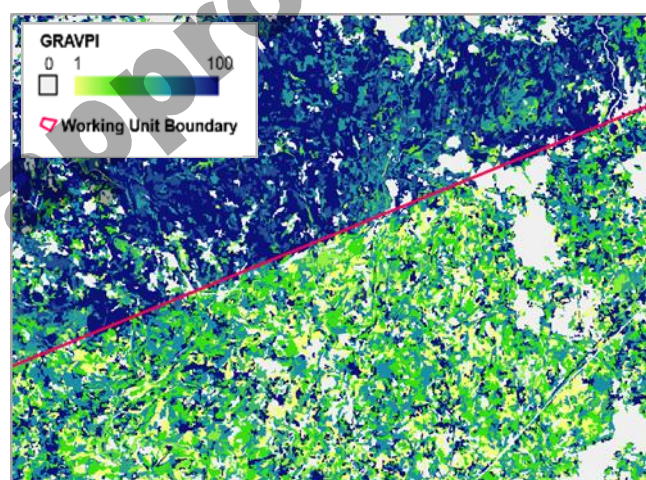


Figure 2-7: Example of GRAVPI from Turkey. The upper Working Unit (WU) provides a high number of adequate scenes for classification and thus a better data base than the WU below. The GRAVPI above consequently shows significantly higher percentages.

Like the main product HRL GRA 2015, PLOUGH and GRAVPI provide a spatial resolution of 20 m and a MMU of 1ha. The prototypes in ECoLaSS are produced at a spatial resolution of 10 m and MMU of 0.5 ha, as a result of the recommendations extracted from the tests.

2.3.3.3 Mapping Mediterranean Grassland with Multi-temporal Earth Observation Data 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 region¹ showed potential for enhancement and fostered a second thought about a methodological adaption. The implementation experiences in ECoLaSS suggest a stratification based on bio-geographical regions or landscapes with similar characteristics for a potential roll-out. This has also been proved to be the case in the agriculture prototypes developments.

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 Value² (Vrahnakis 2016). However, mapping of these and other significant grassland areas by means of EO data 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 adaptations 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.

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).

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. In ECOLaSS, the prototypes implementation in phase 2 has incorporated the SAR features from S-1, in combination with S-2 features. The multisensory approach has been benchmarked against single sensor in terms of accuracy and cost-efficiency performance.

2.3.3.3.1 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.

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 2-8) 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 droughts³ of variable length, and wet and mild winters with relatively high inter-annual variability (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

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.

- 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)

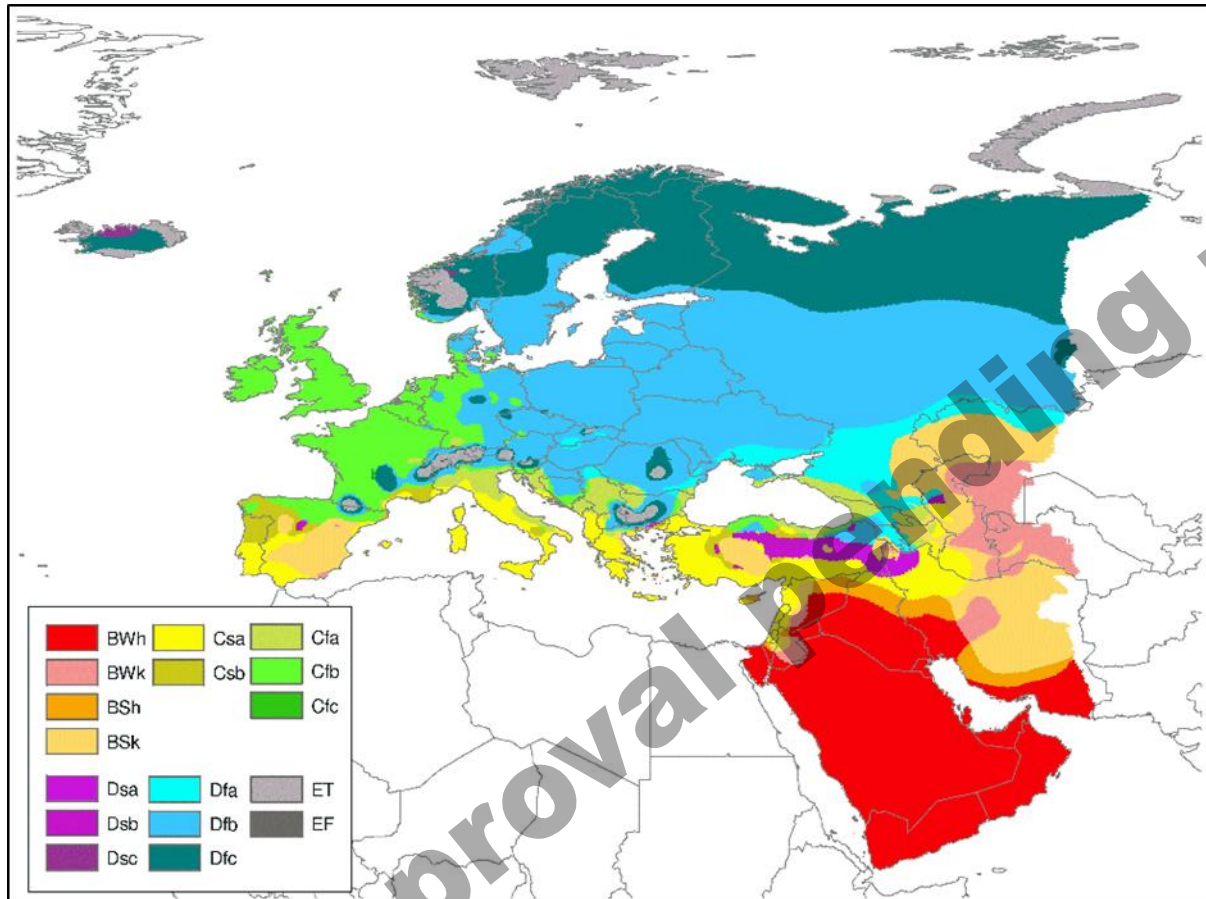


Figure 2-8: Mediterranean climatic map after Köppen & Geiger (Peel et al. 2007).

The tendency to experience (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). 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).

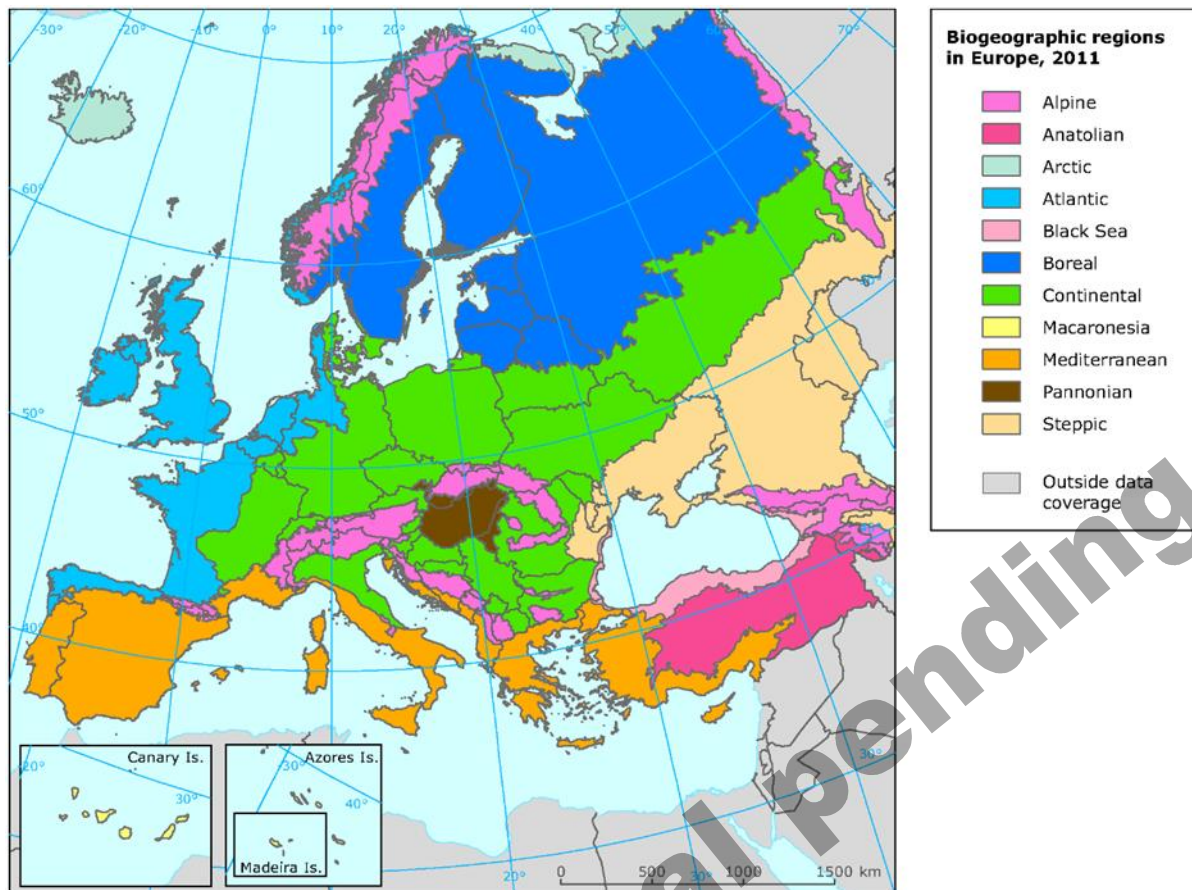


Figure 2-9: Biogeographic regions in Europe 2011 (EEA 2012).

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. In this sense, the consideration of phenology dynamics throughout the year and monitoring of longer periods by means of biophysical indicators and time series indicators is to provide cross-cutting products enriching the binary masks (e.g., as addressed in WP41).

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.

2.3.3.3.2 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 2.3.3.2), 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. 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 grasslands⁴. 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 2006; 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.

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.

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. 2017).

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.

2.3.3.3.3 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.

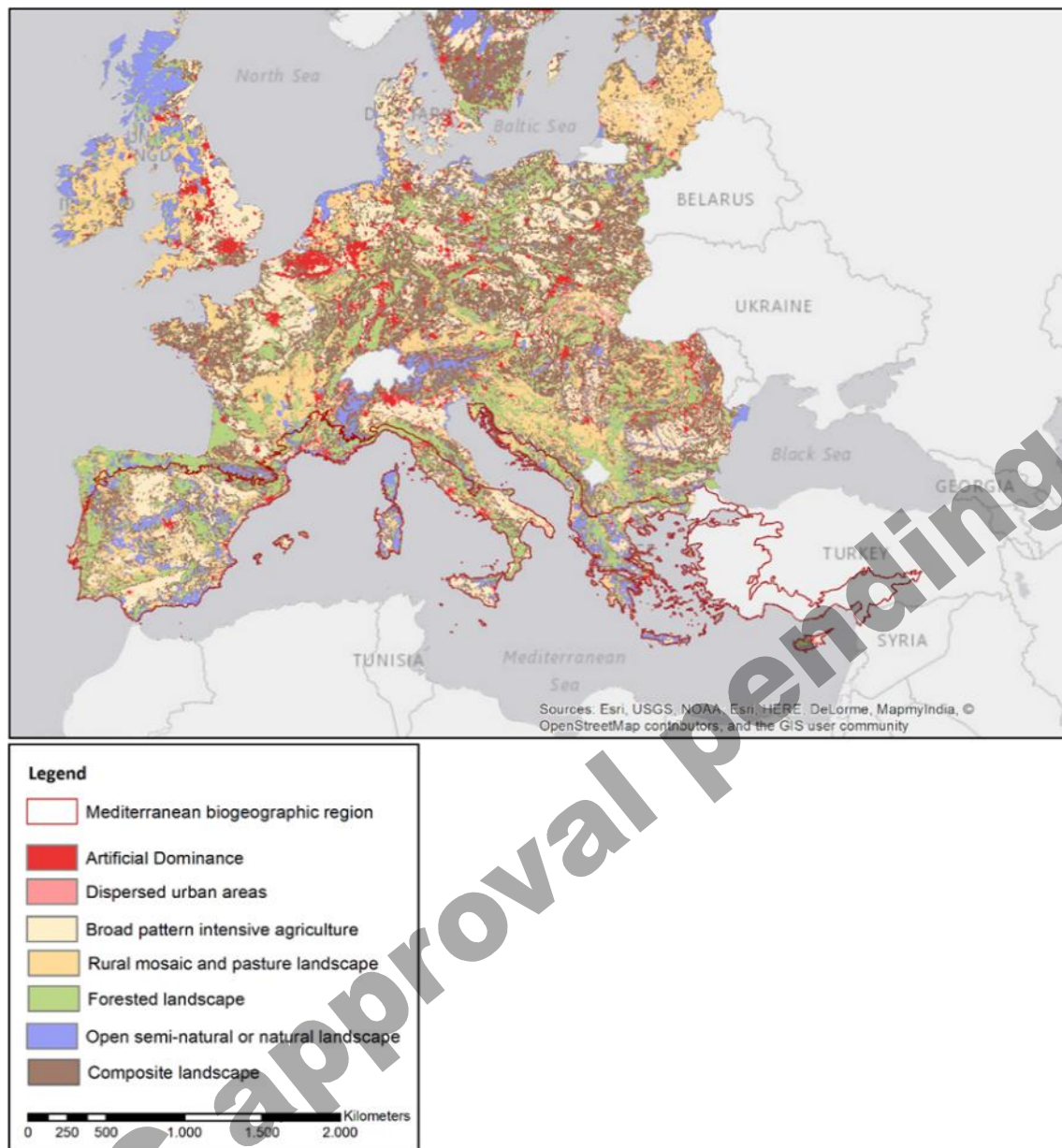


Figure 2-10: Dominant land cover in Europe and the Mediterranean region (red boundaries) (EU 2017 and EUROSTAT 2013)

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; Plantureux 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.

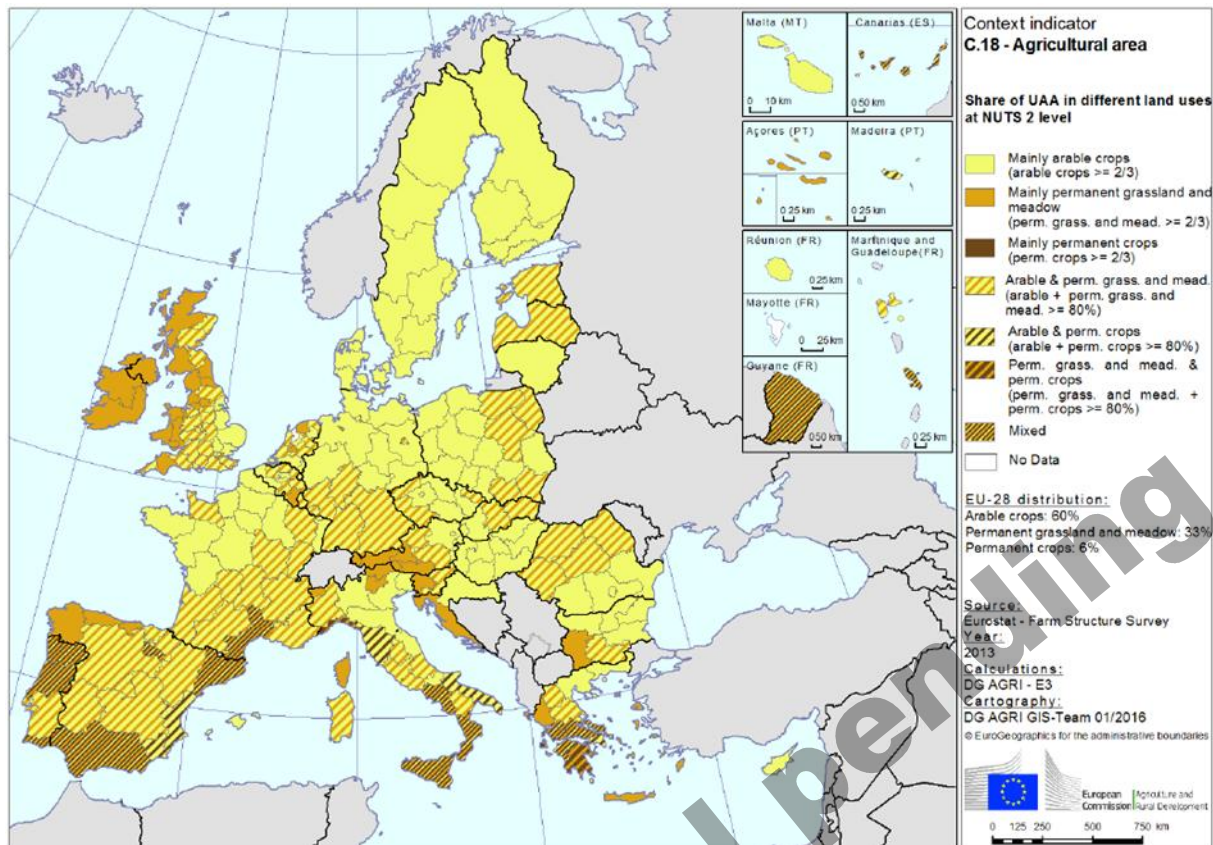


Figure 2-11: Share of utilized agricultural areas (UUA) in different land uses at NUTS 25 level, 2010 (EU2017).

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 et al. 2015; Hervieu 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; 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 consisting of oak trees, shrubs, annual herbaceous species, fodder, winter cereal (Sitzia et al.

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).

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. 2014; Salis et al. 2011; 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 pastoralism⁶ 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.

2.3.3.3.4 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).

2.3.3.3.5 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 agriculture information for a vast range of users at all geographical scales. The gap between agricultural EO-derived data producers and 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 and crop type 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.

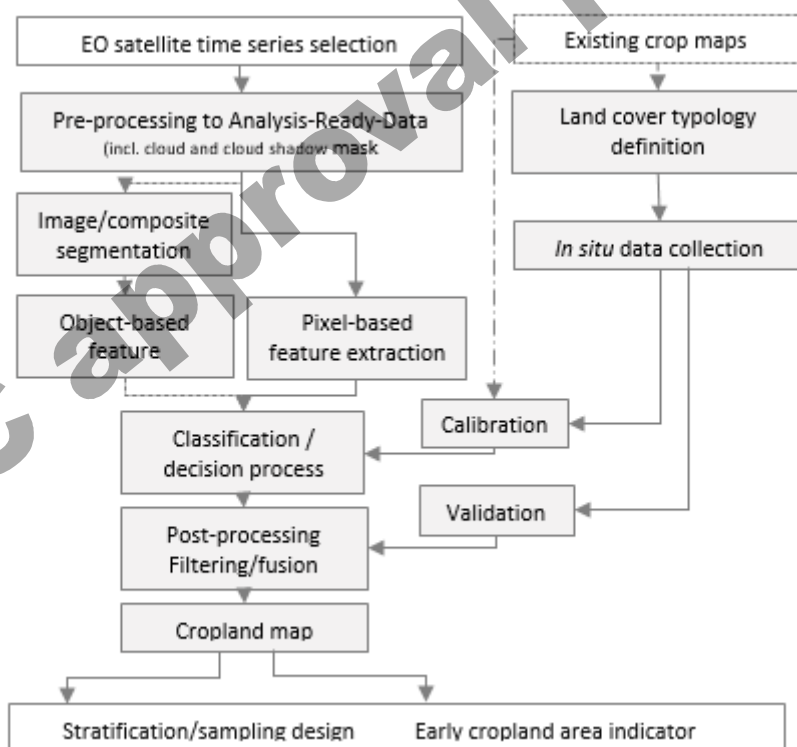


Figure 2-12: Workflow for cropland mapping from satellite observation time series. (Dashed lines correspond to alternative pathways).

This section reviews some key elements of the crop mapping process, as organized according to a standard workflow (Figure 2-12). 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 all land 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 was modified into the Land Cover Meta-Language (LCML) and became an ISO standard, 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. In this sense it shares partly the principles selected for the EAGLE system which is further described in the section 2.3.5.

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. Fortunately, this MMU has been change for the change between two CLC maps to 5 ha (see more info in the 2.3.5). The JECAM definition is found quite relevant for Sentinel 10 to 20 m resolution, insuring that at least some pure pixels are available to label the observation. However, 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. In ECoLaSS, implementation and testing of biophysical time series indicators have been carried out in the frame of WP41. 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 agriculture 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)	<ul style="list-style-type: none"> Simple application Easy to understand and interpret Predicts class membership probability 	<ul style="list-style-type: none"> Parametric Assumes normal distribution of data Large training sample necessary
Artificial Neural Networks (Non-parametric)	<ul style="list-style-type: none"> 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) 	<ul style="list-style-type: none"> Needs parameters for network design Tends to overfit data Black box (rules are unknown) Computationally intense Slow training
Support Vector Machines (Non-parametric)	<ul style="list-style-type: none"> Manages large feature space well Insensitive to Hughes effect Works well with small training data set Does not overfit 	<ul style="list-style-type: none"> Needs parameters: regularization and kernel Poor performance with small feature space Computationally intense Designed as binary, although variations exist
Decision Trees (Non-parametric)	<ul style="list-style-type: none"> 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 	<ul style="list-style-type: none"> 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)	<ul style="list-style-type: none"> Capacity to determine variable importance Robust to data reduction Does not overfit Produces unbiased accuracy estimate Higher accuracy than DTs 	<ul style="list-style-type: none"> 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 increase of temporal and spatial resolutions offered by the Sentinel constellation is expected to 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, that could not be achieved through the current implementation based on mono-temporal VHR snapshots, but that will be clearly within reach thanks to the use of dense time series. Those quicker deliveries will de facto lead to a better monitoring of the different kinds of change or transitions from one LC to another. This aims at creating, for example, 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. This consistency is also the point of focus for the upcoming CLC+ product.

2.3.5.1 Previous attempts

The past 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 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 have called for the emergence of new 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) and 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.

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, 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%.

2.3.5.2 CORINE Land Cover

The CLC inventory was initiated in the 1980s to standardize data collection on land in Europe, mainly to sustain environment policy development. Information is provided on LC, through biophysical characteristics of the surface, which can be determined in a semi-automated fashion, but also LU, which requires the input of a human interpretation, through the use of 44 classes at a level-3 in the hierarchical nomenclature, with a MMU of 25ha and a MMW of 100m. Full maps are freely available online (Copernicus, 2019). The past implementation has relied on the national entities, through a bottom-up process, where each national team independently produces the databases for their own country, before the integration at European level.

Despite its limitations for spatial resolution, which were mainly dictated by the geometrical accuracy expected for the first satellite data used in 1985, CLC remains a widely used dataset - may it be as primary source for the development of various indicators (Gardi, Bosco, Rusco, & L., 2010), for environmental or urban modelling (Siedentop & Meinel) (Gallego, Peedell, & al.) and change analysis in the LC/LU (Feranec, Jaffrain, Soukup, & Hazeu, 2010) at European down to regional levels.

Table 2-4: Key elements regarding the evolution of CLC through the years

	CLC 1990	CLC2000	CLC2006	CLC2012	CLC2018
Satellite Data	Landsat 5 MSS/TM (single date)	Landsat 7 ETM (single date)	SPOT-4/5 IRS P6 LISS III (dual date)	IRS P6 LISS III RapidEye (dual date)	S-2 Landsat-8 (for gap- filling)
Geometric Accuracy	≤ 50 m	≤ 25 m	≤ 25 m	≤ 25 m	≤ 10 m (expected for S-2)
Time Consistency	1986-1998	1999-2001	2005-2007	2011-2012	2017-2018
Thematic Accuracy	≥ 85% (probably not achieved)	≥ 85% (achieved)	≥ 85%	≥ 85% (probably achieved)	≥ 85%
Production Time	10 years	4 years	3 years	2 years	1.5 years
Number of Countries Involved	26 (27 with late implementation)	30 (35 with late implementation)	38	39	39

The choice, through the years, to retain such limiting parameters, detailed in Table 2-4, lies in the EEA's will to maintain a full comparability between consecutive releases. At the moment, the complete CLC dataset is still the most popular EEA databases downloaded. The time series is complemented by a change layer - however, it should be noted that those layers highlight changes with a MMU of 5 ha. This difference in MMU means that the difference between two status layers (at a MMU of 25 ha) will not equal to the corresponding CLC-Changes layer (at a MMU of 5 ha). This "incompatibility per construction" should be addressed in the upcoming CLC+ product.

2.3.5.3 Current state-of-the-art

A wall-to-wall land cover map at country scale – in this study, France – was produced by the Centre d'Études Spatiales de la BIOSphère (CESBio) based solely on Landsat-8 and S-2A 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 LC classes while providing a complementary confidence map at pixel level, listed as:

- Annual summer crops;
- Annual winter crops;
- Broad-leaved forest;
- Coniferous forest;
- Natural grasslands;
- Woody moorlands;
- Continuous urban fabric;
- Discontinuous urban fabric;
- Industrial or commercial units;
- Road surfaces;
- Bare rock;
- Beach, dunes and sand plains;
- Water bodies;
- Glacier and perpetual snow;
- Intensive grasslands;
- Orchards;
- Vineyards.

France land cover classification, from Landsat 8 to Sentinel-2.

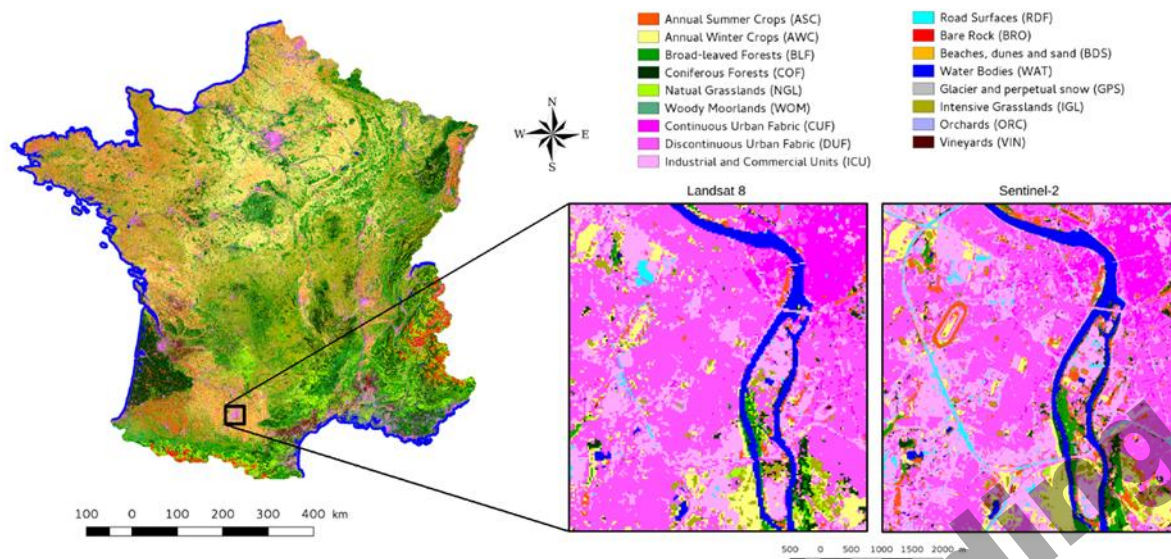


Figure 2-13: Overview of the 2017 cover map produced by the CESBio (Source: <http://osr-cesbio.upstlse.fr/~oso/>).

The following releases of this wall-to-wall land cover map for the year 2018 based on the full time series of S-2A and B images, retreated at level 2A, over France (Inglada, 2019), increasing the number of classes up to 23 between the iterations 2017 and 2018 to slightly enhance the global accuracy, by splitting the annual summer crops into sunflower, corn, rice, tuber or roots, while the winter crops have been separated between straw cereals, rapeseed and protein crops.

Further demonstration of the suitability of S-2 time series of the year 2017 for mapping LC over entire European countries is found in the results of the project S-2 for Global Land Cover (S2GLC), funded by the ESA. The random forest algorithm is used to classify each image of the time series of S-2 images at level 2A, pre-processed by Sen2Cor (see [AD07]) with a customized improved cloud mask which had to be generated for each image, before using aggregation rules to merge them and create the final LC classification (Lewinski et al. 2019). The classes for a complete map of Europe are the following:

- Artificial surfaces and constructions;
- Cultivated areas;
- Vineyards;
- Broadleaf tree cover;
- Coniferous tree cover;
- Herbaceous vegetation;
- Moors and Heathland;
- Sclerophyllous vegetation;
- Marshes;
- Peatbogs;
- Natural material surfaces;
- Permanent snow cover;
- Water bodies.

At a larger spatial resolution (100m) and a global scale, the institute VITO (short for Vlaamse Instelling voor Technologisch Onderzoek) has released new global land cover maps (Copernicus, 2019) (VITO, 2019) based on Proba-V optical data (a medium resolution satellite, with a native resolution at 300m and a 2 days revisit time) merged with S-2 and S-1 data in order to compensate for gap left by clouds - especially in the Intertropical Convergence Zone, known for its very strong cloud cover. Classes are

environment-oriented in order to provide better monitoring for biodiversity preservation and a first validation put the accuracy at 80%. Some of those classes can be expressed as fractional cover layers, and are:

- Forest;
- Shrubland;
- Herbaceous vegetation;
- Moss and Lichen;
- Bare/Sparse vegetation;
- Built-up;
- Cropland;
- Snow and Ice;
- Permanent inland water bodies;
- Seasonal inland water bodies.

VITO's remote sensing team had already begun to work earlier on optical and SAR data fusion and released a map of Belgian types of culture over the whole country (Van Tricht, Gobin, Gilliams, & I., 2018).

2.3.5.4 Toward CLC+

The Environment Information and Observation Network (EIONET) Action Group on Land monitoring in Europe (EAGLE Group) has been tasked by the EEA to develop the design and technical specifications of the second generation of CLC products. Consultations with various stakeholders have taken place and several preliminary versions of those specifications have been released before being refined by the EEA and definitively released in the fresh ITT EEA/DIS/RO/19/012 for the CLC+ Backbone. The annex 7, "Technical specifications for implementation of a new land-monitoring concept based on EAGLE – D5: Design concept and CLC+ Backbone, technical specifications, CLC+ Core and CLC+ Instances draft specifications, including requirements review", has been used in the second phase as a guidance to create New Land Cover prototypes approaching the CLC+ Backbone products, while remaining in the boundaries of ECoLaSS project scope.

Several key requirements have been identified, in particular a backward compatibility with CLC for the new products, which should be produced using the latest state-of-the-art concepts and developments to fulfill user requirements. This led to the design of 4 interlinked elements:

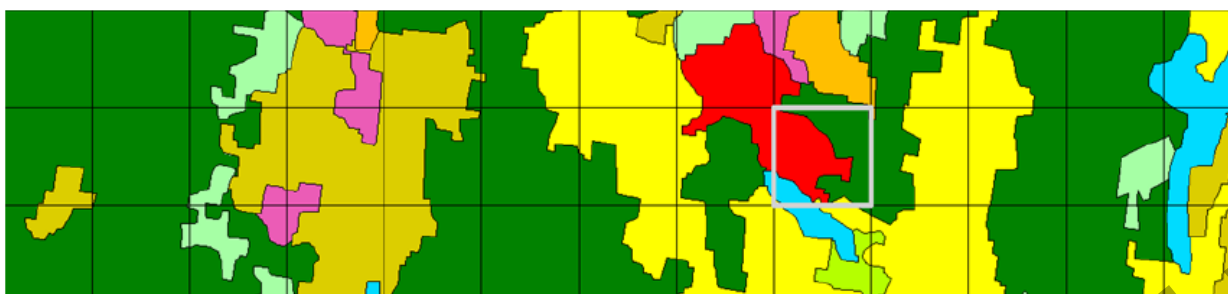
- CLC+ Backbone: spatially detailed, in vector format providing the geometric structure of the landscape, complemented by raster data
- CLC+ Core: Multi-use grid database repository to be populated with various LC/LU ancillary data from CLMS and other sources
- CLC+ Instance: A derived grid product from CLC+ Core, which can be tailored at will for different types of application
- CLC+ Legacy: One of the instances (raster and vector) that can be derived from the CLC+ Core database, populated with ancillary data to retrieve the 44 classes from CLC

The CLC+ Backbone, which is the main focus of the report, is composed of:

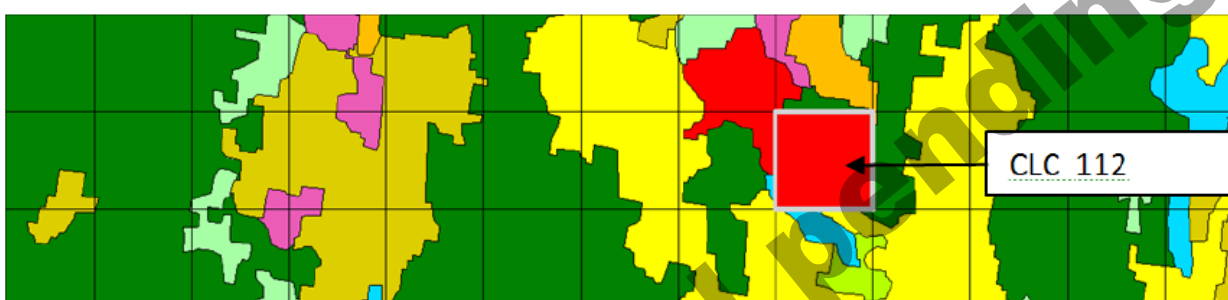
- A 10*10m² raster;
- A vector product with a MMU of 0.5ha.

The envisioned methodology for the second phase is to follow the ITT requirements as closely as possible, using Sentinel datasets as main satellite image sources. Further details are discussed in the section 3.3.5.

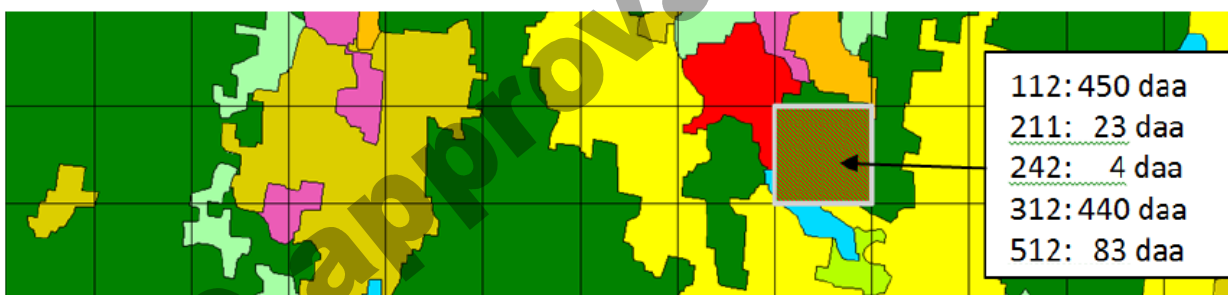
As mentioned as a draft section in this ITT, the next step, the CLC+ Core, will be presented in a new format, between raster and vector shapes: a GRID-based information system. Where each pixel of a raster is classified, i.e. assigned a single particular LC, the grid cells can be characterized by how much it contains of each LC or other information. An example can be seen on Figure 2-14.



CLC with 1 km raster/grid superimposed



Raster representation: A single class is assigned to each pixel



Grid representation: The internal composition of the grid cell is kept as an attribute vector.

Figure 2-14: - Figure extracted from the ITT EEA/DIS/R0/19/012 - caption can be read as: " CLC with a 1 km raster/grid superimposed (top) illustrating the difference between encoding a particular unit as raster pixel (centre) or a grid cell (bottom). "daa" is a Norwegian unit: 10 daa = 1 ha."

The creation of a pan-European HR LC layer will be obtained by superposing together all the currently available layers, in addition to the newly created agricultural layer in this project.

This first superposition step could also be an example of the content expected in a GRID format product based solely on the HRLs. It is expected that the grid size should be for example 10 by 10m, below the MMU of the grid, that could be 0.5ha or lower to match the CLC+ Backbone product specifications. At the moment, the CLC+ Core is expected to be populated with:

- CLC+ Backbone;
- HRLs (not only LC information, but also parameters such as tree density or sealing degree);
- Hot Spot monitoring (Urban Atlas, Natura 2000, Riparian Zones);
- Any other relevant datasets: CLC, LPIS, OSM ...

It should be noted that, to display such GRID interactive format, an online access to a server and its database is required – this cannot be provided as deliverable in the framework of this project.

The next step, which consists in merging the superimposed layers and can be presented as a raster, 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.

No testing or benchmarking is required to produce this fused layer, which is does not require classification methodology, therefore details of the implementation are provided directly in the WP45 report [AD10].

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

- EC approval pending -

2.4 Accuracy assessment principles

This section presents the overarching principles regarding validation procedures in the ECoLaSS project prototypes assessment. These guidelines are taken as reference, to be followed in all implementations in phase 2 for consistency reasons across the different thematic topics, although are further developed where needed, depending on the specifics of each land cover type. The sampling design and statistics described below are common practice in remote sensing applications and land cover studies. Detailed descriptions on computations and background can be found in many papers (e.g., Pontius and Millones 2011, Congalton and Green 2009, Gallego 1995 and 2004, Foody and Arora 1997).

This guideline shall contribute to harmonize the ECoLaSS accuracy assessment approaches. It shall support a standardized presentation of the accuracy results and, in that way, facilitate an evaluation of the various outcomes.

Although accuracy assessment design of each testsite will be to a certain extent product-specific, some basic rules need to be respected that will be laid down in this section. In general, the assessment of each ECoLaSS product's accuracy shall fulfil the basic components of an accuracy assessment and shall describe:

- (i) **sampling design** for selecting the reference sample;
- (ii) **response design** for obtaining the reference land-cover classification for each sampling unit; and
- (iii) **thematic accuracy assessment** carried out.

(i) Description of Sampling design and stratification approach

Stratification and sampling design for each product shall be defined by describing:

- **the Sampling frame:**

Recommendation: A **probability sampling design** based on random sampling techniques is recommended for the ECoLaSS project for its objectivity. Categories included are simple random, stratified random, clustered random and systematic designs. The validation approach preferred should combine **systematic and stratified approaches** in order to benefit from the advantages of both by allocating a pre-defined number of samples per land-cover class within a regular grid. In that way, the geographical spread of the samples is guaranteed.

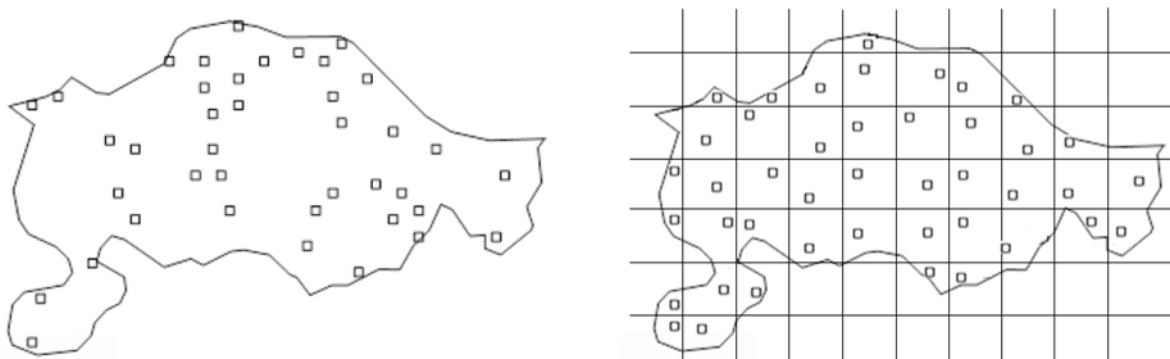


Figure 2-15: Simple random (left) and random systematic (right) sampling designs

- **the Sample Unit:**

Recommendation: As sample unit, point samples are considered as most applicable. Areas may be also used in specific cases, when the geometry of mapped objects needs to be assessed.

- **the Sample Size:**

Recommendation: The number of samples used to assess the accuracy of a product may be calculated as a function of a class's areal proportion in the geographic reference area. In case of small strata, the sampling grid may be systematically densified in order to enable a sufficient number of available grid points for each stratum. Further, sample positions located directly at (often fuzzy) object boundaries should be avoided. For that reason, a minimum distance to a polygon boundary shall be applied.

The required minimum number of sample units per stratum is generally dependent on the number of thematic classes, the spatial extent of the stratum, the expected acceptable error rate and the required precision level (i.e. width of the confidence intervals). A suitable sample size for each stratum (i.e. thematic class) may be estimated based on the expected acceptable error rate. The **standard error of the error rate** can be calculated as follows (Büttner et al. 2012):

$$\sigma_h = \sqrt{\frac{p_h(1-p_h)}{n_h}}, \quad \text{Equation 1}$$

where n_h is the sample size for stratum h and p_h is the expected error rate.

For calculating the sample size n_h the equation can be converted so that the sample size is a function of p_h and the desired standard error σ_h :

$$n_h = \frac{p_h(1-p_h)}{\sigma_h^2}. \quad \text{Equation 2}$$

For the ECoLaSS products, if there was an expected error rate is 15%, a minimum of 51 samples per stratum is necessary with a maximum 5% standard error. Such error rates are expected for grasslands for instance on a general basis for the status layers, whereas a higher accuracy (over 90% is the standard for forest layers), and it can be assumed that a lower accuracy is expected in the change products, totally reliant on the corresponding reference layer, or the crop types (more complex legends, with more classes than the binary presence/absence products). This is also on a general basis in line with the accuracy specifications for the HRL products (only the threshold for grasslands for example was applied in the past at the EU extent and it is now applied at the biogeographic region).

- **the Stratification Approach:**

Recommendation: The stratification approach shall focus on omission/commission strata, where omission strata are understood as areas with a higher likelihood of comprising omission errors, and commission strata accordingly have a higher likelihood of comprising commission errors. The number of strata will depend on the availability of reference data for defining low- or high-probability strata. Existing Copernicus LC/LU classification results like e.g. HRLs, CLC, Riparian or Natura2000 or other information sources may be used for stratification purposes.

(ii) **Response design:**

Data used as independent data source to assess the accuracy of ECoLaSS products shall be documented.

Recommendation: Reference information shall be obtained by using data of higher spatial resolution and quality than the production data. The reference data are independent additional in-situ and ancillary data providing more spatial detail and better landscape context to the assessment than the HR imagery used in the ECoLaSS production. Further details regarding the usage of the ESA DWH data are provided in D12.1 “DWH Use for 2017/2018/2019”. Possible reference data are:

- LUCAS point data: Reference labels obtained by visual re-interpretation of LC/LU at point level, while respecting the product specifications in terms of MMU and MMW and the underlying class definitions.
- Copernicus LC/LU data like HRLs, CLC, Riparian, Natura2000: visual cross-check of derived information shall be applied in order to avoid error propagation and to consider land cover changes in case of differing time stamps
- VHR_IMAGE_2018 datasets, in combination with the previous VHR_IMAGE_2015 imagery applied during visual interpretation of sample points
- Further in-situ data: All existing and accessible complementary data that is of superior quality and matches in terms of spatial, thematic and temporal resolution, including photos, national LC/LU data, biotope maps and topographic data.
- National and regional web map services (RGB and/or CIR imagery with varying spatial resolution).

(iii) **Thematic accuracy assessment**

The final accuracy assessment shall be described based on the following metrics:

• **Error Matrix:**

Recommendation: Unequal sampling intensity resulting from the stratified systematic sampling approach should be accounted for by applying a **weight factor** (p) to each sample unit based on the ratio between the number of samples and the size of the stratum considered. The weighing factor is inversely proportional to the inclusion probability (i.e., the probability that a pixel will be included in the sample) of samples from a given stratum. Within a geographic stratum, the inclusion probabilities of all sample units (u) are the same (π_{uh}^* is constant):

$$\hat{p}_{ij} = \left(\frac{1}{N} \right) \sum_{x \in (i,j)} \frac{1}{\pi_{uh}^*}$$

Equation 3

Where i and j are the columns and rows in the matrix, N is the total number of possible units (population) and π is the sampling intensity for a given stratum. \hat{p}_{ij} is computed for all cells of the error matrix.

To combine sample data over several strata, a weighted estimator of the error matrix is required to account for the different inclusion probabilities among strata. The estimation weight is the inverse of each sample unit's inclusion probability, and the proportion of area for each cell of the error matrix is estimated by formula 3. Else, true map accuracies might result over or under estimated.

• **Overall accuracy (OA):**

Recommendation: The OA is measured by the sum of the diagonal of the weighted confusion matrix divided by the total number of validated points:

$$OA = \frac{1}{\sum C} \sum_i C_{i=j} \quad \text{Equation 4}$$

- **User's accuracy (UA):**

Recommendation: The UA is a measure of the commission error (whereby 100%-UA=commission error):

$$UA_j = \frac{1}{\sum_i C_{i,j}} \sum C_{i=j,j} \quad \text{Equation 5}$$

- **Producer's accuracy (PA):**

Recommendation: The PA is a measure of omission error (whereby 100%-PA=omission error):

$$PA_i = \frac{1}{\sum_j C_{i,j}} \sum C_{i,j=i} \quad \text{Equation 6}$$

- **Confidence interval.**

Recommendation: The standard error is calculated for each stratum and an **overall standard error** is calculated based on the following formula (Equation 7):

$$\sigma = \sqrt{\sum w_h^2 \cdot \sigma_h^2} \quad \text{Equation 7}$$

In which w_h is the proportion of the total area covered by each stratum. The 95% confidence interval is +/- 1.96. σ .

- **F1 score statistic:**

Recommendation: The F1 score is computed per class, as the harmonic mean between precision (i.e., User accuracy) and recall (i.e., producer accuracy), where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

- **Kappa statistic:**

Recommendation: Kappa is a measure of the difference between the actual/chance agreement between provided reference data and an automated classifier/random classifier. Although criticized, it is a widely used statistic useful for benchmarking purposes.

Recommendation:

$$k = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_e} \quad \text{Equation 8}$$

In which p_0 is the relative observed agreement among raters (identical to accuracy) and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category. Kappa value is 1 when there is complete agreement. On the contrary, k is 0 if there is no agreement among the raters other than what would

be expected by change (as given by p_e). The statistic value can be negative, implying that there is no effective agreement between the two raters or the agreement is worse than random.

In addition to these metrics, the look and feel of the prototypes is verified. The expert knowledge of the region and land cover are a qualitative accuracy assessment that even though not resulting in a quantitative measure, is essential to land cover classification products. Each prototype is presented with its associated probability layer: for each pixel, the percentage on the classification choice is given. This is considered a pixel based quantitative quality metric (i.e., spatialized reliability).

In the case of the change products, the sampling for validation focuses on the change. Therefore, a stratified approach is to be based on the initial detected change areas from the overlay of the corresponding status layers (e.g., 2017 and 2018 products). Based on this fit-for-purpose calibration dataset, the relative magnitude of actual changes and, thus, the magnitude of errors (omission and commission from the previous and new time step) are estimated. The estimates obtained for the 2017-2018 change layer provides a basis to target the reclassification of the 2017-2018 changes into real changes or omission and commission errors for 2017 and 2018, respectively. Statistical analysis is performed and if required, reclassification of the status layer is to be considered. Priority is given to an adjustment of the classification threshold based on the class probabilities. However, where this does not lead to satisfactory results a re-processing of the readily computed time-features and, if necessary, the original input imagery is considered. The description and implementation of the accuracy assessment of change products is reported in the final issue of WP34 deliverable on methods compendium of time series analysis for change detection [AD08].

Last, INSPIRE compliant metadata xmls files complete the ECoLaSS prototypes data delivery [AD10, AD11, AD12, AD13, AD14].

3 Methods

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 3.3). 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

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.

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. Such approach has been used in the sampling for the Forest and Grasslands prototypes in ECOLaSS, applying outlier detection and also expert-knowledge in some case as required. Quality of samples and reliability of reference data sources are essential to classification processes, evermore in automation of workflows.

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). This has also been applied in sampling for the ECOLaSS test and demosites. 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 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 $1/\text{\#Features}$, 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.

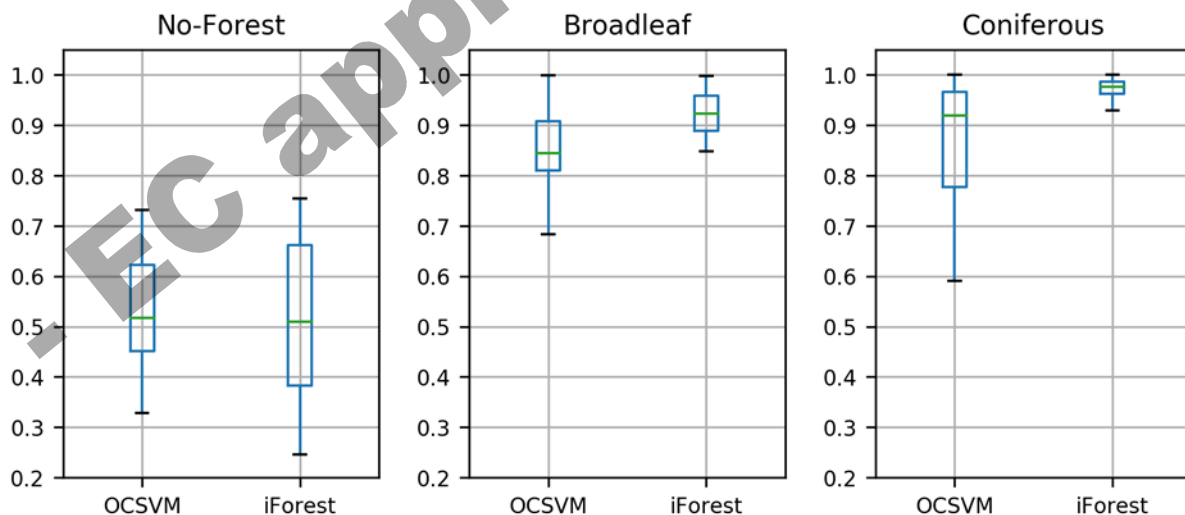


Figure 3-1: Boxplots of AUC values given the class and outlier detection approach achieved over all respective experiments, i.e. varying random replications (5), outlier fractions (10) and assumed outlier fractions (10). Thus, one boxplot is constructed from 500 values.

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.

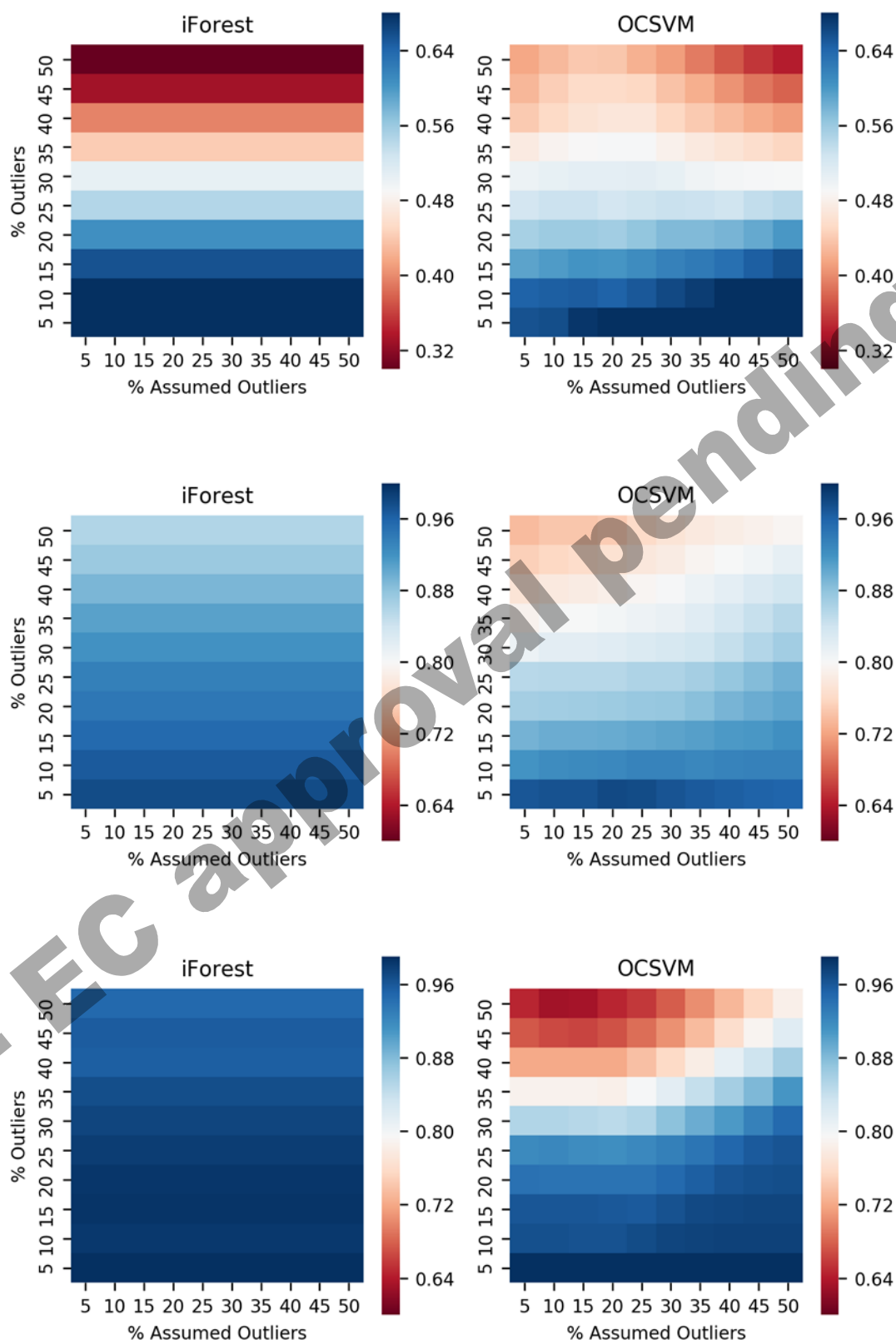


Figure 3-2: Mean AUC for the three classes non-forest, coniferous and broadleaf forest (from top to bottom), the outlier detection approaches iForest (left) and OCSVM (right) dependent on the percentage of assumed outliers (x-axis) and percentage of outliers (y-axis). Each value is the mean AUC of the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

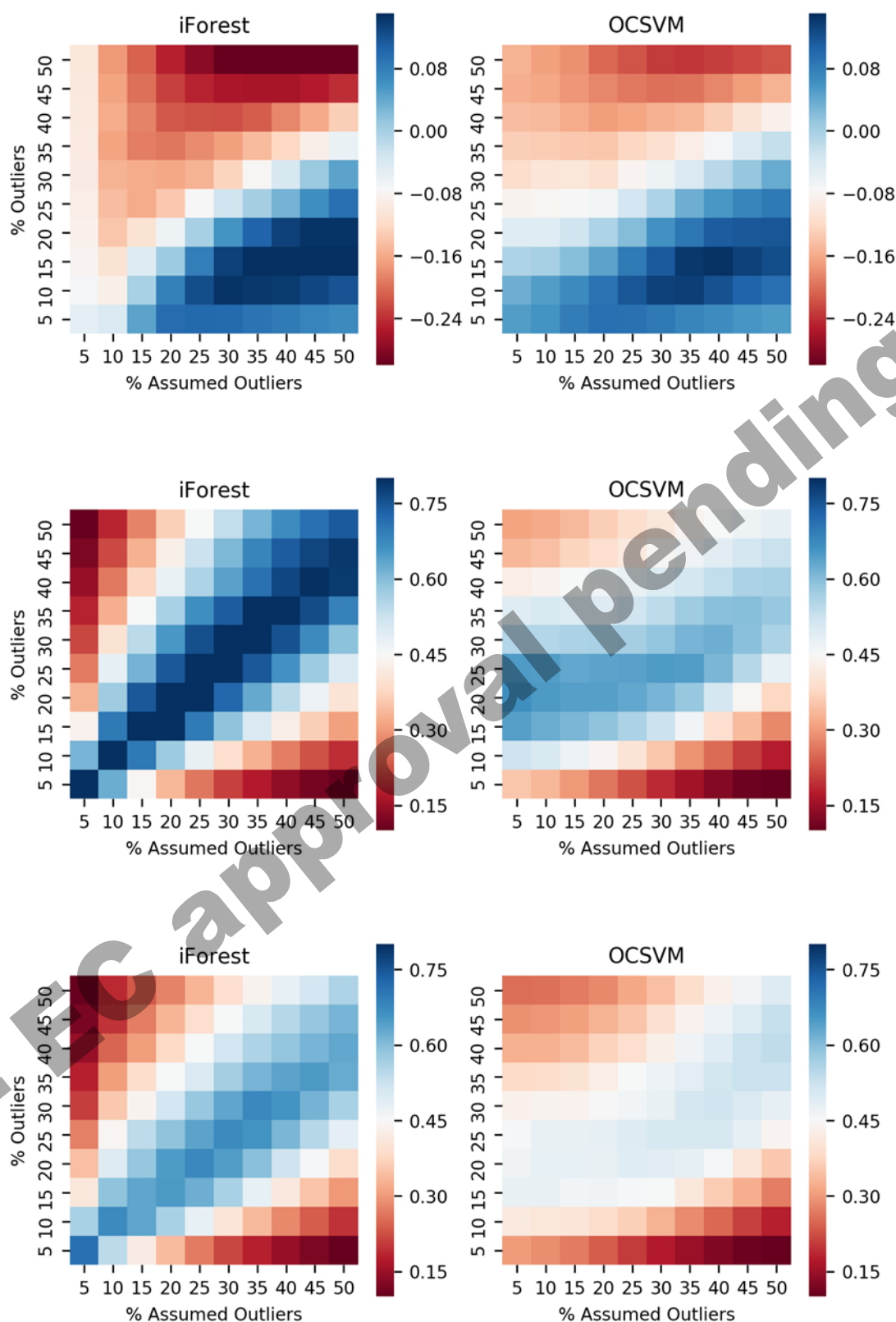


Figure 3-3: Mean kappa coefficient for the three classes non-forest, coniferous and broadleaf forest (from top to bottom), the outlier detection approaches iForest (left) and OCSVM (right) dependent on the percentage of assumed outliers (x-axis) and percentage of outliers (y-axis). Each value is the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

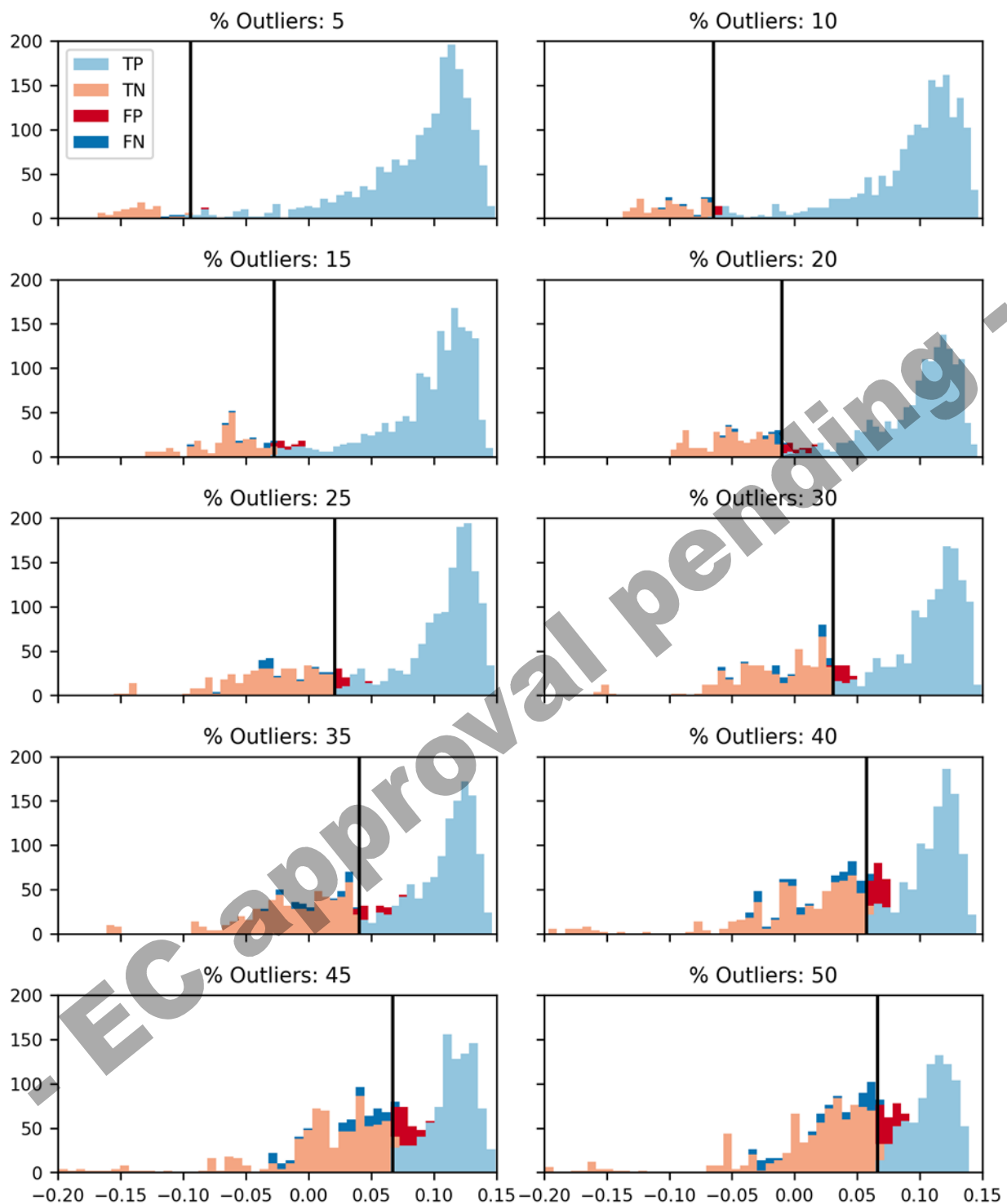


Figure 3-4: Histogram of the iForest decision function values for the coniferous forest class containing different percentages of outliers (see subplot title). The black vertical line shows the location of the threshold when the assumed outlier percentage corresponds to the actual outlier percentage. 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 careful 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 and PROBA-V compositing

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 criterion 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 achieved 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 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.

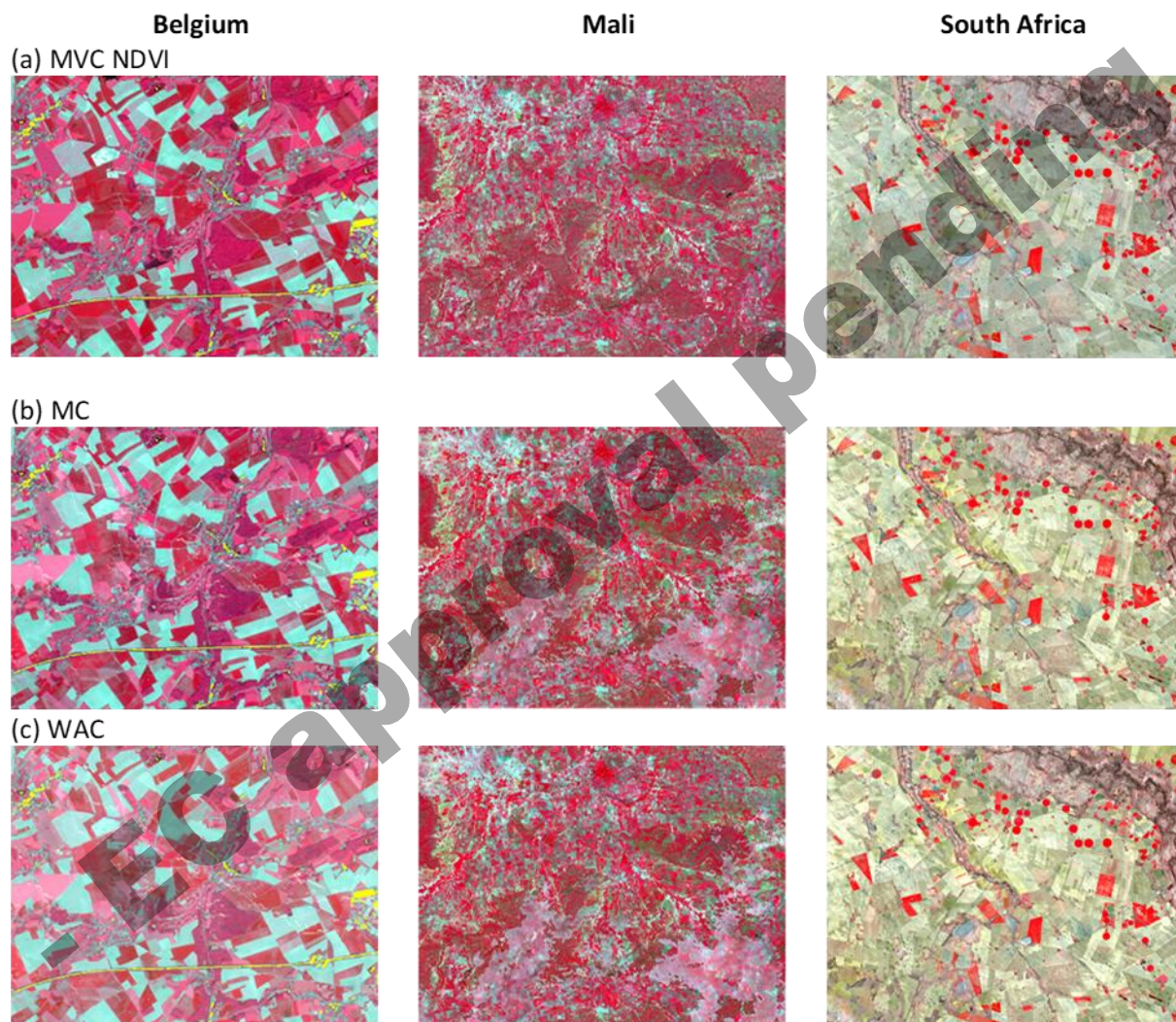


Figure 3-5: False colour (b8, b3, b2) monthly composites over the Belgium site (2017-05), Mali site (2016-08) and South Africa site (2016-09) of the (a) MVC NDVI, (b) MC and (c) WAC algorithms.

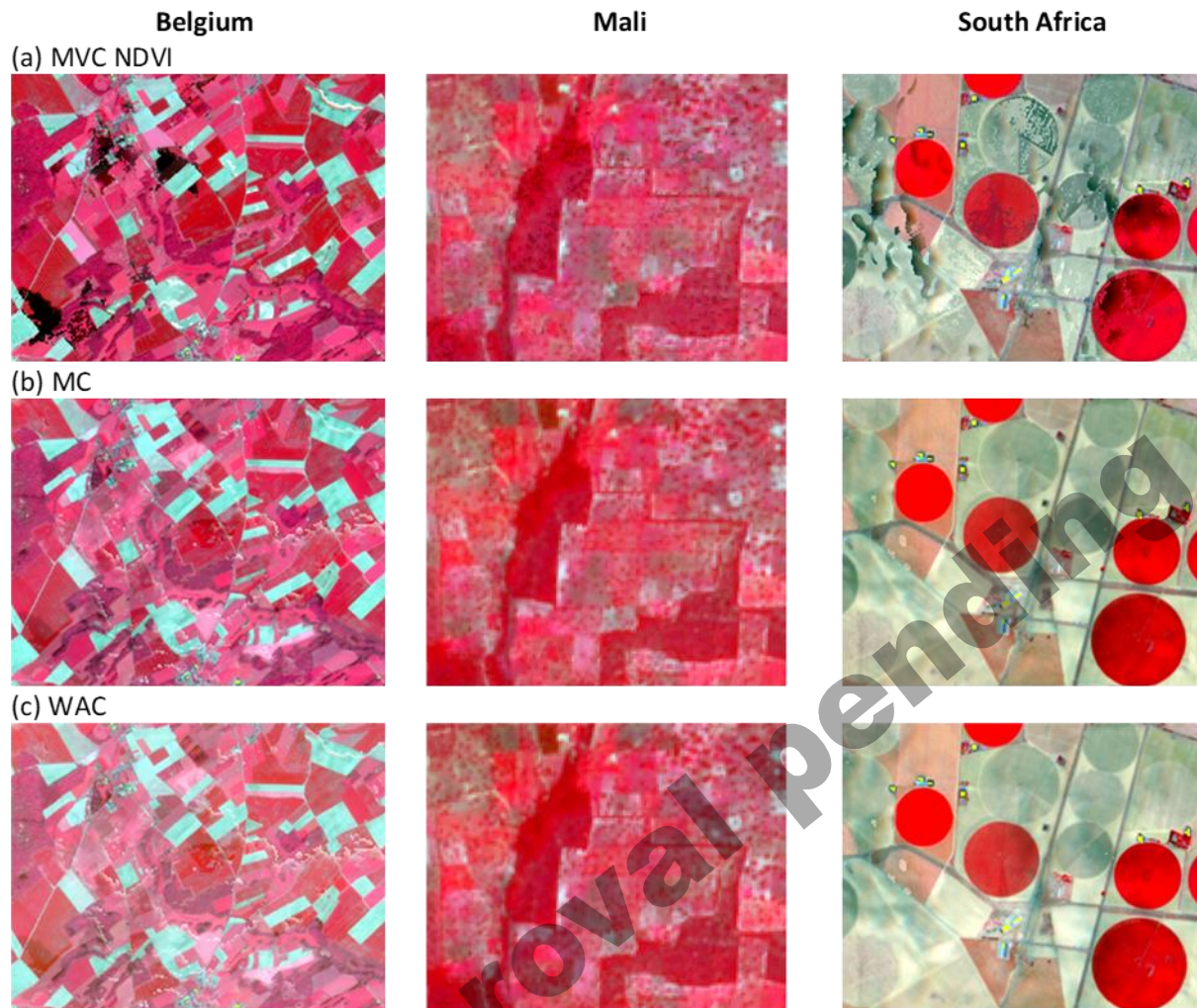


Figure 3-6: False colour (b8, b3, b2) monthly composites over the Belgium site (2017-05), Mali site (2016-10) and South Africa site (2016-10) of the (a) MVC NDVI, (b) MC and (c) WAC algorithms.

However, when zooming at the field scale, such as in Figure 3-6, 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. 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 reflectance. Undetected cloud shadows lead to more artefacts in the MVC NDVI products, while it is smoothed in MC and WAC

composites. 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.

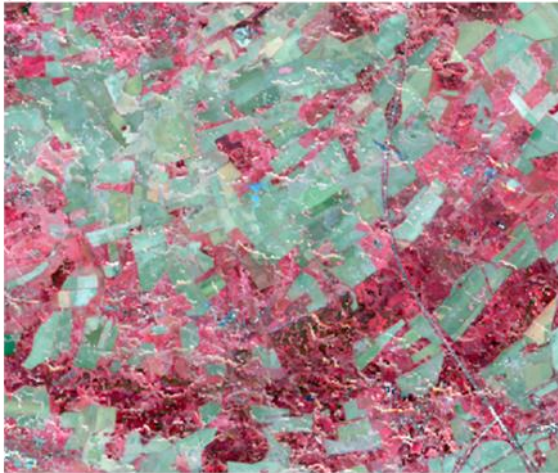
Figure 3-8 and Figure 3-10 show knowledge-based features extracted over the Belgium site and Mali site respectively. They target key phases of the crop cycle such as (a) the bare soil after harvest or before sowing (Maximum Red), (b) the growth rate (Maximum positive NDVI slope), (c) the peak of photosynthetically 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 provides 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



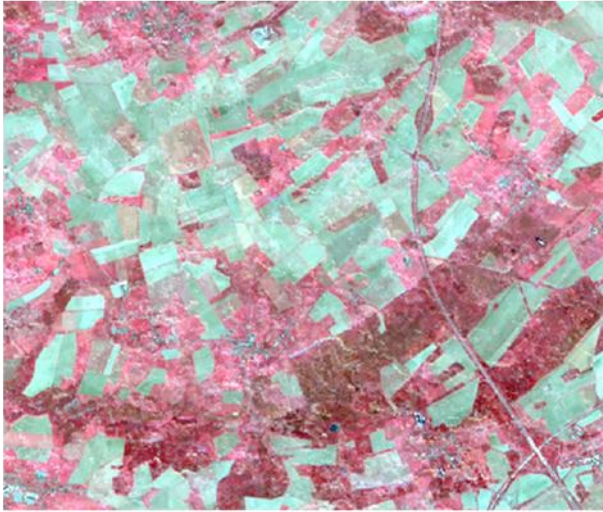
(e) Minimum NDVI



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 (e) Minimum NDVI.

QC - Belgium

(a) Quantile 10



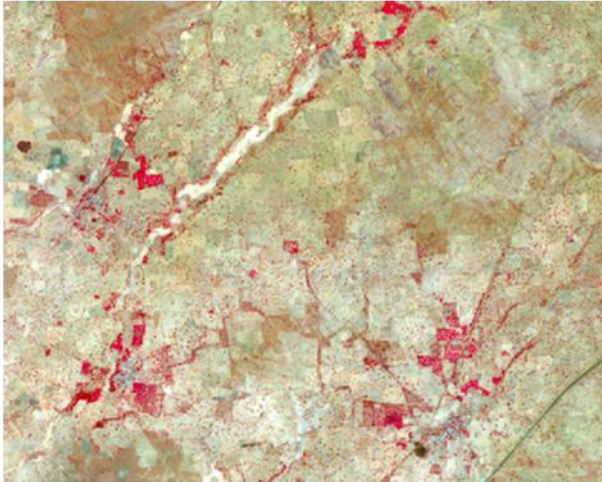
(b) Quantile 90



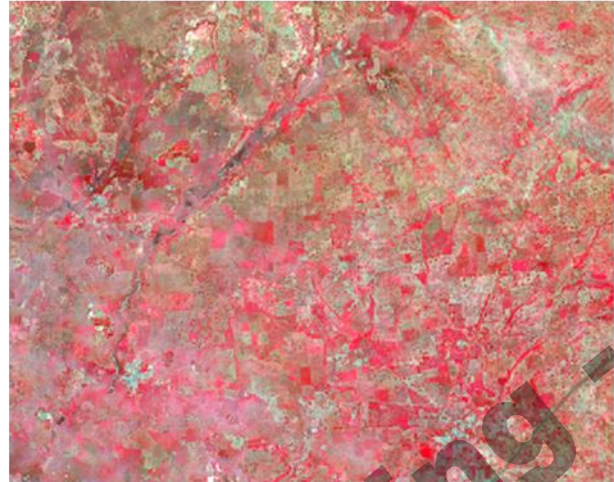
Figure 3-9: False colour (b8, b3, b2) quantile compositing features over the Belgium site: (a) Quantile 10 and (b) Quantile 90.

KC - Mali

(a) Maximum Red



(b) Maximum positive NDVI slope



(c) Maximum NDVI



(d) Maximum negative NDVI slope



(e) Minimum NDVI

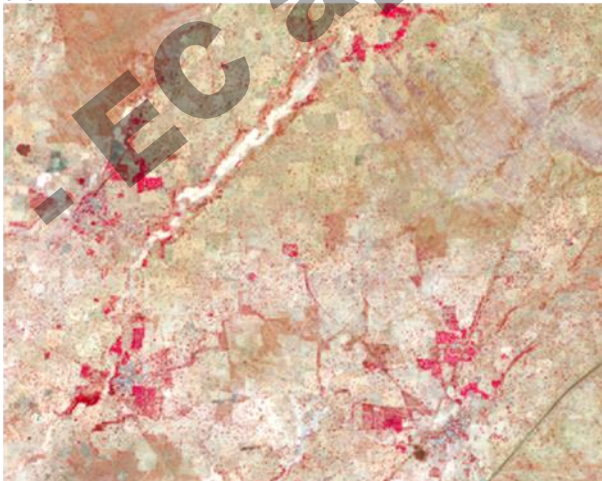
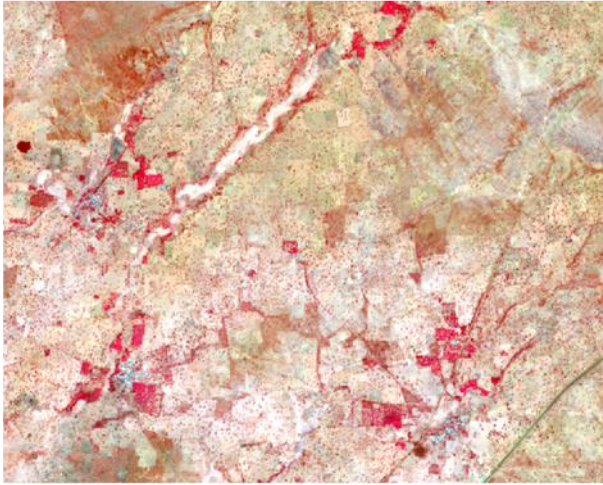


Figure 3-10: False colour (b8, b3, b2) knowledge-based features over the Mali site: (a) Maximum Red, (b) Maximum positive NDVI slope, (c) Maximum NDVI, (d) Maximum negative NDVI slope and (e) Minimum NDVI.

QC - Mali

(a) Quantile 10



(b) Quantile 90

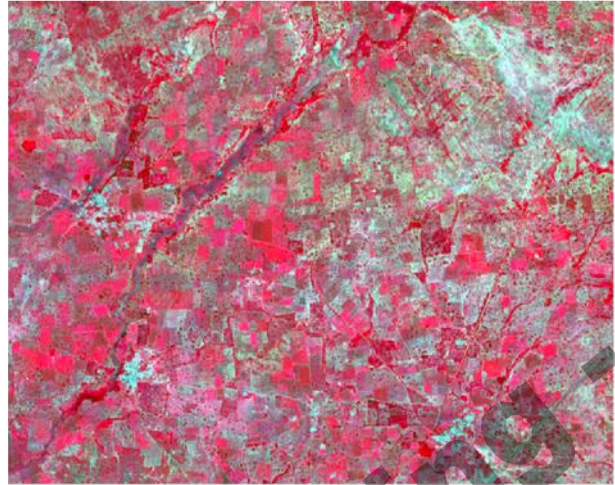
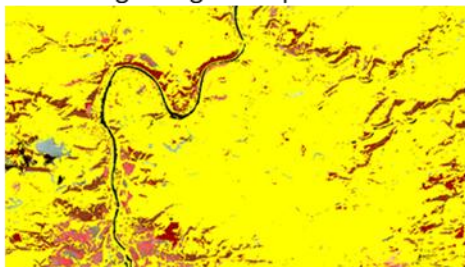


Figure 3-11: False colour (b8, b3, b2) quantile compositing features over the Mali site: (a) Quantile 10 and (b) Quantile 90.

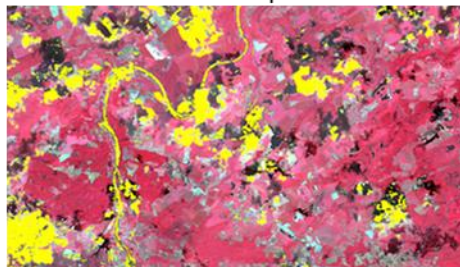
Belgium

(a) MVC NDVI

Beginning of Crop season

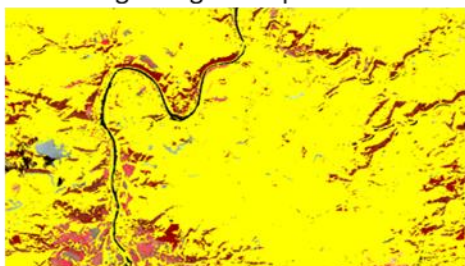


Middle of Crop season

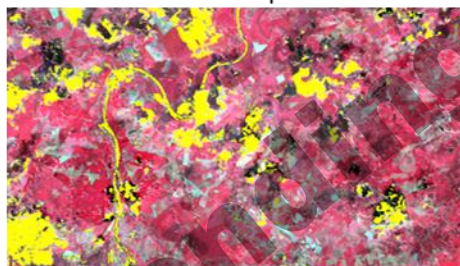


(b) MC

Beginning of Crop season

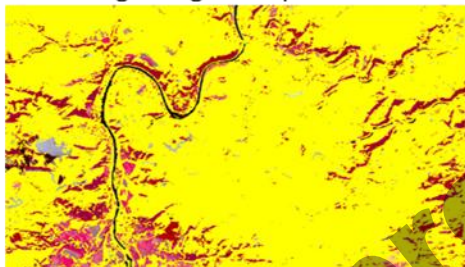


Middle of Crop season



(c) WAC

Beginning of Crop season

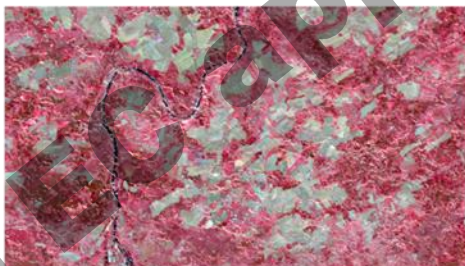


Middle of Crop season



(d) KC

Maximum Red

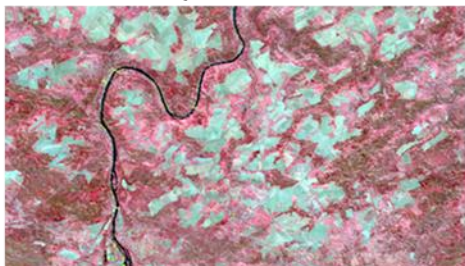


Maximum NDVI



(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).

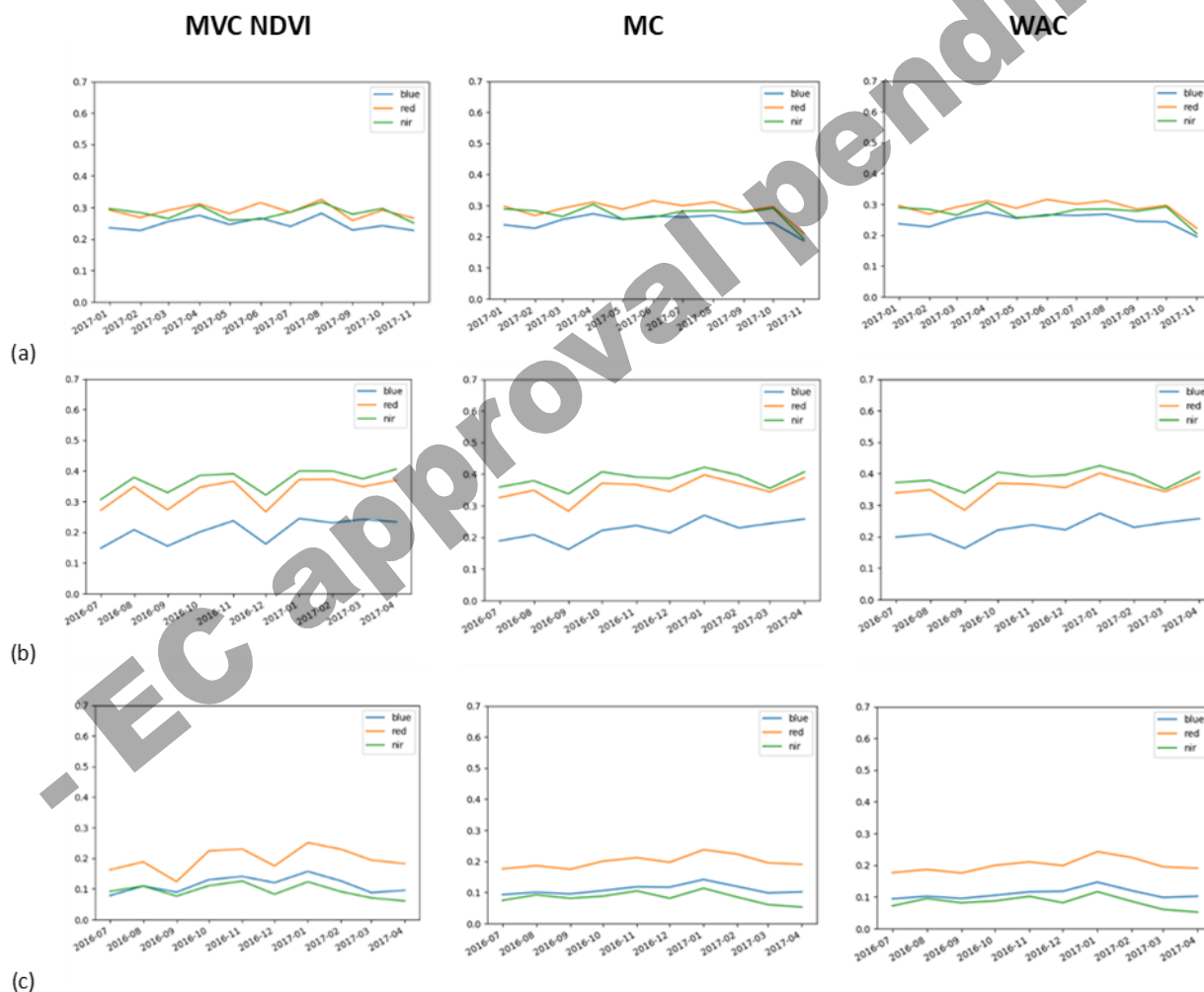


Figure 3-13: Temporal profiles of average surface reflectance for (a) roof top in Belgium, and (b) bare soil and (c) water in South Africa for MVC NDVI, MC and WAC composite time series.

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.

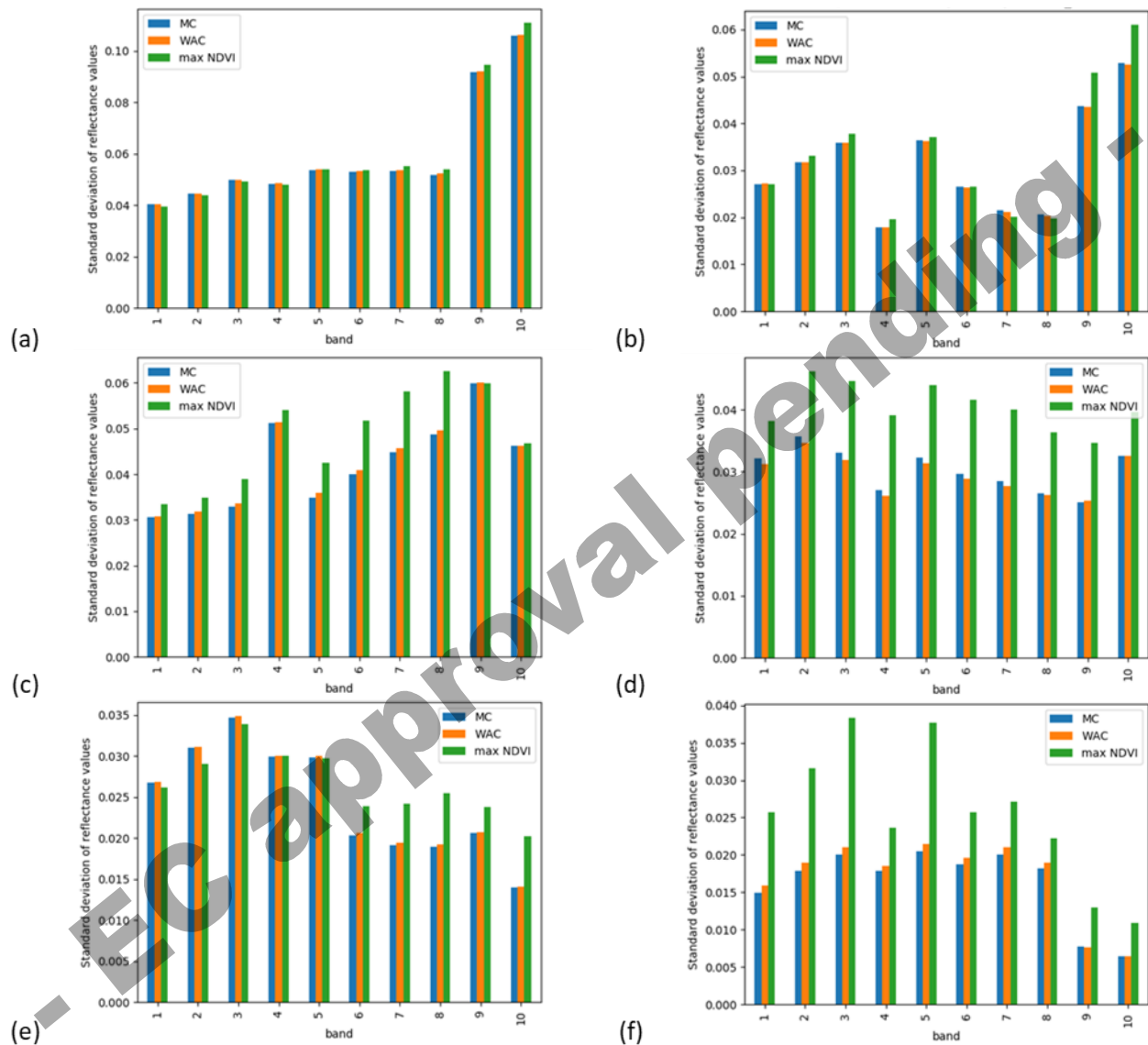


Figure 3-14: Standard deviation of average surface reflectance over roof top in (a) Belgium and (b) Mali, bare soil in (c) Belgium and (d) South Africa, and water in (e) Belgium and (f) South Africa, derived from the three time interval algorithms.

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.

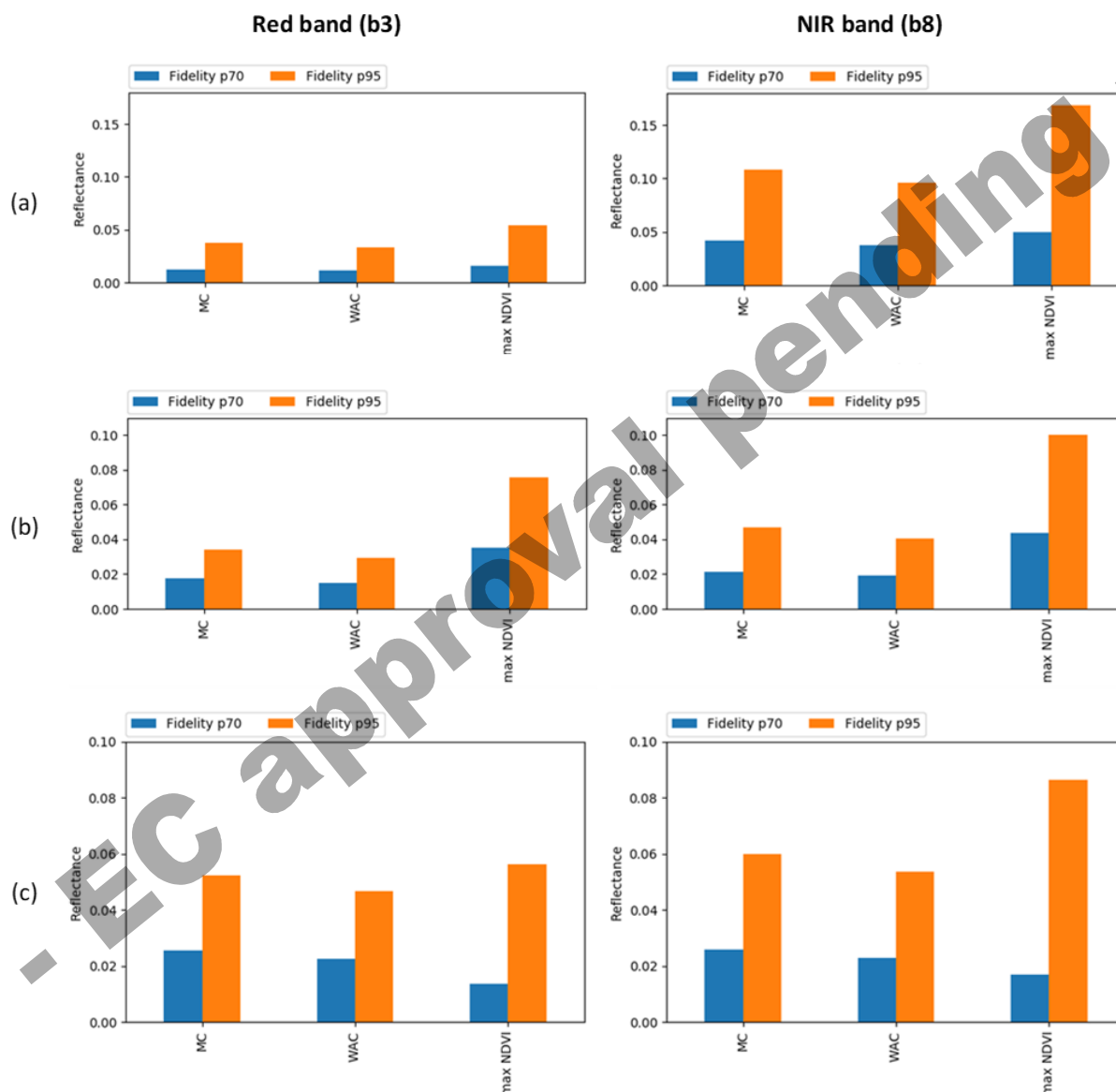


Figure 3-15: Fidelity to central date in the Red and NIR bands for MVC NDVI, MC and WAC for (a) the Belgium site, (b) Mali site and (c) South Africa site.

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.

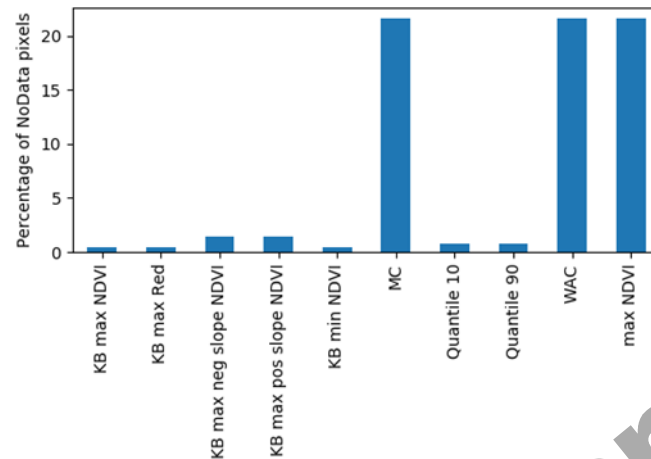


Figure 3-16: Average percentage of data gaps remaining in the composites for the Belgium site.

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.

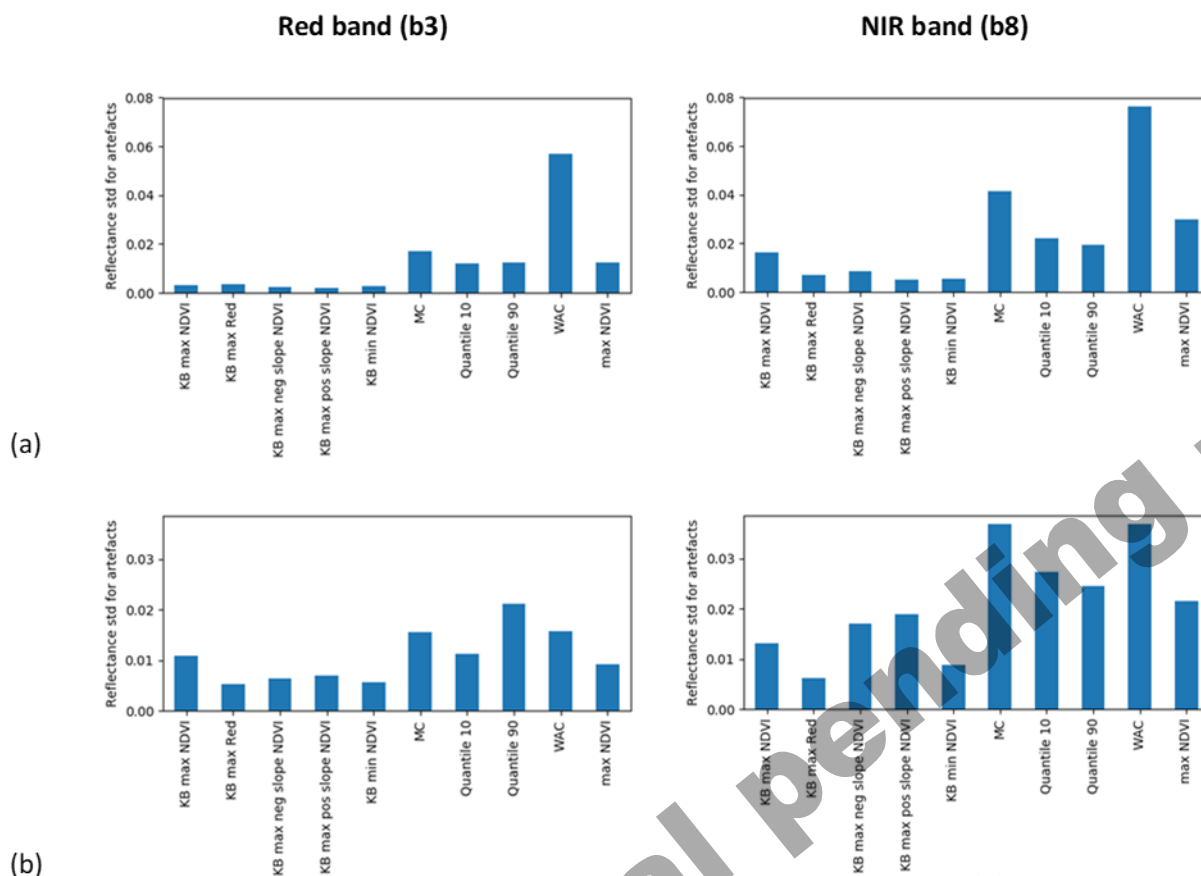


Figure 3-17: Artefacts in the Red and NIR bands for the five selected algorithms for (a) the Belgium site and (b) the Mali site.

3.1.2.4 PROBA-V Compositing

The PROBA-V time series was a good candidate to replace the Sentinel-3 dataset initially planned in this project. The PROBA-V Collection 1 currently provided by VITO corresponds to the most recent reprocessing of the PROBA-V archives completed in 2018 and was expected to improve significantly the cloud and cloud shadow screening. Unfortunately, several artefacts are clearly detected during the compositing process whatever is the compositing period or interval. Indeed, the cloud detection algorithm flags systematically the very bright pixels corresponding to industrial areas, bare soils or beaches.

Figure 3-18 illustrates the MC image for the first half of July showing in white the areas without data due to this cloud detection algorithm. The compositing process further enlarges these areas without data due to the erroneous clouds detection which is systematic but does not flag exactly the same pixels but sometimes the neighbouring pixels as well.

For the sake of demonstration of this shortcoming of the current PROBA-V algorithm, the Figure 3-19 highlights the false cloud and cloud shadow detection induces by the Collection 1 cloud screening algorithm which corresponds to an improvement with regards the previous one. These findings also observed on the 300 m resolution prevent the advanced exploitation of the PROBA-V time series.

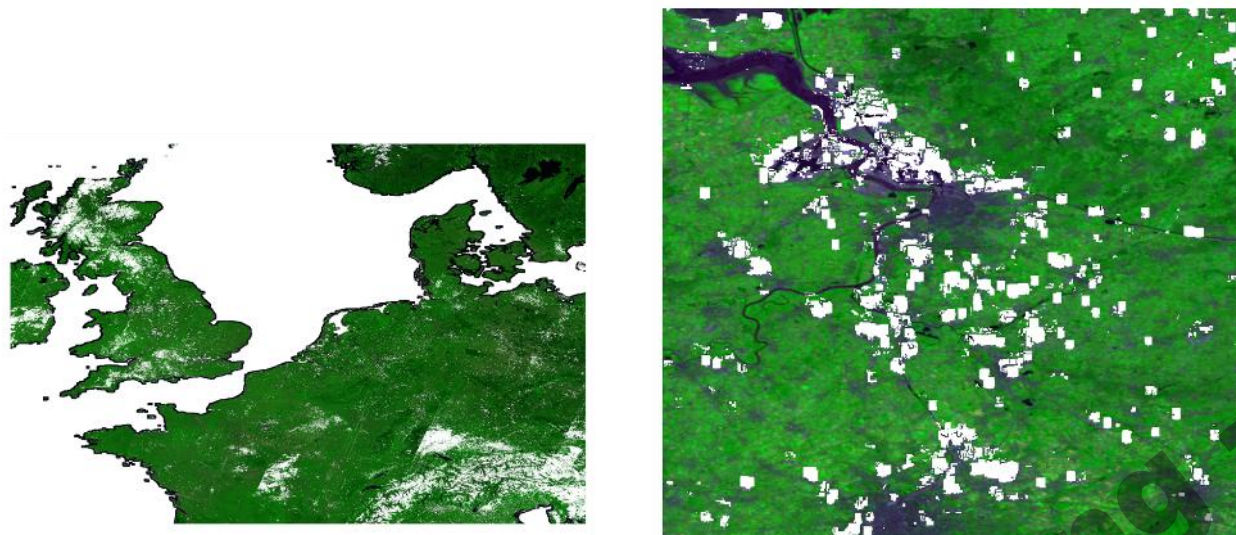


Figure 3-18: PROBA-V 100 Composite from a time series acquired the 1st to the 15th July 2018. The mean compositing was applied on the Collection 1 cloud flag recently reprocessed. The white pixels in the zoom to Antwerpen (Belgium) on the right corresponds to features permanently flagged as cloud.

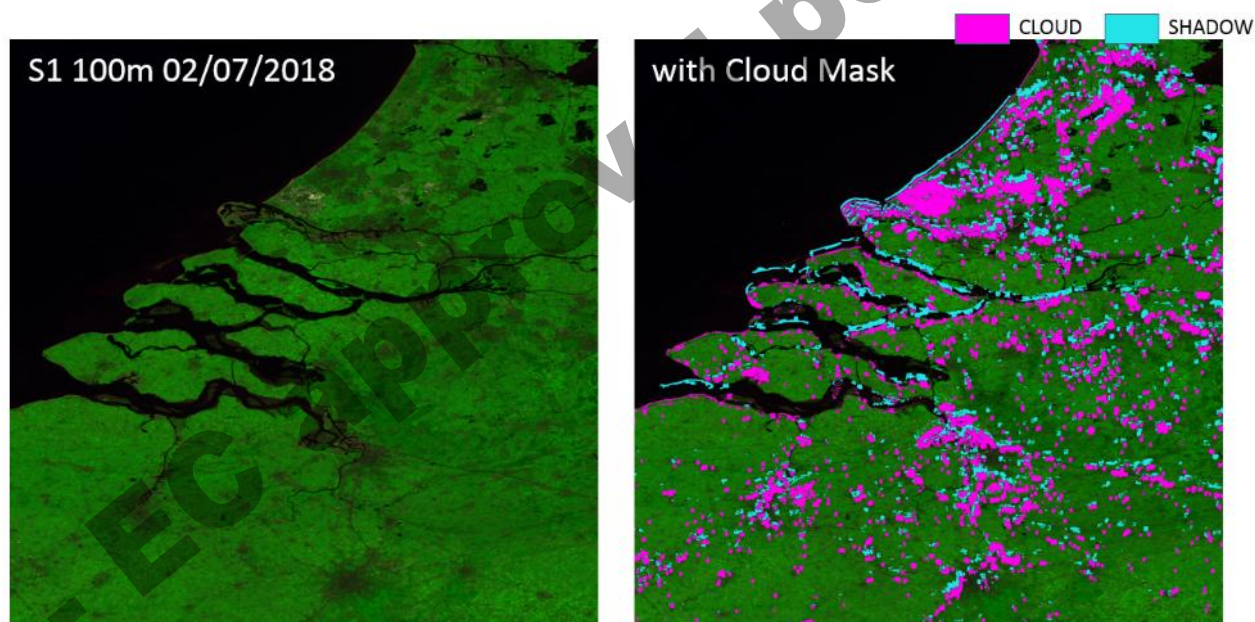


Figure 3-19: PROBA-V cloud free image acquired on the 2nd July 2018 over Belgium and The Netherlands (left image). The corresponding cloud and cloud shadow flags as detected by the Collection 1 algorithm.

3.1.2.5 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 (see [AD07]). 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 D31.1b [AD06]. In phase 1, the focus was set on the following indices, among the most used, the NDVI and the NDWI (also named NDMI).

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). As proved all throughout the tests and demomaterials productions, temporal metrics are decisive to extract most of the time series information.

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), NDRE1 and NDRE2 (Normalized Difference Red Edge Index), MSAVI2 (Modified Soil-adjusted Vegetation Index), SWIRMean (mean of SWIR1 and SWIR2), TCG (Tasseled Cap Greenness), TCB (Tasseled Cap Brightness), Clrededge (Chlorophyll Index Red Edge), and Clgreen (Chlorophyll Index Green) (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. WP31 provides further details of the integration between Sentinel-1 and Sentinel-2 (i.e., optical and radar), plus other sensors, together with info related to derived indices.

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

Sensor	Bands / Indices	Time features
Sentinel-2	<ul style="list-style-type: none"> 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) NDRE1 and NDRE2 (Normalized Difference Red Edge Index) MSAVI2 (Modified Soil-adjusted Vegetation Index) SWIRMean (mean of SWIR1 and SWIR2) TCG (Tasseled Cap Greenness) TCB (Tasseled Cap Brightness) Clrededge (Chlorophyll Index Red Edge) Clgreen (Chlorophyll Index Green) 	min, max, mean, std, p10, p25, p50, p75, p90, pdiff75/25, pdiff90/10, CoV maxmean, activity, difmin3, difmax3, difdif3mean Postrend(NDVI only), negtrend (NDVI only)
Sentinel-1	<ul style="list-style-type: none"> VV (Gamma0) VH (Gamma0) Norm. Difference VV/VH Ratio VV/VH 	min, max, mean, std, p10, p25, p50, p75, p90, pdiff75/25, pdiff90/10, CoV 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: 10th (p10), 25th (p25), 50th (p50), 75th (p75) and 90th (p90), and the differences between the 90th and 10th (pdiff90/10) and 75th and 25th percentiles (pdiff75/27) of the time series.

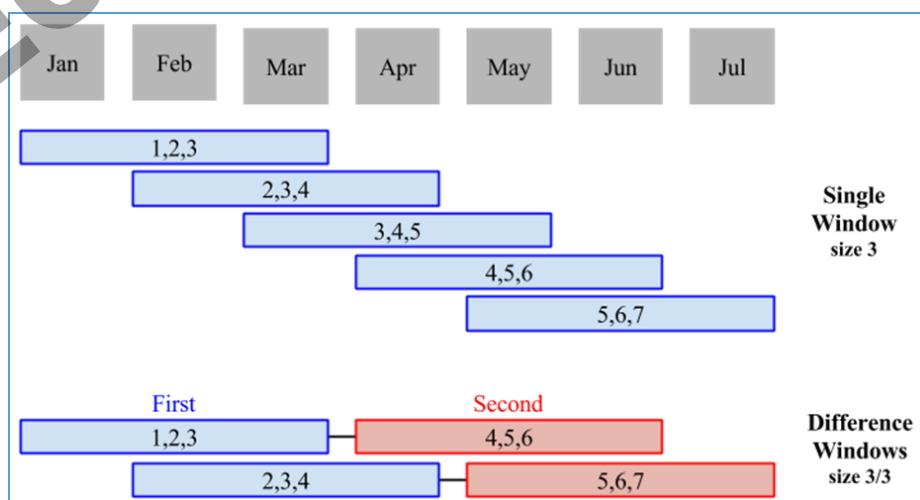


Figure 3-20: Temporal window concept: Single sliding temporal window (e.g. for calculation of mean_max) (top) and difference sliding temporal window configuration (e.g. for calculation of dif_max (bottom)); both examples have a window size of 3 consecutive observations.

The "complex" time features are calculated by the application of a temporal sliding window from the time series stack (Figure 3-20). 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 3-21. 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.

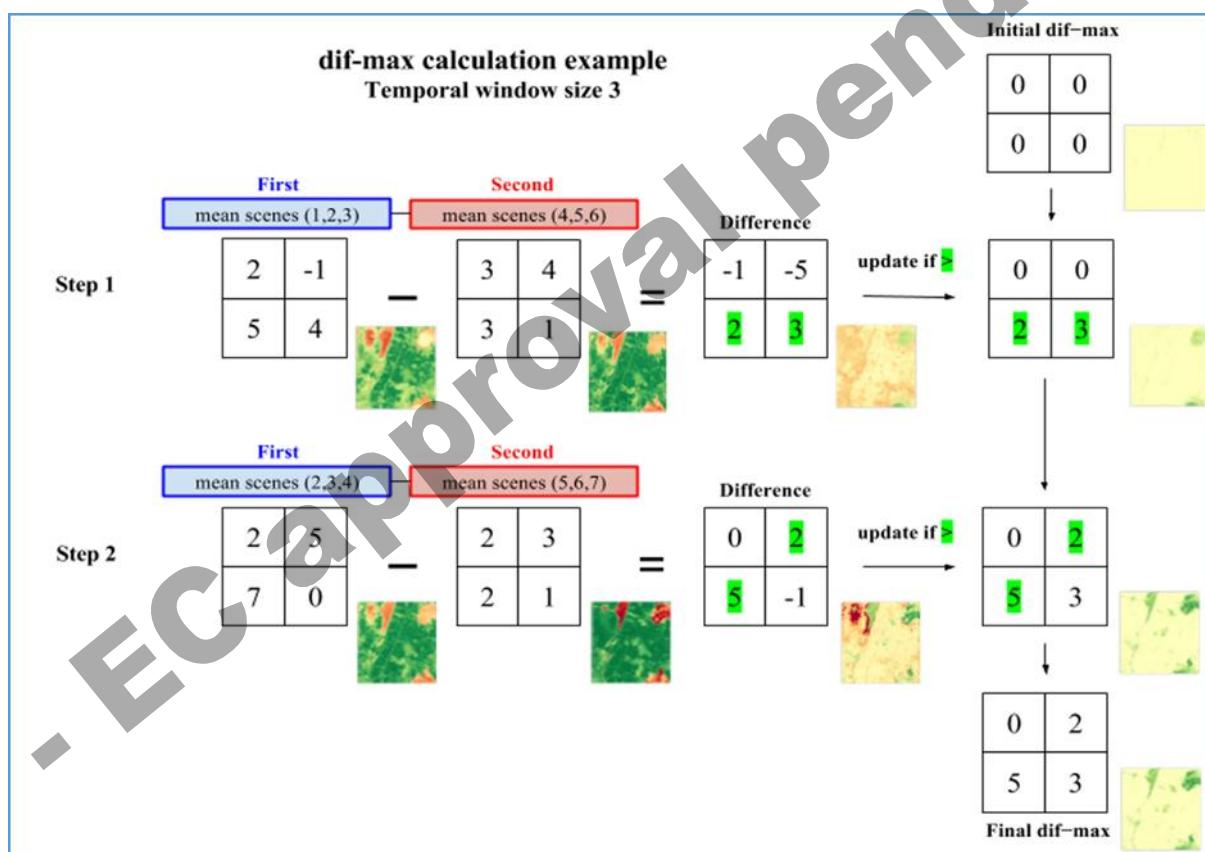


Figure 3-21: Concept of the calculation of a complex time feature shown for the dif_max time feature.

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 3-22) explicitly addresses feature selection before the feature calculation for the full dataset is carried out. In the corresponding sections, benchmarking of computation times versus achieved accuracy during implementation of the tests is applied.

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).

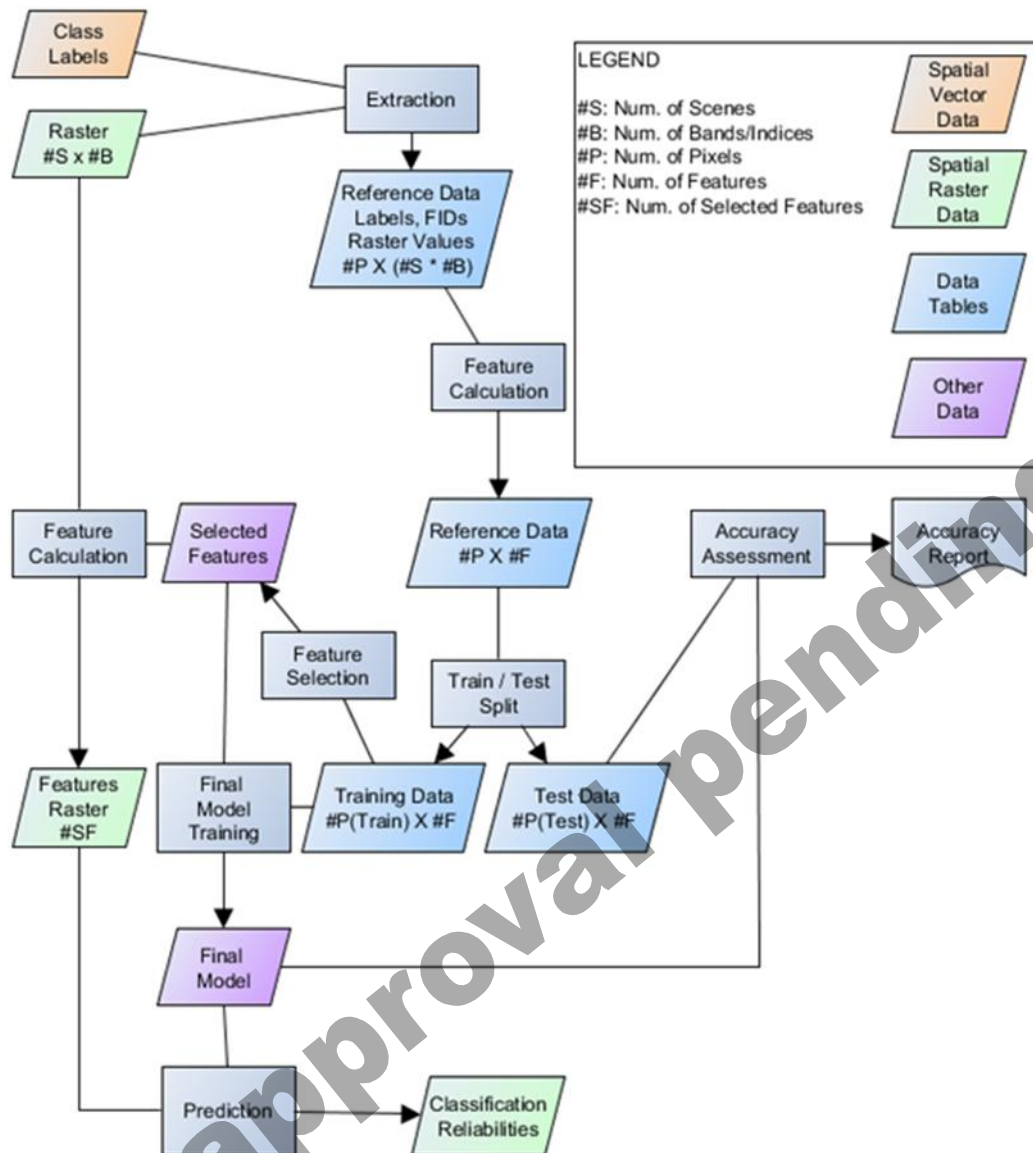


Figure 3-22: Classification workflow.

3.2 Indicators and Variables

This section presents methods developed and applied on the Demonstration site West for determining first generic land cover metrics, then crop growth conditions, and finally multiannual trends and potential changes for the specific changes of land covers, i.e. the HRL Grassland and Forest layers. The prototype dealing with the multiannual trends and potential change detection is based on S-1 time series.

The method for the phenological prototype directly related to agriculture which corresponds to the crop emergence date detection is described also in this section and will be applied on the Demonstration site South Africa.

3.2.1 Method for generic LC metrics

In this section, it is proposed to create phenological products that includes several layers over the West demonstration site. This approach seeks to determine phenological parameters, such as “Phenological Start of Season” (PSS) or a “Maximal Monthly Activity” (MMA), based on robust series of dense multi-temporal images and derived phenological parameters, such as the spectral optical index NDVI. Based on the maximal monthly value of the NDVI, an unsupervised classification with an arbitrary number of classes is launched, in order to regroup pixels exhibiting the same phenological behavior. Parameters such as “Phenological Start of Season” (PSS), “Phenological Peak of Season” (PPS), “Phenological Length of Season” (PLS), are manually detected for each of those classes and the resulting images of the unsupervised classification is reclassified for each of those parameters.

Normalized Difference Vegetation Index or NDVI

The “Normalized Difference Vegetation Index” (Rouse et al., 1974; Tucker, 1979) is used as an indicator to monitor vegetation health and can be used as a proxy for photosynthetic activity, as detailed in WP31. It is calculated as:

$$\frac{\rho_{NIR} - \rho_{Red}}{\rho_{NIR} + \rho_{Red}}$$

By design, the NDVI varies between -1 and 1, where:

- dense vegetation exhibits values between 0.9 and 0.6;
- grasslands or senescing crops gives values between 0.5 and 0.2
- soils are characterized by small positive values usually between 0.1 and 0.2;
- deep water and clouds yield negative values.

The NDVI is widely used to qualitatively detect the presence of vegetation and monitor qualitatively its growth without requiring any further in-situ data.

Temporal feature

For a given pixel of the demonstration site, the maximum value of the NDVI was determined for a selected month, for all the considered years, from 2013 to 2017. When clouds were too present, a fusion of several months has been applied: the winter maximal NDVI has been computed using images from November, December, January and February – and the resulting image has been labeled as ‘months of winter’.

The maximum value of NDVI are then stacked into a multiband image used for the classification.

K-means

The K-means clustering algorithm is a classifier which assumes that features associated with each class are distributed according to a Gaussian distribution. However, this 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. The algorithm performs two steps iteratively in order to reduce the variability within each cluster:

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

The classes identified by the K-means classification, based solely on the spectral signature of their pixels, can then be associated with a type of LU to produce the map. For the K-means to deliver those classes, a given amount of them has to be given as a parameter.

3.2.2 Method for crop growth condition

The growing condition of any crop can be assessed by the trajectory of the Leaf Area Index (LAI) when the LAI can be observed on a regular basis. Unlike the NDVI, the LAI is a biophysical variable which can be retrieved by various sensors. The LAI is here not only defined by the half of the leaves area as commonly accepted but rather by the Green Area Index (GAI). The GAI is indeed a more appropriate term when working with cereals because the main aerial organs (leaves, ears and stems) are photosynthetically active. For the sake of clarity, the LAI acronym will however be used while this means GAI. The LAI retrieval is based on the BVnet algorithm using artificial neuronal network trained on simulated LAI and reflectance values. The reflectance values are simulated using the ProSail radiative transfer model for the Sentinel-2 bands at 10 m and 20 m-resolution except the blue band (B2) and the B8 due to its overlap with B7 and B8a.

The retrieved LAI values are averaged over the entire field minus an inner buffer of 2 pixels from the field boundaries. Then the Whittaker smoother is applied on the LAI time series assuming a continuous evolution of the LAI.

Based on the Land Parcel Identification System which is available on the crop type (either from the crop map obtained in WP4.4 or from the LPIS whenever available), the LAI values of all the fields of the same crop located within a radius of 3 km far from the field border are averaged along the season. That is mainly to compare the crop growth condition for any given field of interest. The average does not include the field of interest and is not available when no field of the crop of interest are grown within the 3 km radius. The 3 km radius was found relevant because of the similarity of external factors, typically the meteorological conditions and the agro-climatic zone. The average LAI profile of the crop of interest and the LAI profile of the field of interest can then be visually compared in terms of crop development (earliness, maximum, maturity, etc.). Both profiles are also quantitatively assessed through a simple metric corresponding to area under the curve for three different crops.

3.2.3 Method for multiannual trends and potential changes based on SAR data from S-1

In this section, an approach to describe multiannual trends and potential changes for the specific HRL layers forest and grassland is explained into detail. The approach relies on the idea that based on a series of multi-temporal images for a given study area, the remotely sensed temporal dynamics of a specific HRL class are sensibly different to those of all other classes. For instance, in the case of radar data the backscattering temporal mean of urban areas (due to double bounce reflection) is higher than that of forest areas (which might result in high backscattering in one/few acquisitions due to specific conditions, but in general exhibit lower values). Further general assumptions of this approach are that (1) a specific class of the HRLs might change from one year to another and (2) that the classes within the HRL are homogenous at the local to regional scale of the test sites despite having different characteristics at the pan-European scale. Considering these assumptions, the following method is based on calculating statistical distributions for different seasonal and annual metrics derived from Sentinel-1 time series data (see section 3.2.3.1) for each class of the HRL Forest (two classes: broadleaved and coniferous) and HRL Grassland (one class: grassland (GRA)) within the demo site. In the next step, potential changes within the HRL are detected at pixel-level, based on these seasonal and annual metrics. Here, a statistical test is applied describing if a specific pixel belongs to the considered class at a certain significance level. Pixels

identified as not belonging to the class are labelled as candidates for a HRL update. Therefore, pixels within a certain HRL are assumed to have similar backscatter values over time and space. This might become more critical when the considered spatial extent covers larger areas and different biogeographical zones.

The approach comprises two steps, which are described into detail in the following.

3.2.3.1 Pre-processing on the input data

This section describes the pre-processing and preparation of the input data.

Creating the master HRL

Primarily, the HRLs require harmonization with respect to a) their original spatial resolution of 20m to the resolution of the Sentinel data (i.e., 10m) and b) their spatial coverage (adaptation to the Sentinel-2 tile system). Figure 3-23 shows the harmonized HRL Grassland of 2015 for the Belgium test site. The harmonized HRL Grassland describes the situation at time t_0 (in this case 2015).

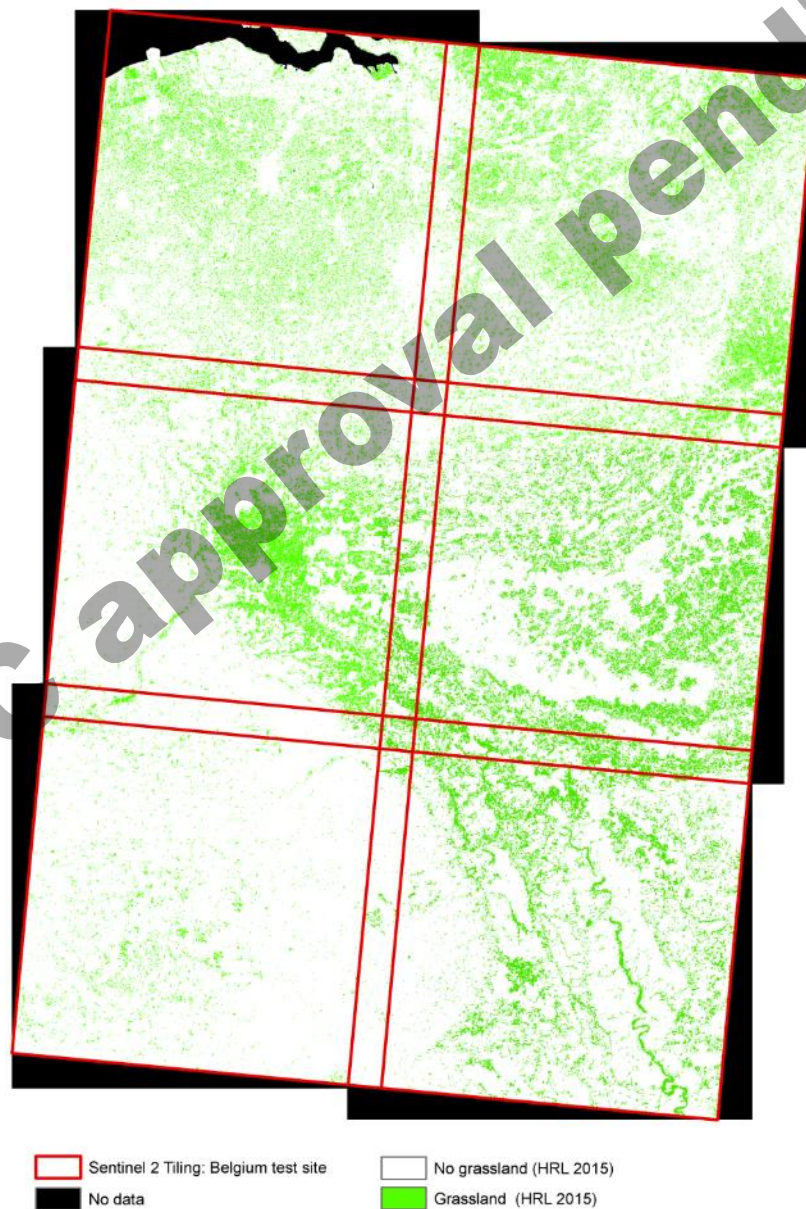


Figure 3-23: Sentinel-2 tiles for Belgium test site and harmonized grassland HRL of 2015

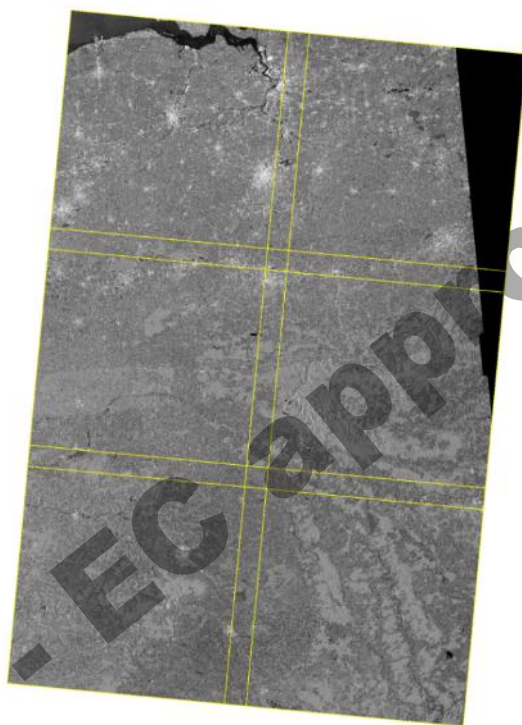
S-1 preprocessing and feature extraction

S1 IW GRDH data acquired both in ascending and descending pass from 2015 to 2018 was pre-processed by means of the S1TBX/SNAP software as described in AD07 Methods Compendium: Time Series Preparation. Specifically, this included the usual steps of orbit correction, thermal noise removal, radiometric calibration, Range-Doppler terrain correction and conversion to dB values.

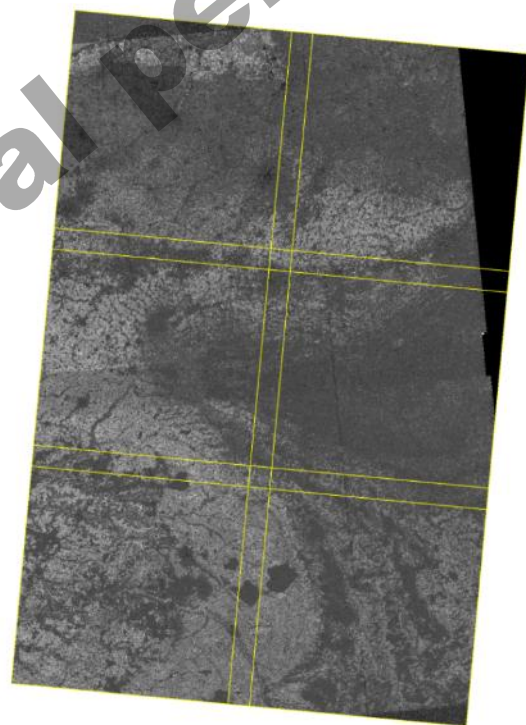
After pre-processing all the available data for the given region, four key temporal statistics have been extracted for each pixel for the seasons a) March, April, May (MAM), b) June, July, August (JJA), c) September, October, November (SON) as well as separately for the complete years 2015, 2016, and 2017, namely:

- backscattering temporal maximum;
- backscattering temporal minimum;
- backscattering temporal mean;
- backscattering temporal standard deviation;

Ascending and descending orbits were considered separately due to the strong influence of the viewing geometries on the backscattering process. Figure 3-24 shows examples of the backscattering temporal statistics of 2015 over the Belgium demo site.



(a) backscattering temporal mean



(b) backscattering temporal standard deviation

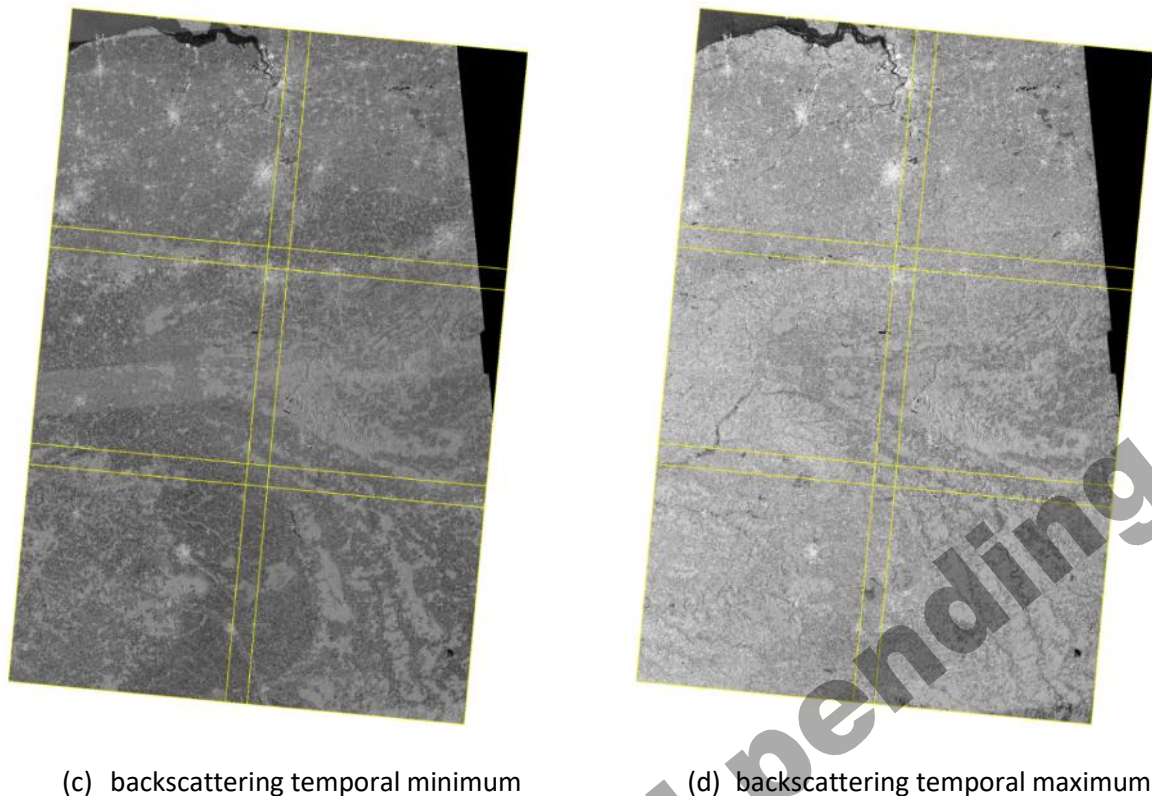


Figure 3-24: Examples of temporal S-1 metrics: Backscattering temporal statistics of 2015 of the S-1 relative orbit 161 in ascending pass over the BELGIUM demo site: (a) backscattering temporal mean, (b) backscattering temporal standard deviation, (c) backscattering temporal minimum, (d) backscattering temporal maximum (outlines of S-2 granules in yellow)

Reducing time series data to seasonal and annual metrics has the advantage of enabling a fast and effective processing without noise influence. The preference of using metrics instead of single observations for land classifications is widely applied because of its ability of characterizing certain land cover classes with such features (Huettich et al., 2009).

3.2.3.2 Statistical analysis of seasonal and annual metrics within the HRL classes to identify potential change

The approach relies on the basic assumption that all pixels of the considered HRL class have similar characteristics in the feature space (described in section 4.3.1) which considerably differ from all other classes. Differences between the feature values of a pixel and their distribution for a certain class can be used as an indicator for change. Therefore, three steps have been implemented to identify pixels which might need an update.

1. Firstly, the statistical mean and standard deviation of all available S-1 backscatter metrics were calculated for each HRL to derive the general behaviour of that specific land cover class.
2. Based on these statistics, the distance in the feature space between each pixel and the characteristic class mean was calculated.
3. The distance of pixel values to the class mean was discretized into three “potential change” categories:

- a. The 1st category suggests that the corresponding pixel does not differ from the class mean and thus is considered as stable with no change:

$$Distance_{stable} : \text{for pixel} < Class_{mean} \pm Class_{stddev}$$

- b. The 2nd category implicates that the pixel is placed between \pm two standard deviations:

$$Distance_{1stddev} : \text{for pixel} < Class_{mean} \pm 2 * Class_{stddev} \text{ AND } > Class_{mean} \pm Class_{stddev}$$

- c. The 3rd category implicates that the pixel is out of the range of \pm two standard deviations

$$Distance_{2stddev} : \text{for pixel} > Class_{mean} \pm 2 * Class_{stddev}$$

4. In a fourth step, the results for each feature were combined to a final change plausibility estimate, which is robust against infrequent outliers. The approach considers the distance classes throughout all features of one year and calculates the frequency of one pixel belonging to one of the three categories defined in the second step. By using all metrics and making a decision based on the cumulative analysis of each pixel's distances robust change estimates could be derived.

These change plausibility frequencies may subsequently be discretised further with class-specific thresholds to highlight only change candidates at the pixel level. Empirical testing suggests a suitable thresholding in that either a) less than 50% of considered metrics belong to category $Distance_{stable}$ or b) more than 33% of considered metrics belong to category $Distance_{2stddev}$.

By applying these two thresholds, all pixels featuring a higher distance to the statistical mean of one particular HRL class are labelled as a potential pixel for update. A flow-chart of the entire workflow is presented in Figure 3-25.

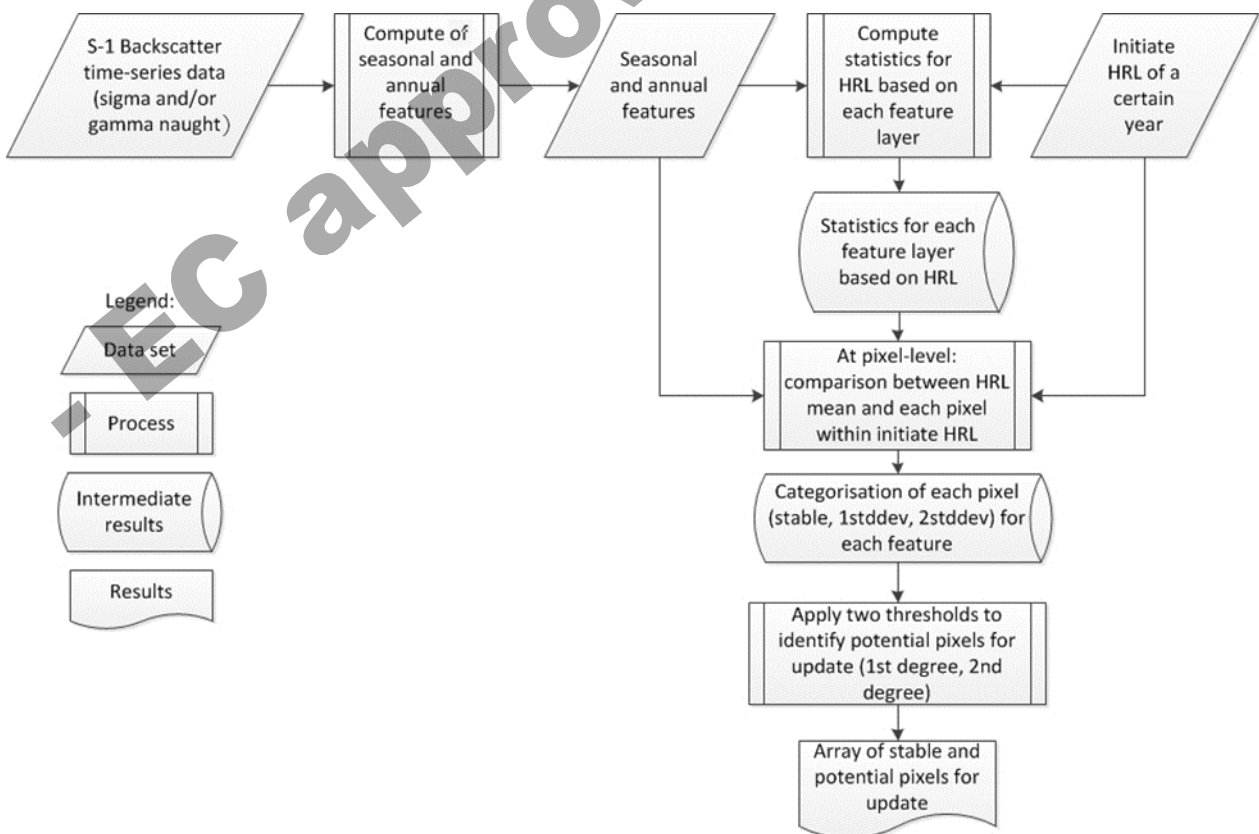


Figure 3-25: Workflow for identification of potential pixels for update within a certain HRL

3.2.4 Methods for emergence date detection

The crop emergence date is a prototype which requires method developments which are specifically targeted to this phenological parameter. Therefore several methods are described to be tested on a calibration dataset. The challenge of the emergence date detection in an agriculture context is to deliver the information as early as possible in the season. Unlike most methods for phenological parameters retrieval, this means that the phenological analysis take place much before the completion of the full growing cycle.

3.2.4.1 Emergence date as phenological parameter

The estimation of the emergence date is based on the detection of the first phenological stages (crop emergence, first leaf development) from satellite remote sensing (i.e. Sentinel-2) time-series. The aim is to provide early-season information, to allow an easy operational implementation and further generalization of the method. The detection is performed at the field level on early-season satellite images which presents a clear advantage against detection methods requiring full season images to provide initial results.

The phenological stages are commonly defined according various classification systems such as the Biologische Bundesanstalt and Bundessortenamt und Chemische Industrie (BBCH) scale. The standard satges are selected to provide a continuous scale ranging from 0 to 100 which is relevant to any crop and location.

In particular, a specific BBCH stage is reached when at least 50% of the plants are within the definition of that stage (Lopez-Sanchez et al., 2012). The onset date of the stages presents an important inter-annual variation. It can be explained by environmental and climatic factors as well as farm-level management decisions (crop variety, crop rotation, input availability, etc.) (Sakamoto et al., 2010). Generally, temperature and water are the main climatic factors impacting the development of the majority of species. A large number of species is also impacted by the length of the photoperiod. In temperate climates, light is generally the primary limiting growth factor. In humid climates, light and nutrients are both limiting. In tropical or dry subtropical climates, water is the main constraint but the absorption of nutrient is also reduced.

3.2.4.2 VIs and hue time series as candidate data sources

Vegetation Indices (VIs) can be seen as a proxy of the Fraction of Absorbed Photosynthetically Active Radiation (FAPAR) as they relate "greenness" with the measure of the absorption characteristics of the vegetation in the red and NIR spectral bands. NDVI first relates to total green biomass and is sensitive to low to moderate LAI values but saturates at high values (Nguy-Robertson et al., 2014). VIs do not present a straightforward biophysical interpretation, although they are strongly correlated with biophysical variables (White et al., 1997; Eklundh et al., 2003). LAI retrieval algorithm can also be considered but the underlying assumption and the higher computing cost of such an algorithm prevent considering it at this stage.

Studies showed that the NDVI gives good estimates of the vegetation dynamics when the vegetation is photosynthetically active (Palacios-Orueta et al., 2012). Other VIs based on the Short-Wavelength Infrared (SWIR) are better for assessing low vegetation density zones. In semi-arid areas, the information contained in both the Mid-Infrared (MIR) and blue regions relates to soil properties which helps in distinguishing vegetation type classes (Hüttich et al., 2009).

Pekel et al. (2011) studied the detection of green vegetation in semi-dry and dry areas. The image color is transformed from a Red-Green-Blue (RGB) to a Hue-Saturation-Value (HSV) representation. As the MIR and NIR regions present several advantages for soil discrimination, the three MIR-NIR-Red bands are used instead of RGB. The Hue component is the only parameter conserved because it is able to discriminate land cover type where Saturation and Value fail. Marinho et al. (2014) applied the method

developed by Pekel et al. (2011) and tested it for sowing date estimation which had not been investigated yet. The study aimed at estimating green-up onset dates in arid and semi-arid regions (i.e. the Sahel region) from MODIS 250 m resolution images and RFE 8 km resolution rainfall estimate and, then, comparing it with ground-truth data.

To reduce the noise and fill the gap in the time series, two interpolation methods are tested. First, a simple linear interpolation method is applied between all the successive observations, between October and the end of April. Alternatively, a logistic interpolation method commonly used to reduce noise in the vegetation profiles is also selected to interpolate the satellite observations during plant growth (i.e. between minimum and maximum index values): a four-parameter logistic function.

3.2.4.3 Candidate detection methods

The candidate methods for phenology studies are often grouped into four main categories: threshold, moving window, function fitting and model fitting methods (Zeng et al., 2016; de Beurs and Henebry, 2010).

(1) The threshold method (Figure 3-26a) is based on linking a phenological event with the crossing of a certain value of the VI curve. The threshold can either be fixed or dynamic and varies with land cover, soil background, view and solar angle (Reed et al., 1994). However, they do not rely on an underlying biophysical meaning. For example, White et al. (1997) identified the onset and end of greenness when the NDVI ratio of the smoothed curve exceeds or falls below 0.5 respectively. Lobell et al. (2013) defined the green up phase as the date when the double-logistic fitted function exceeds 10% of the year's maximum amplitude. The main drawback of the fixed threshold is the disability for reflecting the spatial changes of larger study area and inconsistency for a wide variety of land covers (de Beurs and Henebry, 2010; Reed et al., 1994). Plethora of thresholds have been used based on the long-term VIs mean, yearly VIs, NDVI ratios, Normalized Difference Water Index (NDWI), etc. The ratio approach has the advantage of being independent from the geographic location and land cover of the area. As such, the NDWI is particularly indicated for heavily snowed areas.

(2) The moving window method can be derivative or backward-looking moving average. The derivative method (Figure 3-26b) is founded on the assumption that the fastest green-up or greatest leaf expansion corresponds to the most ecologically relevant SOS (White et al., 1997). In other words, the maximal increase and decrease of NDVI tally with SOS and EOS (de Beurs and Henebry, 2010). Moving windows of a certain temporal extension are applied on each pixel and the slope (or derivative) is calculated. The highest positive and lowest negative derivatives are then easily extracted.

Some methods retrieve the second derivative and determine the SOS as the time point combining a positive slope and a local maximum. Cong et al. (2013) defined green-up onset date as the highest positive relative change of the average NDVI time-series for a 15-day moving window. Moulin et al. (1997) identified the beginning of the vegetation cycle (b_date) on three conditions: (i) NDVI value at b_date is close to a bare soil value, (ii) left derivative (before b_date) should be equal to zero because NDVI is assumed constant before the growth season, (iii) right derivative (after b_date) should be positive on two weeks' time window. de Beurs and Henebry (2010) reported that this method gives good results where the NDVI curve displays a sharp increase and a steep decrease.

The backward-looking moving average method identifies the onset of greenness as the date when the VI curve crosses the moving average function which represents a significant change in the growth trend. The moving average is built as the average of the last i observations. The choice of the temporal window (i.e. number of i observations) is crucial and arbitrary as it introduces a time lag: a large time interval is less sensitive whereas a small interval may take insignificant variations into account (Reed et al., 1994; de Beurs and Henebry, 2010).

These two first methods present the advantage of being able to retrieve multiple growing seasons (Verhegghen, 2013). However, they are not good at distinguishing the basic temporal variations of the vegetation reflectance (noise) from the relevant seasonal changes. That, the date retrievals based on local minima, maxima, or fixed thresholds can be completely shifted if observation errors contaminate the original dataset. For instance, atmospheric constituents, bi-directional reflectance distribution function, cloud coverage, and the mixed-pixel effect often influence MODIS images (Sakamoto et al., 2010).

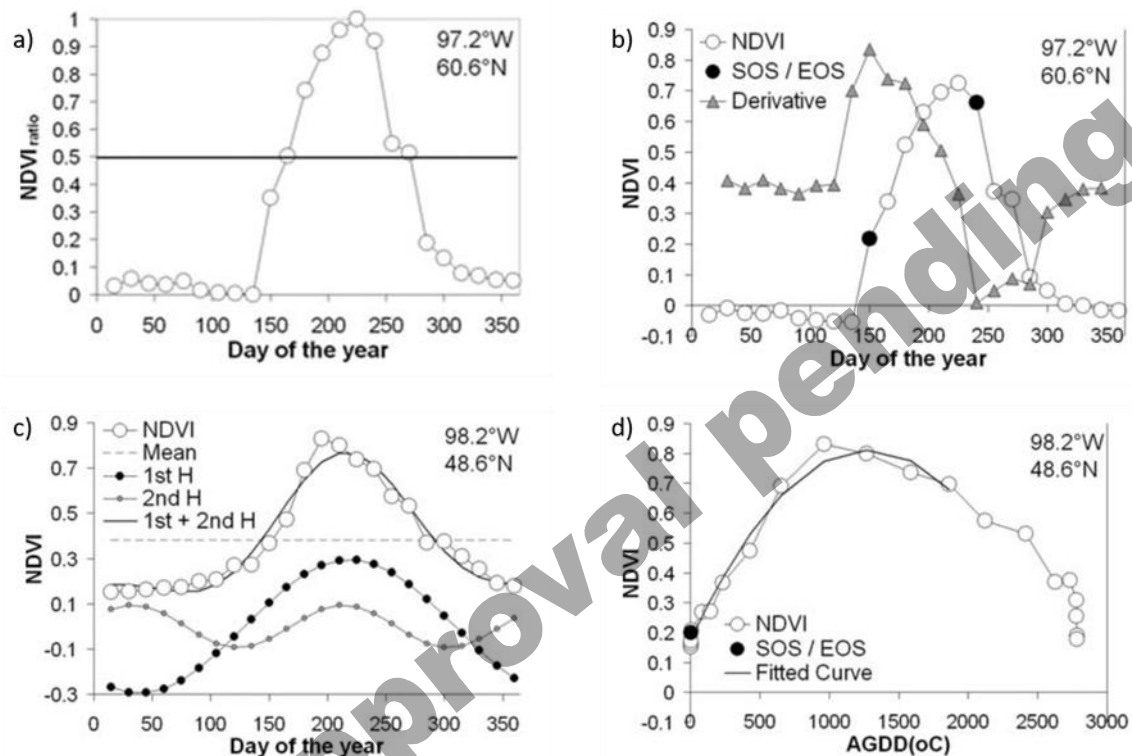


Figure 3-26: Four methods based on the NDVI to detect start and end of the season. a) fixed threshold, b) derivative, c) Fourier transform, d) quadratic fitting based on AGDD (de Beurs and Henebry, 2010)

(3) The function fitting method (Figure 3-26c) applies a mathematical function to a given VI curve to smooth or filter the data and extract the main information. Different functions have proved to be useful: Fourier analysis, wavelet transformation, Principal Component Analysis (PCA), Canopy Structure Dynamic Model (CSDM), etc. The Fourier analysis which decomposes a complicated curve into a sum of sinusoidal waves, is able to approximate a VI (de Beurs and Henebry, 2010). This segmentation is sensitive to systematic changes and reduces the non-systematic data noise. To interpret the new curve, the first Fourier harmonic is considered to represent the mean NDVI. The wavelet transform also decomposes the VI time-series into a set of small local waves (named wavelets) assuming the fact that the noise components have higher frequencies than the main seasonal changes (Sakamoto et al., 2010). An important aspect is that this type of frequency decomposition performs better on long time-series showing periodic changes. Consequently, the source observations should be measured at a regular time interval or require gap filling to be adequately processed (de Beurs and Henebry, 2010).

To retrieve phenological events from a fitted curve, the procedure of Sakamoto et al. (2005) can be used: the minimal or inflection point earlier than 60 days before the maximum value (defined as heading date) is selected, then, the later of the two points is identified as planting date. However, the Root Mean Square Error (RMSE) of 12.1 days for planting date estimate is not satisfactory. Another way to account

for key information lies on the PCA. Through a linear combination of the original observations, the primary factors explaining the main variance of the dataset are kept. Again, the interpretation of the resulted parameters (eigenvectors) is not self-evident and does not remain consistent over the years limiting the comparative power of the method (de Beurs and Henebry, 2010). The advantage of those fitting methods is to reduce noise and adjacency pixel problem (between pixel effects) and their ability to derive phenological metrics in a consistent way (Palacios-Orueta et al., 2012).

The model fitting method fits a model to the remote sensed observations. These models can be simple (logistic models, etc.) or more complex (Gaussian Local Functions, etc.) and are previously defined or dynamically built with input data (de Beurs and Henebry, 2010; Zeng et al., 2016). The number of input parameters compared to the amount of observations available for their identification and the need of large-scale ground-truth data is crucial when assessing the scope and implementation of these model fitting (Duchemin et al., 2008).

Accumulated Growing Degree Days (AGDD) can be interpreted as a measure of the accumulated heat above a specified base temperature from the beginning of the season: maize base temperature is estimated around 10°C. Modelling vegetation growth under AGDD instead of anthropocentric calendar time (Figure 3-26d) is more relevant especially during the first half of the growing season when day length and water stress are not the main contributors yet (de Beurs and Henebry, 2010).

These four methods can be combined. Threshold methods are generally applied on smoothed function to reduce data noise.

Then, a logistic function was applied to each identified increasing or decreasing period and key phenological dates were then retrieved from the fitted curve. For a single growth cycle, the following logistic function modelled the curve:

$$y(t) = \frac{c}{1 + e^{a+bt}} + d$$

Where t is time in days, $y(t)$ is the VI value at time t , a and b are fitting parameters, $c + d$ is the maximum VI value, and d is the initial background VI value.

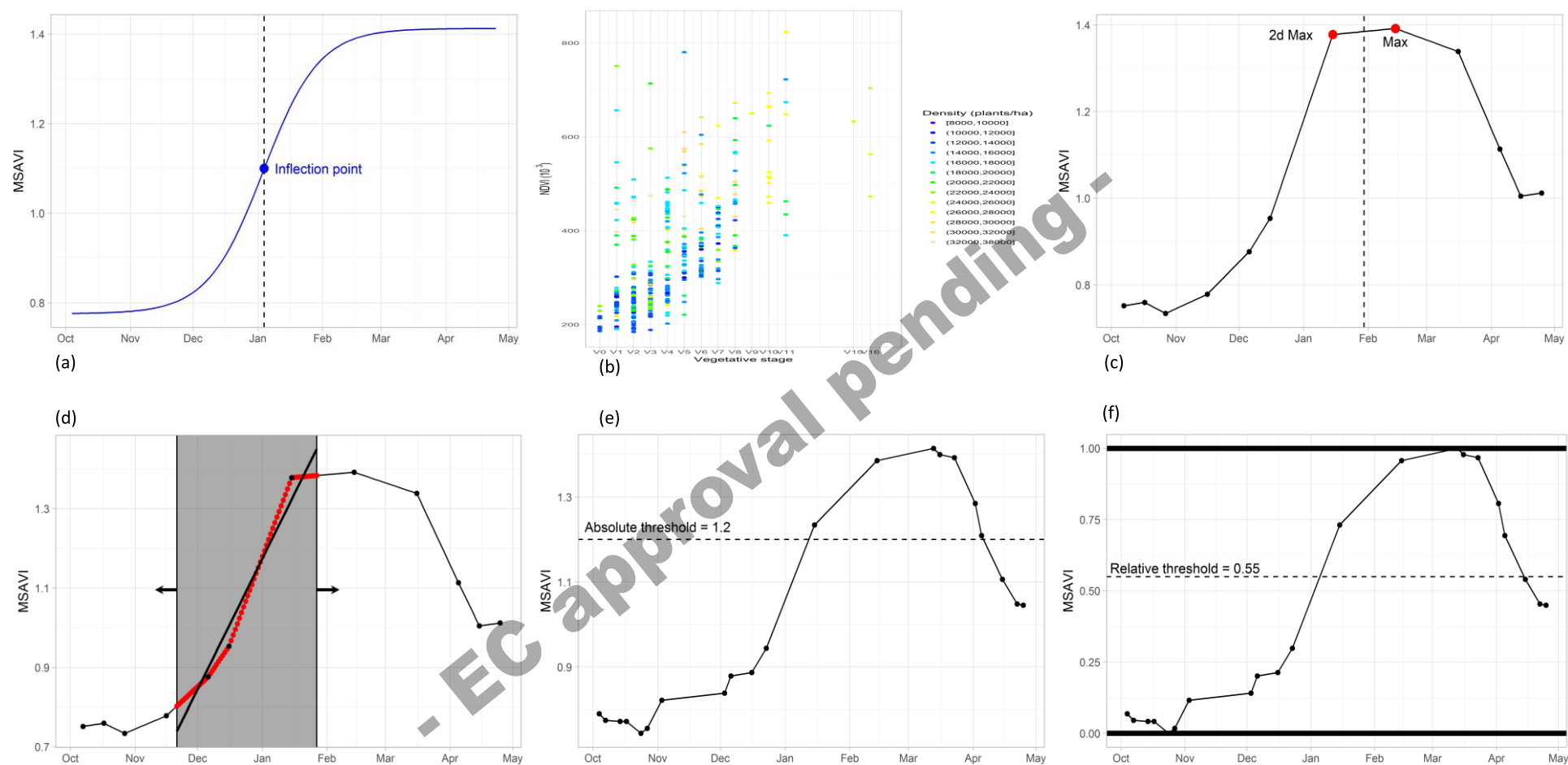


Figure 3-27: Overview of the six emergence estimation methods. Methods without parameterization: (a) inflection point, (b) base logistic (c) maximum value. Methods with parameterization: (d) highest slope, (e) absolute threshold, (f) relative threshold. The threshold methods are both tested on the linear and logistic interpolated observations. Presented vegetation profiles are examples from the maize calibration sample

3.2.4.4 Performance indicators

Furthermore, the comparison of the results of a systematic combination of detection methods and vegetation proxies is necessary for detecting the best combination. The systematic analysis of three approaches (i.e. threshold, derivative method and model fitting) and three vegetation proxies (NDVI, soil-adjusted MSAVI and Hue index (MIR-NIR-Red)) interpolated using two functions (linear, logistic), and the assessment of their relative performances against ground-truth measurements are achieved.

It is important to notice that each tested method aim to detect different time metrics (threshold intersection, highest slope, inflection point, maximum value and base logistic value) to estimate emergence date. Obviously the emergence date is then estimated by using a correction coefficient. The correction coefficient is named the time-lag which corresponds to the interval between the emergence date and the temporal metric specific for each method. Hence, the estimated emergence date from each profile corresponds to the temporal metric date minus the mean time-lag for this method. Subsequently time-lag stability (between the profiles) is the criteria to assess the performance between methods and VIs profiles.

Given that stable time-lag reducing the spread around its value is targeted, two statistical indicators are used to discriminate the different methods and VIs: Standard deviation (SD) and median absolute deviation from the median (MAD). In particular, SD is used for measuring the dispersion of the emergence date estimations around the mean time-lag but it is very sensitive to outliers. Therefore, MAD is incorporated as addressed by Varmuza and Filzmoser (2016).

$$MAD = b \times \text{median} (| (x_i - \text{median} (x_n)) |)$$

with, x_n , the n original observations and b , a multiplicative parameter of 1.4826, assuming the normality of the data and disregarding the abnormality induced by outliers.

Although, utilizing both SD and MAD avoid the outliers affect and treat the data as it has Gaussian distribution, significant difference are shown by the two indicators (Leys et al., 2013).

3.3 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 3.3.1), Forest (section 3.3.2), Grassland (section 3.3.3), Agriculture (section 3.3.4), and new land cover products (3.3.5). For each of the thematic classifications, different inputs, classification methods and parameters are assessed.

3.3.1 Imperviousness

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

3.3.1.1 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 JR, as detailed in WP32.

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. The Sentinel-2 sensor system has an overall number of 12 bands from 10m to 60m spatial resolution;
- 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) such as the Pantex index extensively used for the extraction of the built-up areas (Pesaresi, et al., 2008).
 - **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, , 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 and was testing in phase 1 and implementing in phase 2 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 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 (Florczyk et al. 2019, Pesaresi et al., 2013) and European Settlement Map (ESM Release 2019) (Sabo et al. 2019).

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 selected (cloud-coverage based);
- The multi-temporal pixel-based classification, which is performed on a full stack of this selection of 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 1984), Support Vector Machines (Mountrakis, Im, & Ogola, 2011) and ensembles of classifiers such as Random Forest (Breiman 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 were applied to create the updated built-up mask 2017 in phase 1. 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. During phase 2, Active Learning (AL) method firstly tested over the South-West demo site was successfully implemented to produce the 2018 masks over other demonstration sites.

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 identifies 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 et al. 1996, Pesaresi et al. 2008, Pesaresi et al. 2013, Kemper et al. 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 et al. 2012, Breiman, 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 et al. 2015) or (Lefebvre et al. 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 10 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, Kanevski, & 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 exploiting multi-temporal Sentinel-1 (S1) radar and Sentinel-2 (S2) optical data. The rationale of the adopted methodology is that the temporal dynamics of human settlements in remote sensing imagery are distinct from 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).

Regarding the SAR 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. As is typical for urban applications, the VV polarization contains most of the relevant information regarding urban structures; hence the classifier relies on VV data only. 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. The temporal features are complemented with texture information which are helpful in 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 5x5 spatial neighborhood.

Concerning optical data, only Sentinel-2 scenes with cloud cover lower than 60% are taken into consideration. Data are calibrated and atmospherically corrected using the Sen2Cor software. Next, a series of six spectral indices suitable for an effective delineation of settlements are extracted; these include 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 five 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.

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 most 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, since all three 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 are lower than 5 are excluded.

Regarding SAR 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 result in a high number of candidate points, for AOIs up to 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 is worth noticing 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 are 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 as it is 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 NDVI is extracted for each image. Since the NDVI is 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, 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 module is then employed for properly correlating the resulting training information with the temporal maximum NDVI to finally derive the percent impervious surface (PIS) for the entire area of interest.

3.3.1.2 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 3-28, 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, 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, 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 3-28.

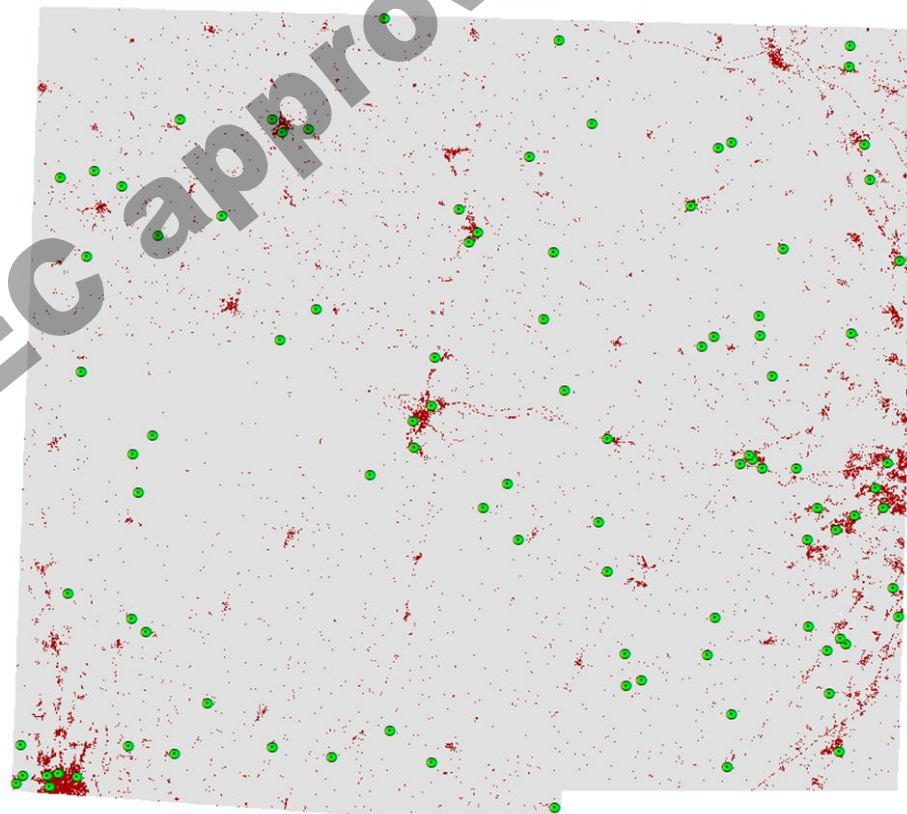


Figure 3-28: Validation samples overlaid on the HRL IMD 2015, reference map.

- (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.

A threshold is applied to the density values for reference and map data to produce binary attributes of built-up for both the reference and map data layers. For IMD, the threshold is set to 30 % with density values lower than 30 % classified as 0 (non-built-up) and density values greater than or equal to 30 % classified as 1 (built-up). The minimum acceptable thematic accuracy of 90 % should be reached for both omission and commission errors for class 1 (built-up).

Regarding the density values, a scatterplot extracted from the sample units for both the reference and prototype is made with a view to assess the correlation between reference and map values and identify any systematic bias (slope and intercept of the regression line significantly different for 1 and 0 respectively). A scatterplot is a way of displaying data against Cartesian coordinates to show and compare values for two variables within a dataset. The data is displayed as a series of points, where the x and y locations relate two variables assigned to a particular recording instance, in this case a PSU. The available measurements for each PSU are the reference data and the mapped value from the product. To quantitatively summarise the results displayed in the scatterplots above a linear regression analysis is performed to estimate the relationships between the reference and mapped product information. The analysis produces a coefficient of determination (R^2) which gives information about the goodness of fit of the estimated regression model. In this case as the reference and map information are meant to represent the same information then it is useful to also consider the slope and intercept of the estimated regression model. The slope should therefore approach 1 and the intercept should be close to 0 for the required relationships. Deviations from the expected values will give an indication of the correspondence of the reference and mapped imperviousness data.

3.3.1.3 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 of the benchmarking has been done in phase 1 on the test site in South-West site of France, over the tiles 30TYP and 31TCJ. Based on the outcome of the phase 1, the section 3.3.1.4 describes the experimental setup of the phase 2 on the 3 test sites South-West, Central and South-East.

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 3-4;

the classification algorithms themselves, as listed in

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

Table 3-4: 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 3-5: 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 3-6. Selection of the best input dataset based on the results given by various classifications.

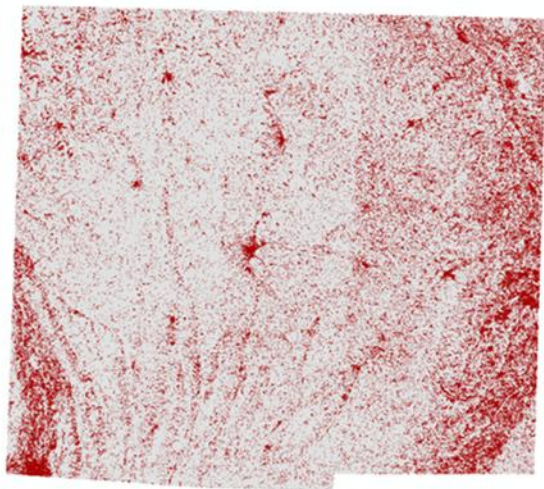
Test	Input Data	Fusion of classifications	Classification algorithm
1	Full dataset – all bands	Dempster-Shafer	Support Vector Machine/Random Forest/ Artificial Neural Network
2	Selection of the 36 images – all bands	Dempster-Shafer	Support Vector Machine/Random Forest/ Artificial Neural Network
3	Selection 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 3-7: Selection of the best sensor dataset based on the results given by SVM.

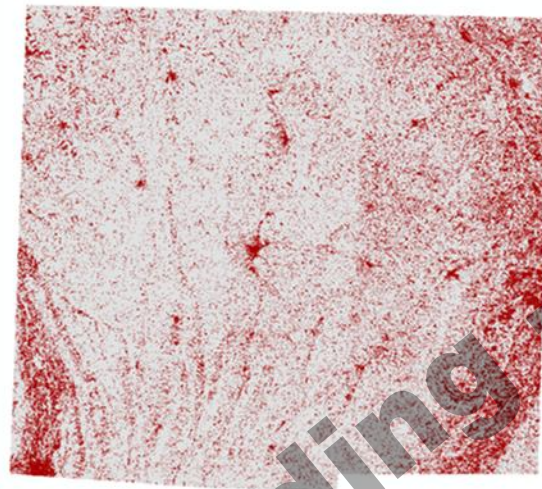
Test	Input Data	Fusion of classifications	Classification algorithm
1	Selection of Sentinel-2 images – all bands	Dempster-Shafer	Support Vector Machine
2	Selection of Sentinel-1 images – all bands	Dempster-Shafer	Support Vector Machine
3	Selection of Sentinel-1 and 2 images – all bands	Dempster-Shafer	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 3-9 and visually assessed in the Table 3-8.

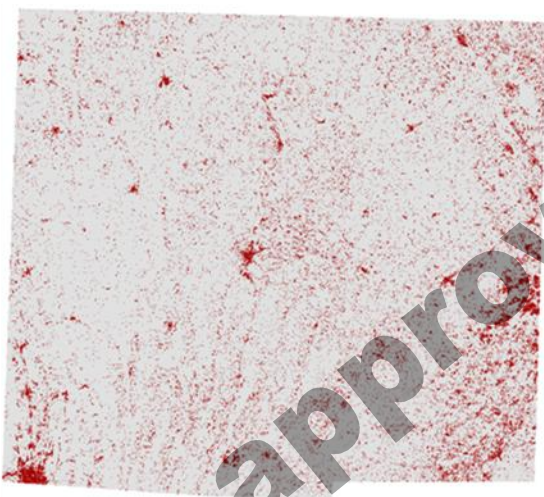
Table 3-8: Visual check for the Demspter-Shafer fusion algorithms based on the precision rate, the recall rate, the overall accuracy and the kappa coefficient – the D-S fused result using the overall accuracy is the closest to the HRL IMD for 2015.



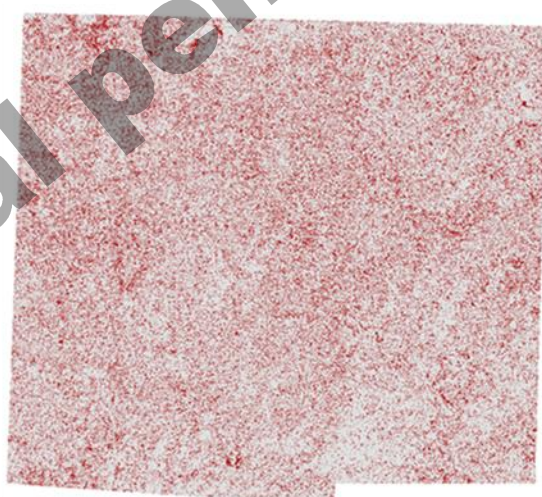
Precision rate



Recall rate



Overall accuracy



Kappa coefficient

Table 3-9: User and producer accuracy for the diverse Dempster-Shafer algorithms.

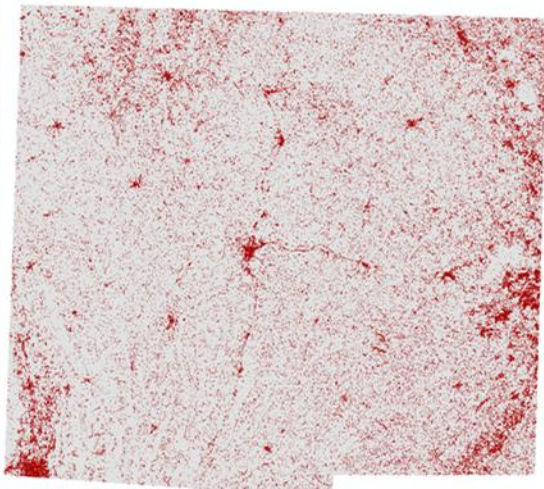
Test	Fusion algorithm	Visual checking	User Accuracy	Producer Accuracy
1	Dempster-Shafer - Overall Accuracy	Yes	74.19%	88.46%
2	Dempster-Shafer – Kappa coefficient	No	42.86%	34.62%
3	Dempster-Shafer – Precision rate	Yes	63.41%	100.00%
4	Dempster-Shafer – Recall rate	Yes	65.79%	96.15%

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.

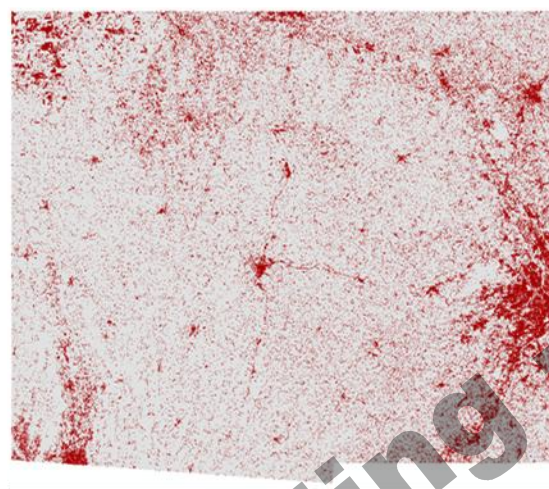
The results of the tests for the determination of the best classification are quantified in the Table 3-11 regarding the use of the full dataset for one year and in the Table 3-13 for a reduced dataset input while being visually assessed in the Table 3-10.

EC approval pending

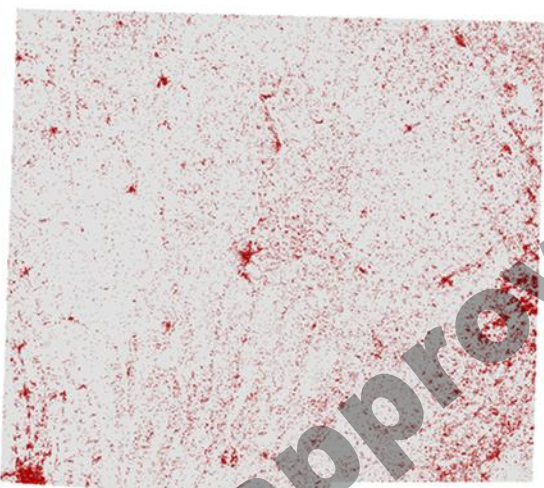
Table 3-10: Visual check for the various classification algorithms and different input datasets – the SVN classifier gives the best result compared to the HRL IMD layer for 2015.



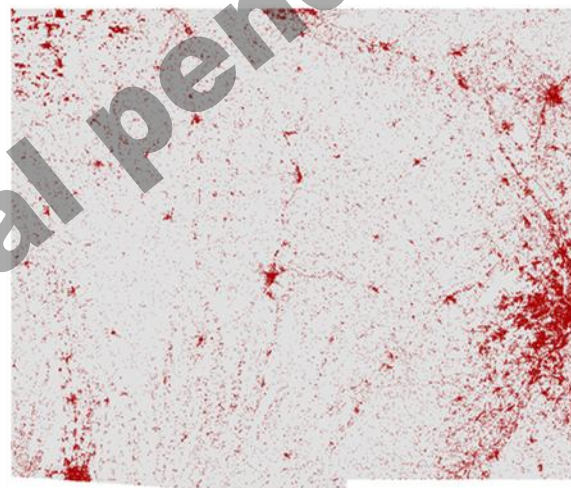
RF applied on full dataset



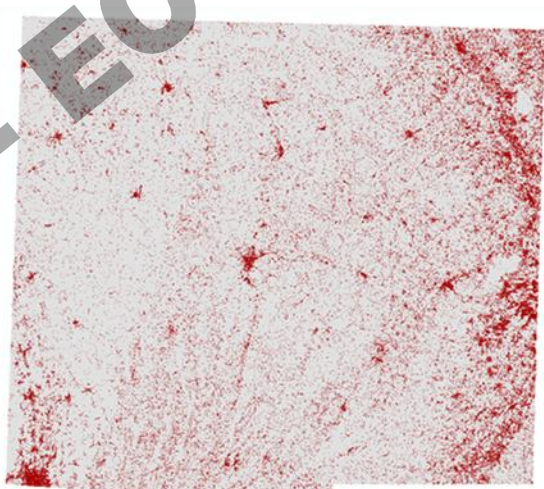
RF applied on subset



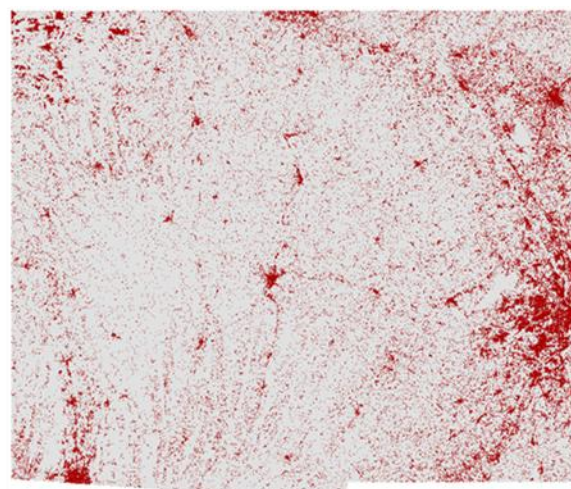
SVM applied on full dataset



SVM applied on subset



ANNs applied on full dataset



ANNs applied on subset

Table 3-11: 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 3-12: DLR Settlement Extent and Growth Classifier

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

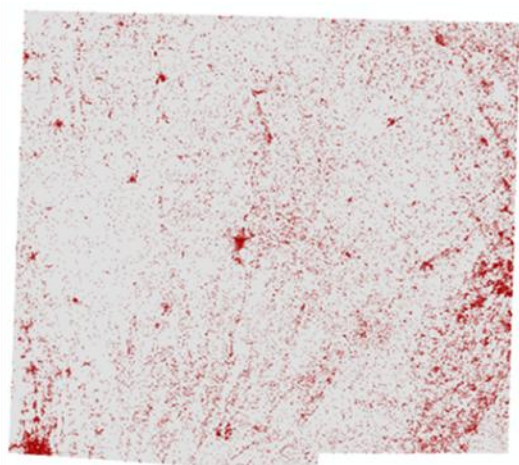
Table 3-13: Subset dataset (36 best images) with all spectral bands results

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%

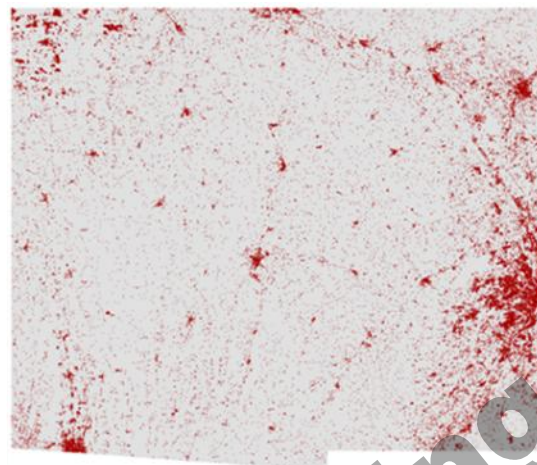
The best classifier appears to be the Active Learning which shows a good balance between the user and the producer accuracies. Then, the Support Vector Machine shows the next best results but with high commission errors. The random forest and neural network classifiers present high producer accuracy 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 3-15 while being visually assessed in the Table 3-14.

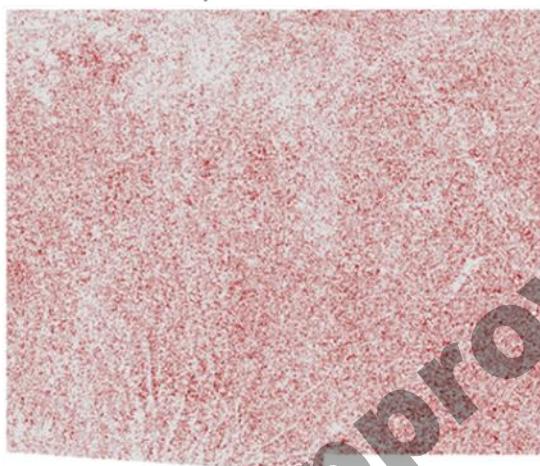
Table 3-14: Visual check for different input datasets – the full dataset input gives the best result compared to the HRL IMD layer for 2015.



Full Sentinel-2 pre-processed dataset with all spectral bands



Selection of the best 36 Sentinel-2 pre-processed images with all spectral bands



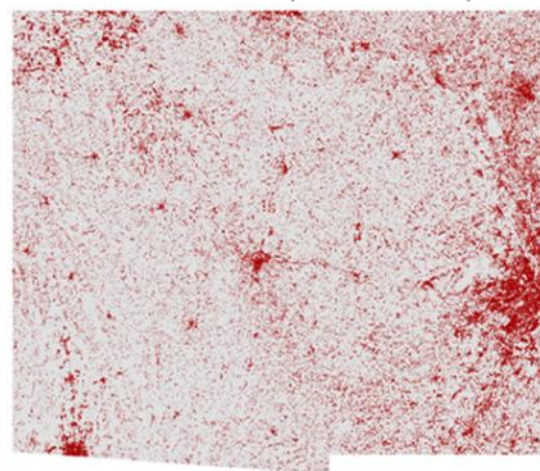
Selection of the best 36 images with a spectral subset (bands 2-3-4-7-9)



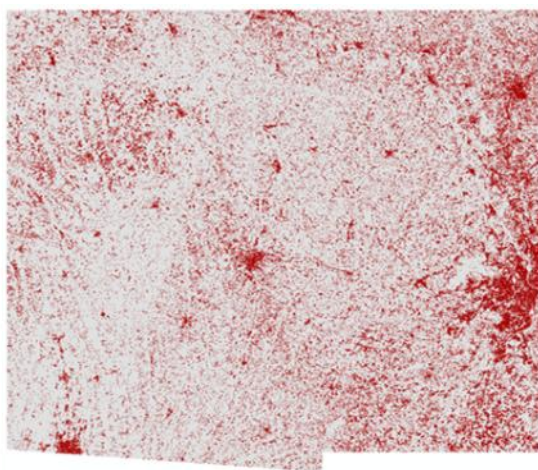
Selection of the best 36 images based on spectral indices combined (NDVI and NDBI)



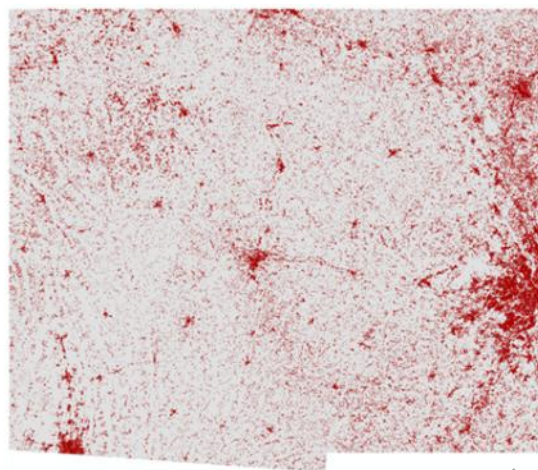
Selection of the best 36 images based on spectral subset (bands 2-3-4-7-9) and indices combined (NDVI and NDBI)



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 3-15: 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 3-29. The layer is a continuous raster with values between 0 and 100 indicating high (red) and low (green) density of impervious surface area.

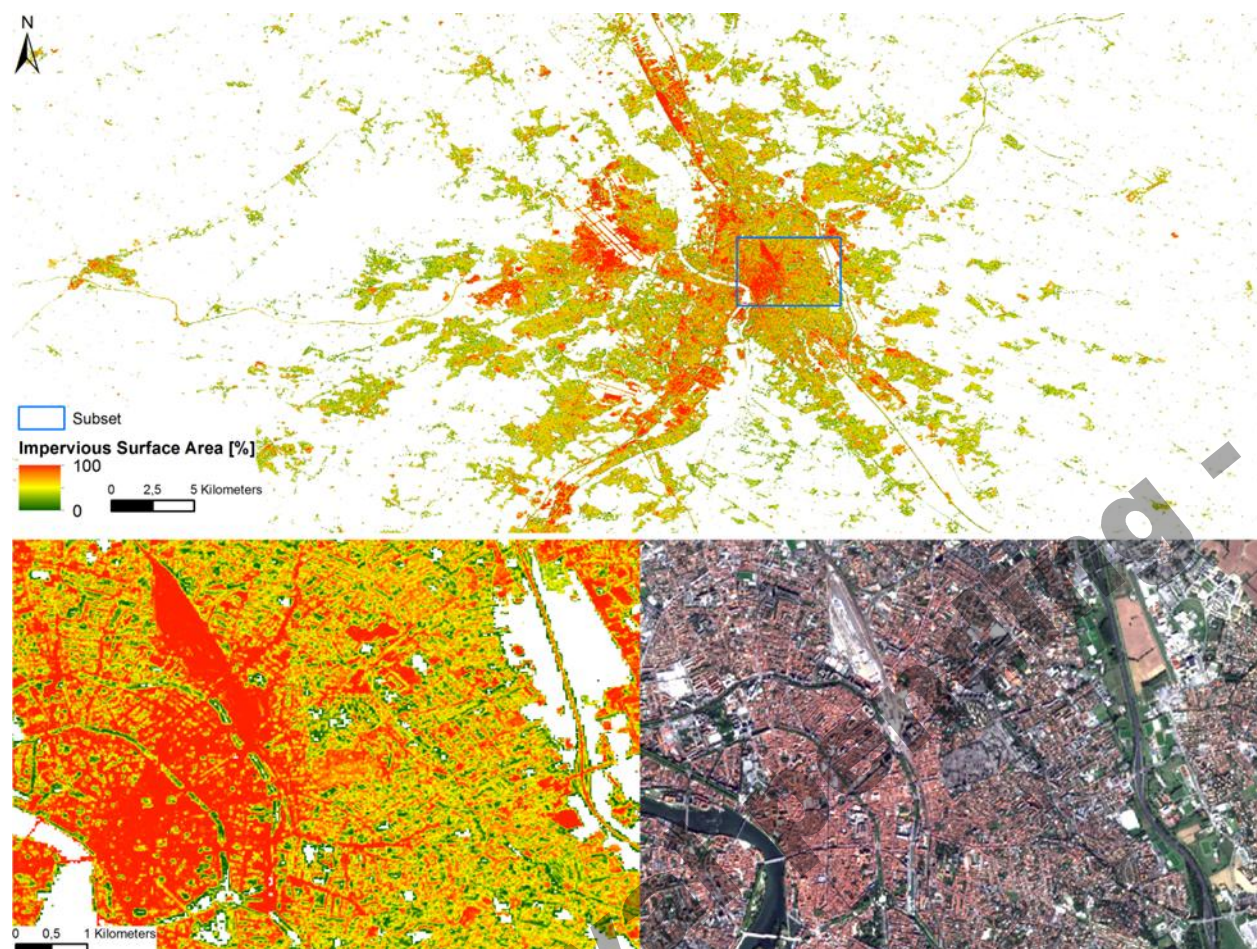


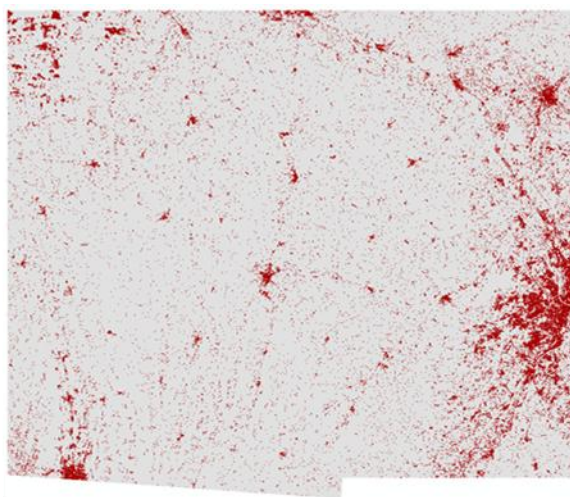
Figure 3-29: Subset of Imperviousness Layer compared with Sentinel-2 imagery.

The results of the tests for the determination of the best sensor are quantified in the Table 3-16 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 3-17.

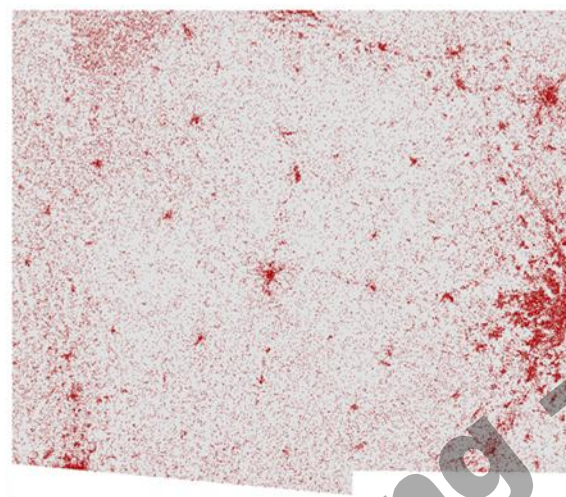
Table 3-16: Impact of the sensor used for the SVM classification

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%

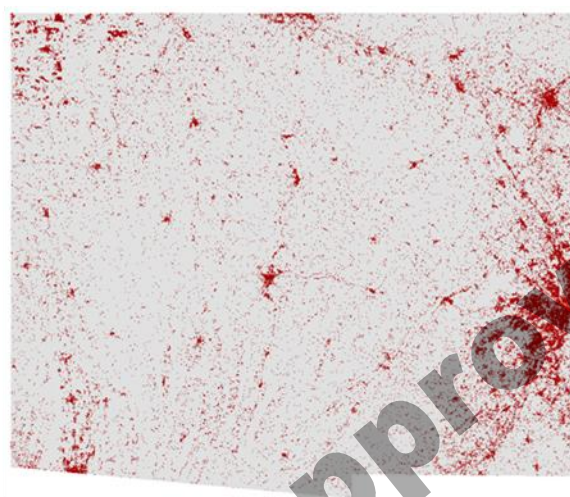
Table 3-17.



Sentinel-2 pre-processed dataset with all spectral bands



Sentinel-1 pre-processed dataset (gamma 0)



Fusion Sentinel-1 and 2 pre-processed dataset

Figure 3-30: 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.

3.3.1.4 Experimental Setup for phase 2

This section shows the phase 2 implementation of the Imperviousness products over the 3 test sites, both the status layer IMD and the built-up layer IBU. Firstly, the integrated EO and ancillary data are described, followed by explaining the pre-processing steps, the demonstration of the classification results of the actual test sites including the accuracy assessment.

3.3.1.4.1 Input Data and Data Integration

Based on the outcomes of the phase 1 (respectively Task 3 and Task 4), a multi-sensor approach combining Sentinel-1 and Sentinel-2 was adopted in 2 to perform the classification that finally leads to the impervious prototypes.

SAR DATA - SENTINEL-1

The Sentinel-1 sensor system has an overall number of 2 bands (both polarisation signals VV and VH) at 10m pixel spacing. Pre-processing has been performed following the processing chain as detailed in WP 32 [AD 07]. Selected scenes cover the time frame from 01-January to 15-November 2018 and represent a total of 1 836 Sentinel-1 images which were used to produce the impervious prototype.

OPTICAL DATA - SENTINEL-2

The South-West test site comprises two adjacent Sentinel-2 tiles (30TYP, 31TCJ).
 The Central test site is composed of two adjacent Sentinel-2 tiles (32UNU, 32TN,).
 The South-East test site is made of two adjacent Sentinel-2 (34TFM, 34TFL).
 All Sentinel-2A+B data in 10m resolution have been pre-processed.

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 bands are used, which are in total 4 bands. The list of the used bands with their central wavelengths and abbreviations is shown in Table 3-18.

Table 3-18: Used Sentinel-2 reflectance bands

Sentinel-2 Bands	Description	Central Wavelength (µm)	Stack number
Band 2	Blue	0.490	1
Band 3	Green	0.560	2
Band 4	Red	0.665	3
Band 8	NIR	0.842	4

Selected scenes cover the time frame from 01-January to 14-November 2018 and represent a total of 775 Sentinel-2A+B images which were used to produce the impervious products for all three test sites.

3.3.1.4.2 Pre-Processing methods for optical time series

As mentioned in the WP 32, the processing methods for optical time images include the generation of spatio-temporally consistent optical images with top of atmosphere reflectance values. Therefore, the following pre-processing steps are applied:

- Atmospheric correction,
- Topographic normalisation,
- Cloud, cloud shadow and snow masking.

ATMOSPHERIC CORRECTION

The Sentinel-2 data produced by CNES' Theia Land Data Centre and available for download are corrected for atmospheric effects, including adjacency effects. These atmospheric corrections include compensating the light absorption by air molecules and the light scattering by molecules and aerosols.

Several models may be used to perform atmospheric corrections. In the case of the MAJA software, the MACCS processor is the model used. It pre-computes "Look-up Tables" using an accurate radiative transfer code (Successive Orders of Scattering), that simulates the light propagation through the atmosphere. The MACCS/MAJA method combines different approaches to obtain robust estimates of aerosol optical thickness.

TOPOGRAPHIC NORMALISATION

A topographic correction is necessary if the test sites are characterized by mountainous terrain as it is the case for the South-West Demonstration site. The topography can significantly influence the radiometric properties of the signal received from the satellite (see Wulder and Franklin, 2012). This effect is caused by the different lighting angles resulting from the topography (cf. Gallaun et al., 2007). The aim of a topographical correction is to compensate for the differences in reflectance intensity between the areas with varying slope, exposure and inclination and to obtain the radiation values that the sensor would measure in the case of a flat surface.

The Sentinel-2 data using the MAJA software and available for download are corrected from the topographic effects.

CLOUD, CLOUD SHADOW AND SNOW MASKING

The MAJA cloud detection method is based on a number of threshold tests using the cirrus band (B10). Additionally, multi-temporal tests are carried out to detect clouds by measuring a steep increase of the blue surface reflectance. Finally, the correlation of the pixel neighbourhood with previous images is calculated to avoid over detections based on the assumption that two different clouds at the same location on successive dates will not have the same shape. If a large correlation is observed, the pixel is excluded from the cloud mask as it is likely to be a bright land surface.

3.3.1.4.3 Pre-Processing methods for SAR time series

Pre-processing has been performed with the Remote Sensing Software Graz (RSG) module “Space Suite”. It comprises the following processing steps:

- Image ingestion: bulk import of original images to RSG *.rsx files, orbit update (precise orbits), automated combination of adjacent scenes
- Image pre-processing: definition of image frame extent (based on selected granules), full image resolution, no speckle filtering, no multitemporal filtering, radiometric terrain correction to gamma naught based on SRTM 4.1 model (Central demonstration site: also tests with sigma naught), combine polarizations in one image stack (band1: VH; band2: VV)
- Orthorectification: based on an interpolated Digital Elevation Model (DEM) (SRTM 4.1), output image resolution is 10m, output image resampling method (nearest neighbour), coordinate system: UTM WGS84
- Calculation of incidence angle map

3.3.1.4.4 Experimental Setup

The developed processing chain is able to process a large amount of input data within a reasonable amount of time to provide the classification results. The achieved level of automation ensures the effective application of the process to map impervious areas of almost the entirety of Europe.

The workflow/methodological steps for the production of the Imperviousness products is listed hereafter:

1. Set-up of reference databases for calibration
2. Production of the Imperviousness 2018 (phase 2)
 - a. Data preparation (Sentinel-1, Sentinel-2)

- b. Biophysical variables and additional image parameters (NDVI, textural metrics for S-2, time features for S-1)
 - c. Derivation of classification training samples from additional reference data (HR layers)
 - d. Production of initial built-up masks for 2018 by automated supervised classification (Active learning)
 - e. Fusion of S-1/S-2 built-up masks
 - f. Absolute calibration of IMD2018
 - g. Post-processing (filtering, contextual analysis based on change probability)
 - h. Validation
- 3. Production of the Built-up 2018
 - a. Biophysical variables and additional image parameters (NDVI, Pantex, textural metrics for S-2)
 - b. Derivation of classification training samples from additional reference data: Open Street Map (OSM) and European Settlement Map (ESM)
 - c. Production of initial built-up masks for 2018 by automated supervised classification (Active learning)
 - d. Post-processing (filtering, contextual analysis)
 - e. Validation

3.3.1.4.5 Set-up of reference databases for calibration

The development of a dataset for calibration of the IMD layers 2018 is needed to provide a reference dataset for the absolute calibration of the HRL2018 10m status layer Imperviousness degree (1-100%).

The reference imperviousness density values are collected for selected sample cells (PSU of 1ha) within the Sentinel-2 tiles. Imperviousness degree levels from 1-100% are obtained for each PSU. The sealing Information, sealed surfaces vs. non sealed surfaces, is collected through Secondary Sampling Units (SSUs – 5x5 grid) within each PSU.

In order to be a representative methodology, the approach chosen combines random and stratified approaches and benefits. The stratification is based on the previous 2015 Imperviousness layer (IMD density value [1-100%]).

3.3.1.4.6 Production of the Imperviousness 2018

DATA PREPARATION (SENTINEL-1, SENTINEL-2)

This step includes all the pre-processing required to prepare the data which can be listed as: downloading, data and metadata extraction, best-scene selection (based on cloud coverage), layer stacking, preprocessing, for S-2 or S-1 – and finally cloud masking for S-2.

For the purpose of the calibration task, the NDVI is derived per single Sentinel-2 image, then mosaicked to a maximum NDVI as shown in Figure 3-31.

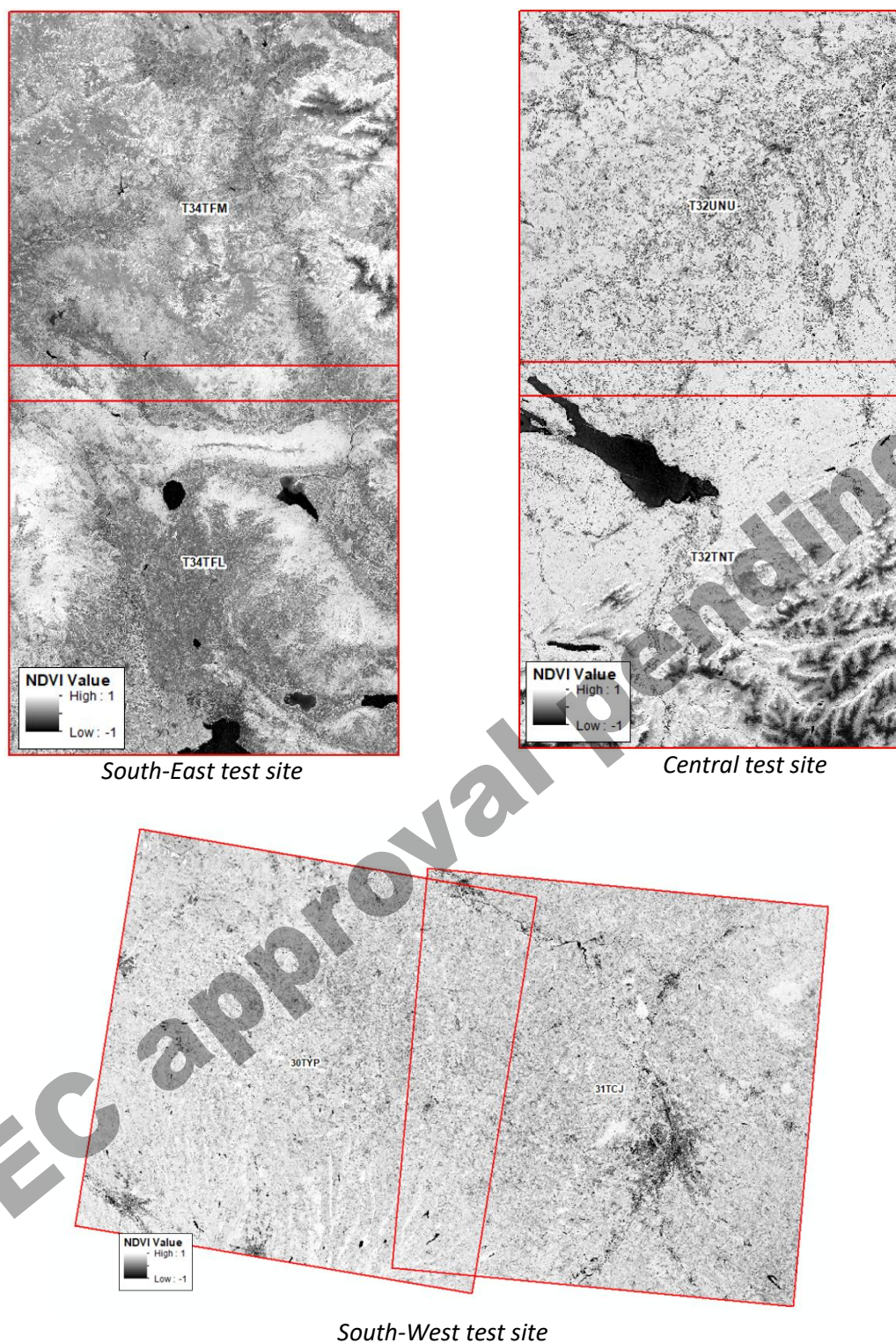
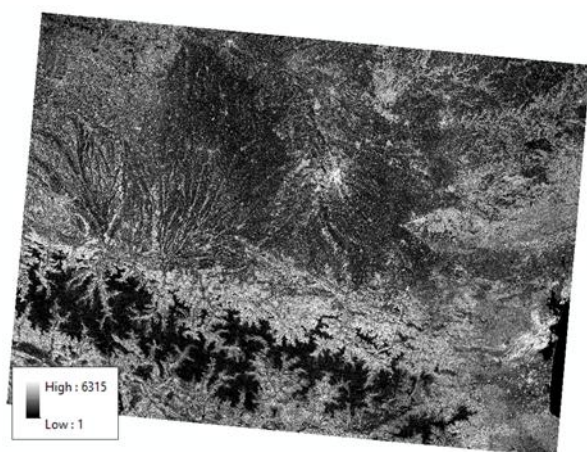


Figure 3-31: 2018 NDVI Sentinel-2 based maximum feature for the year 2018

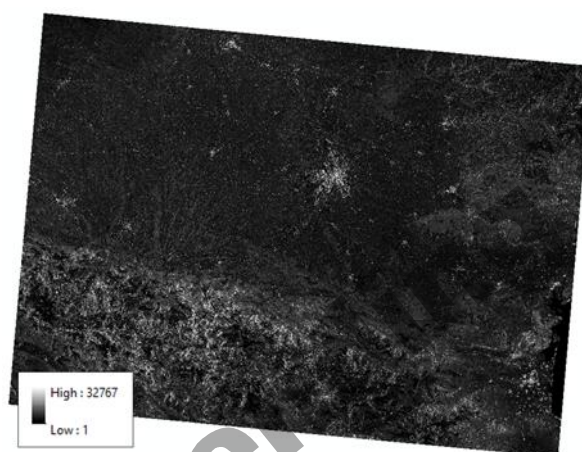
Following annual SAR features are generated using S-1 data and both polarisation signals (VV, VH) including 751 images from 01.01.2018 to 15.11.2018, covering the test sites. Examples for such statistical features are presented in Figure 3-32 Table 3-19 and Figure 3-32.

Table 3-19: SAR annual statistical features.

feature	description
MIN	Minimum
MAX	Maximum
MEAN	Mean
STD	Standard deviation



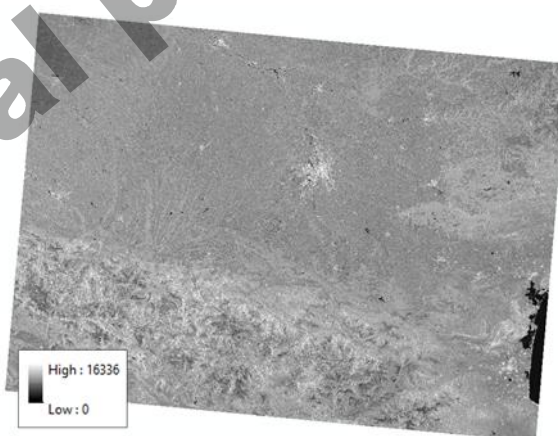
SAR annual minimum feature



SAR annual maximum feature



SAR annual mean feature



SAR annual standard deviation feature

Figure 3-32: SAR statistical features (NB: the 4 features have different value ranges and scaling)

AUTOMATED DERIVATION OF CLASSIFICATION TRAINING SAMPLES

As input for these machine learning algorithms, a set of training data is required. 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 historical High Resolution Layers have been used as training data:

Reliable training samples have been derived from relevant in-situ sources: historical HRL 2015 Imperviousness, Forest, Grassland, Water and Small Woody Features. In order to best reflect the different imperviousness classes, an automated random point sampling within buffered IMD 2015

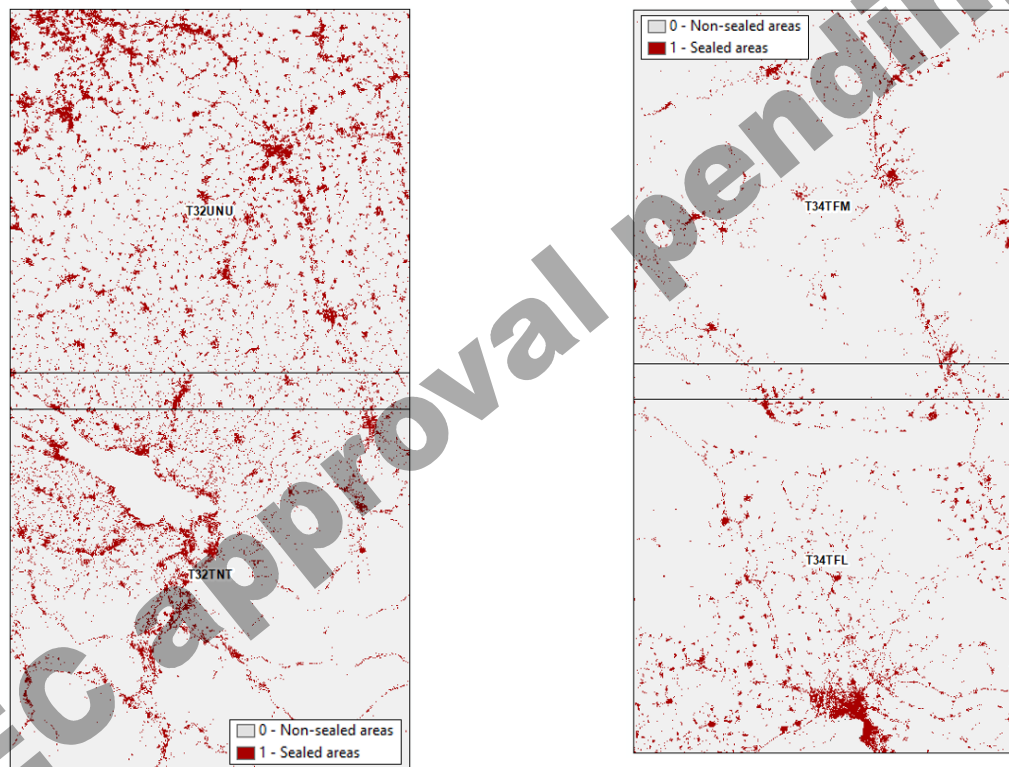
has been applied. Samples in non-built-up areas have been selected in different land cover classes such as grassland, bare soil, vegetation and water in order to obtain a representative distribution of non-imperiousness samples.

Based on the spectral information, biophysical indicators and texture parameters at the training sample points, the algorithm ‘learns’ how to classify the features (Tan et al. 2006, Camp-Valls, 2009) and identifies the most significant combinations of input parameters to differentiate built-up areas from other land cover.

For the purpose of the automated derivation of the training sample, a stratified random approach, based on the HRLs, has been preferred.

PRODUCTION OF INITIAL SEALED AND NON-SEALED MASKS FOR 2018

The results of the initial Sentinel-2 based sealed and non-sealed mask are shown in Figure 3-33.



Central test site

South-East test site

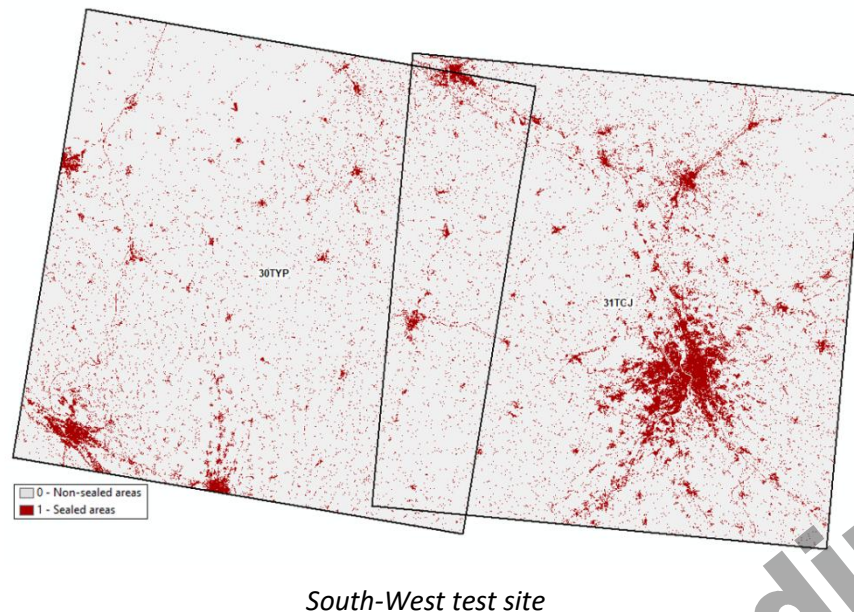


Figure 3-33: Sentinel-2 based initial sealed and non-sealed mask for 2018 for the test sites

ABSOLUTE AND RELATIVE CALIBRATION OF IMD2018

Besides the production of initial sealed and non-sealed masks for 2018 (previously described), one key step in the HRL Imperviousness production is the estimation of the degree of imperviousness and linking these IMD measurements over time. Each single pixel in the built-up mask will be assigned an imperviousness density value of 1 to 100%. The linkage between the biophysical variables and the IMD measurements will be done through an absolute (linking the biophysical variables to IMD) and relative calibration procedure. This combination improves the accuracy of imperviousness density estimates, correct any over-/underestimation of values and assure comparability and consistency over time.

The reference calibration database serves as calibration input for an absolute calibration of the 2018 IMD measurements. For the prediction of the imperviousness degree, a linear regression method is used to model the relationship between the collected reference samples and meaningful metrics from the biophysical variables (e.g. NDVImax) derived from the seasonal image composites. The established linear equation is applied to transform the input data into imperviousness degree values between 1 – 100%. This results in absolutely calibrated IMD measurements derived from the 2017 and 2018 imagery.

Then, the calibrated 20m IMD 2015 status layer will be used as input to adjust the imperviousness density values of 2018 by relative calibration. Indeed, despite the absolute calibration based on a well-established procedure (with the use of a reference calibration dataset), there will always remain some obvious and local issues in the imperviousness density derivation which will lead to wrongly detection changes in the change Layers. The relative approach is so needed to correct these local artefacts.

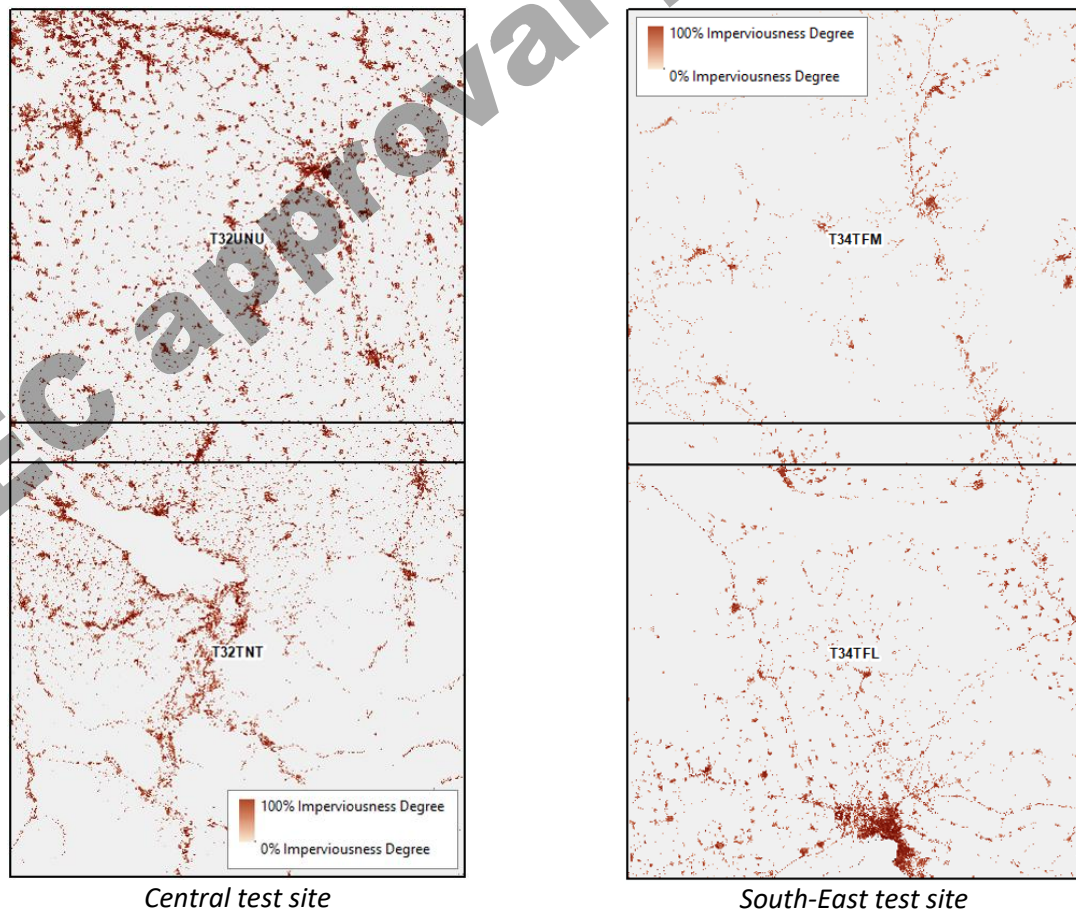
The IMD 2018 values, limited to the newly created 2018 sealed mask, are re-analysed by an automatic cross-calibration approach: the IMD 2018 values are compared to the IMD 2015 values resampled to 10 meters spatial resolution and further corrected using a rule-based approach. Indeed, a filtering approach is needed to adequately map sealing changes. In 20m spatial resolution, changes of imperviousness density within sealed areas are not that frequent compared to changes

of the sealed area. Despite all the calibration efforts in image pre-processing and subsequent adaptation procedures, there will always remain a certain error budget for sealing change detection mainly caused by:

- Persisting spectral differences due to even subtle deviations in illumination, shadow effects, atmospheric conditions and vegetation status, often in conjunction with:
- Geometric misalignments of the IMAGE databases. This occurs quite frequently and often exceeds a range of 1 pixel (>20m).

Hence, in order to derive a reliable and realistic picture of sealing changes (within existing built-up areas), thresholds are applied. Differences of >20% of sealing increase will be considered acceptable if a contiguous area of at least 16 (10m x 10m) pixels is concerned. The threshold of 16 contiguous pixels permits to overcome the scaling issue (10 vs 20m spatial resolution). Differences $\leq 20\%$ sealing increase will be considered as stable. The special case of imperviousness decrease is rare and, if occurring, it will rather be due to a re-greening (full de-sealing) of an impervious surface than an actual decrease. With regard to this assumption sealing decrease within built-up areas will only be accepted as valid if a remarkable change of 80% decrease takes place. Differences $\leq 80\%$ sealing decrease will be considered as stable. In phase 2, for the South-West prototype, the same rule-based approach was applied to correct the 2018 values with comparison to the 2017 IMD values obtained in phase 1.

The final results of the implemented Imperviousness layers 2018 for the 3 test sites is shown in Figure 3-34.



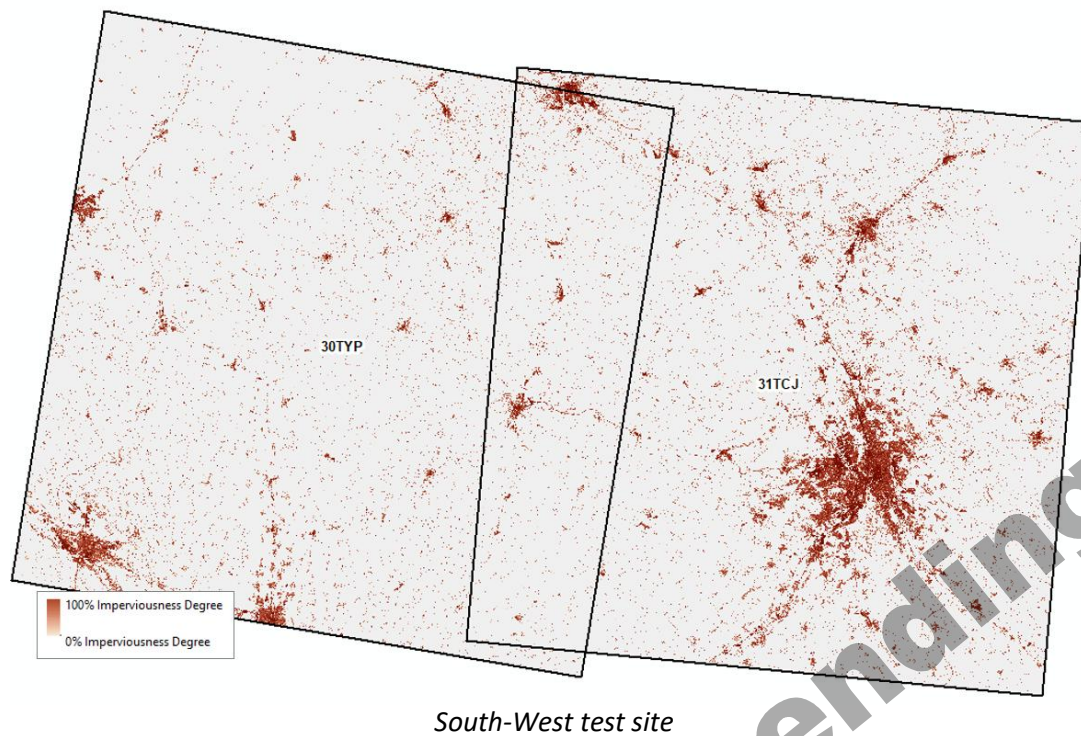


Figure 3-34: Final HRL Imperviousness 2018 layers for the test sites

3.3.1.4.7 Production of the Built-up Layer 2018

DATA PREPARATION (SENTINEL-1, SENTINEL-2)

This step includes all the pre-processing mentioned in section 3.3.1.4.2 and 3.3.1.4.3 for Sentinel-2.

For the purpose of the BU layer classification, the PanTex is derived per single S-2 image, as shown in Figure 3-35. The PanTex is used as input data along with the 10 m spectral bands for the classification.

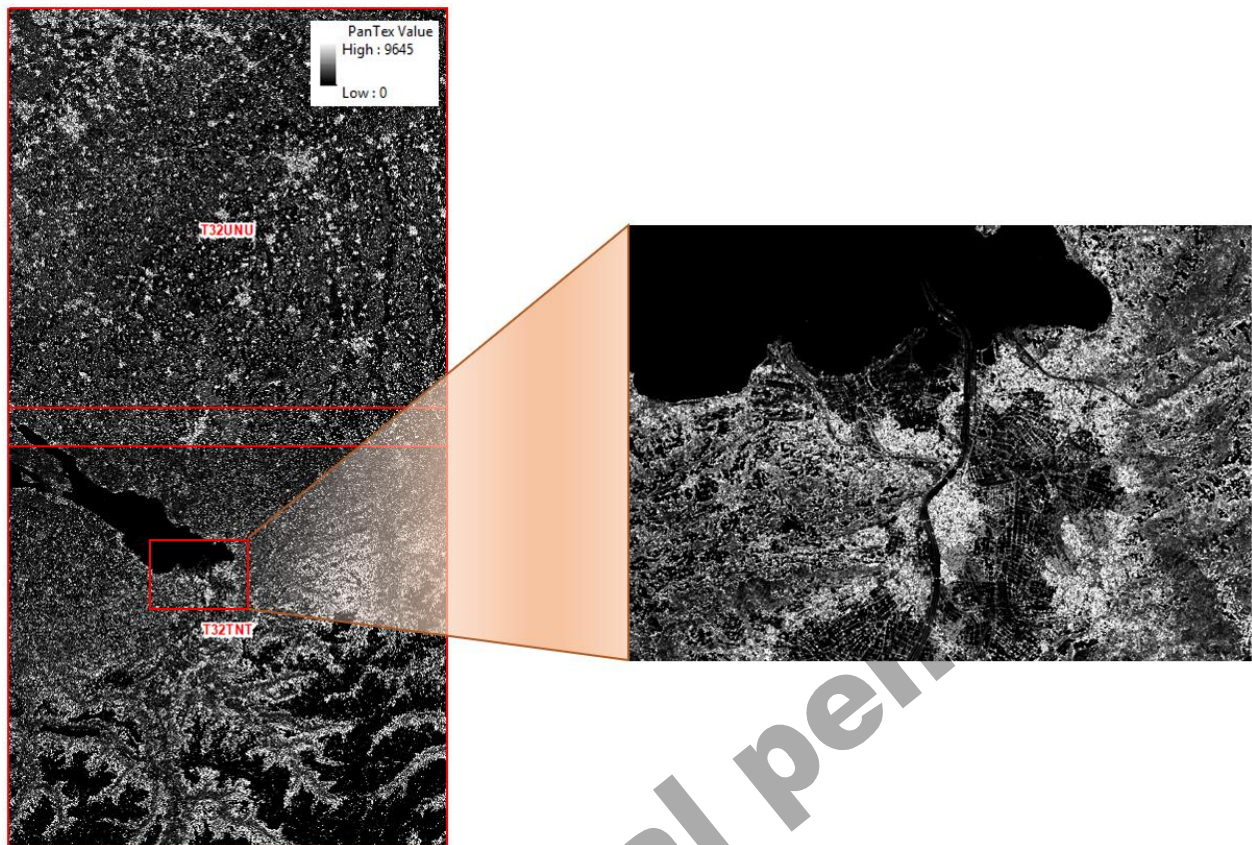


Figure 3-35: 2018 PanTex Sentinel-2 based feature for the Central test site

AUTOMATED DERIVATION OF CLASSIFICATION TRAINING SAMPLES

As input for the machine learning algorithms, a specific set of training data, different from the one implemented for the sealed surface classification, is also required for the Built-up layer. The Open Street Map (OSM) and European Settlement Map (ESM) have been used as training data.

In order to best reflect the different built-up features, an automated random point sampling has been applied. Samples in non-built-up areas have been selected in different features/classes such as roads, railways or parking lots to obtain a representative distribution of non-built-up samples.

Based on the spectral information, biophysical indicators and texture parameters at the training sample points, the algorithm 'learns' how to classify the features (Tan et al. 2006, Camp-Valls, 2009) and identifies the most significant combinations of input parameters to differentiate sealed areas from other land cover.

For the purpose of the automated derivation of the training sample, a stratified random approach, based on the HRLs, has been preferred.

PRODUCTION OF INITIAL BUILT-UP AND NON-BUILT-UP MASK FOR 2018

The results of the initial Sentinel-2 based built-up and non-built-up mask are shown in Figure 3-36.

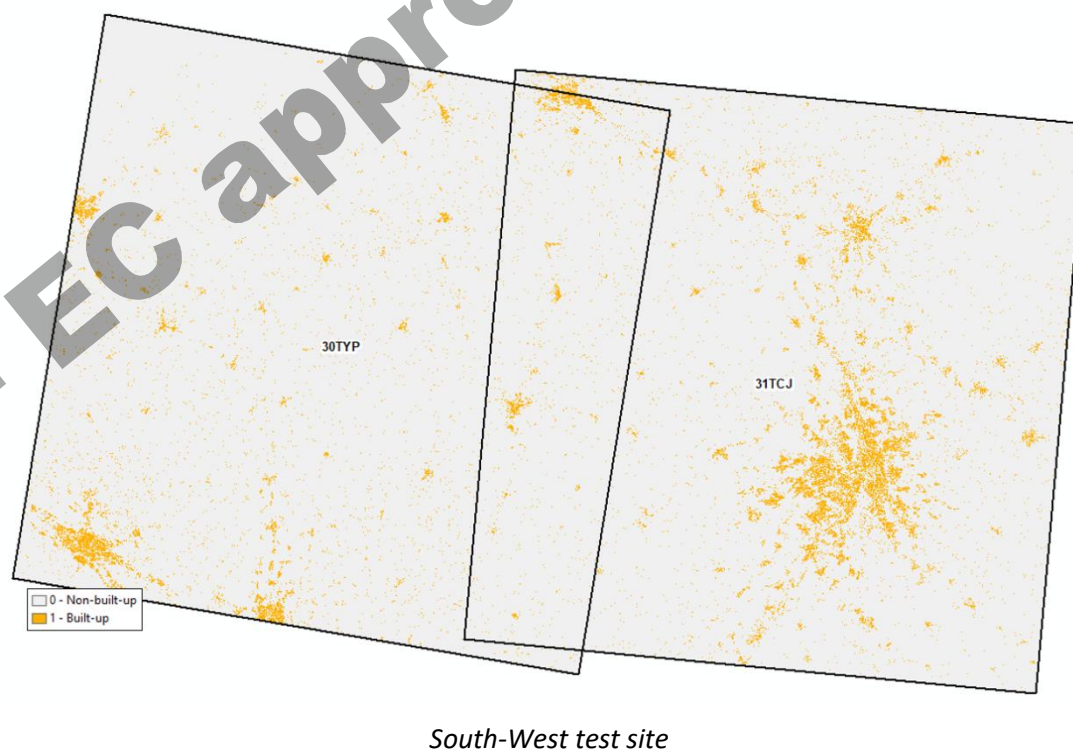
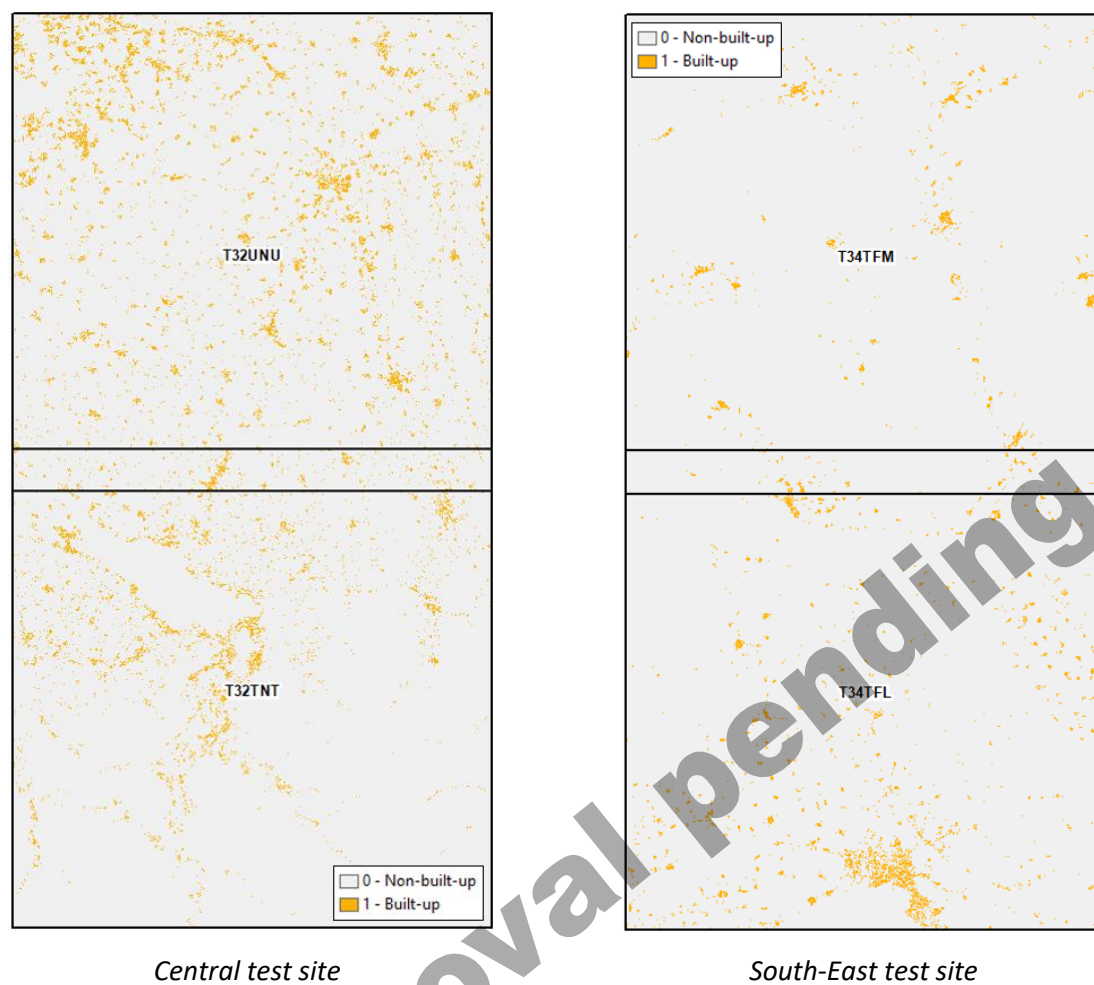


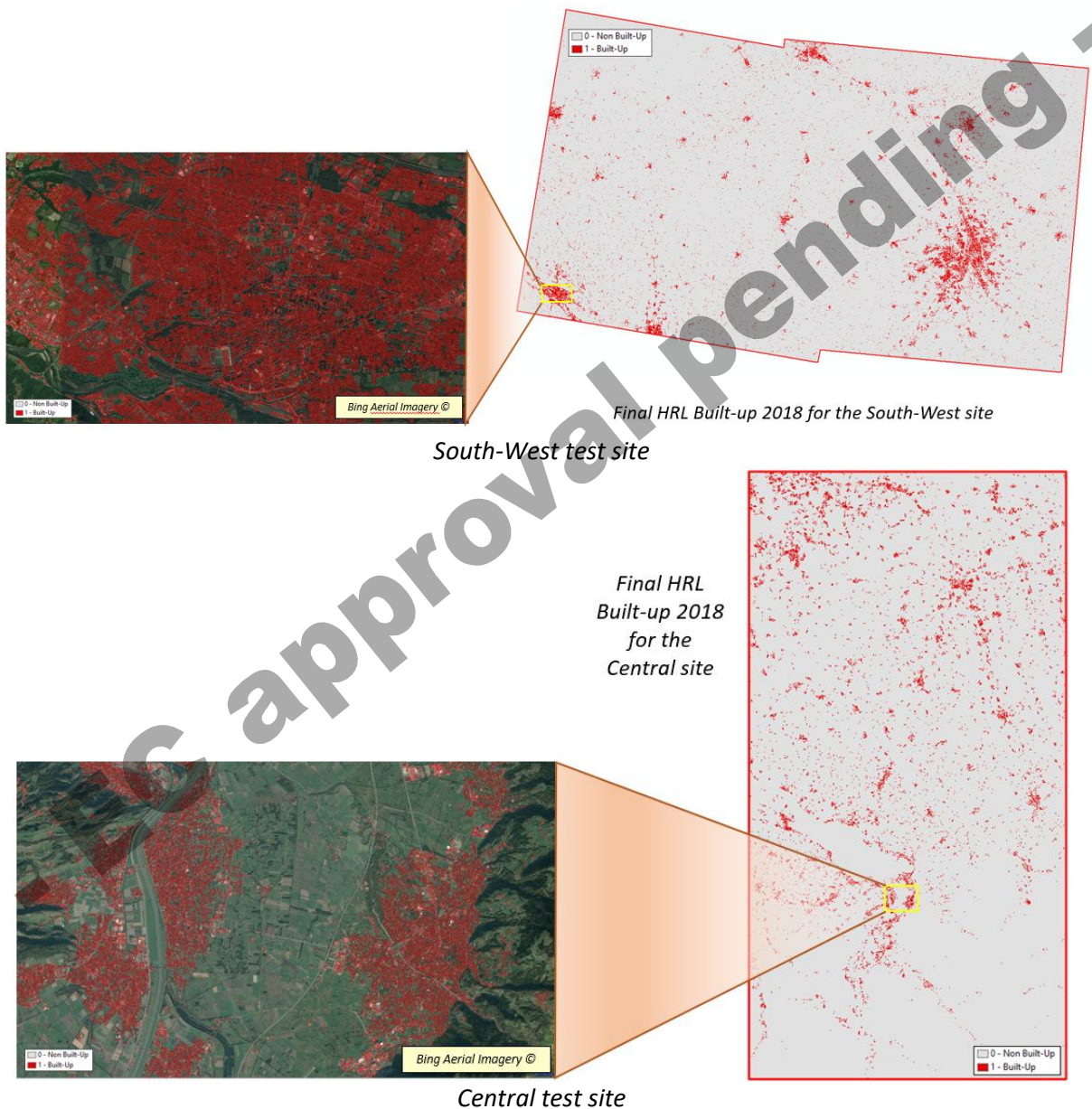
Figure 3-36: Sentinel-2 based initial built-up and non-built-up masks for 2018 for the test sites

POST-PROCESSING

The post-classification implies post-processing of the layers in order to be spatially consistent with the HRL Imperviousness (IMD) layers including:

- Post-processing filtering using the sealed mask. Indeed, built-up pixels which are not sealed in the IMD classification should be removed in a post-classification step to ensure the spatial consistency between products.

The final results of the implemented Built-up layers 2018 is shown in Figure 3-37.



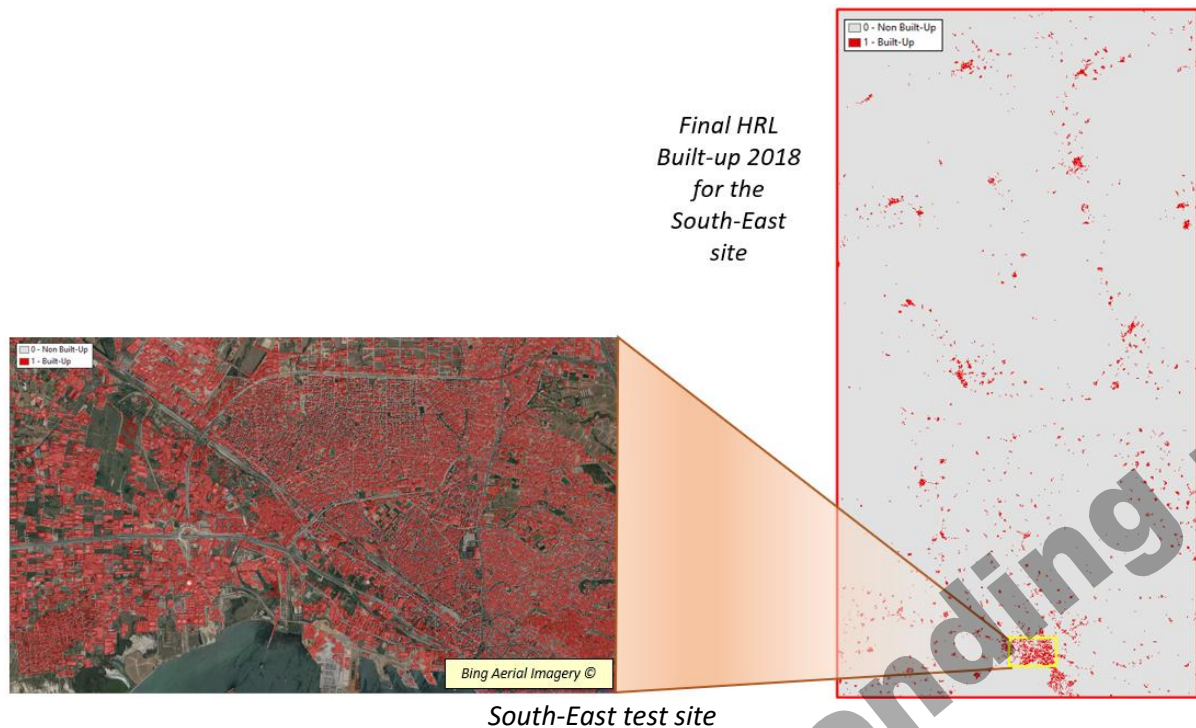


Figure 3-37: Final HRL Built-up 2018 layers for the test sites

3.3.1.5 Classification Results and Validation

This chapter depicts the results of the classification as well as their validation. Firstly, the regression analysis is performed (see section 3.3.1.5.1). Then, the thematic accuracies are summarized (see section 5.1.2.2). The thematic accuracies are followed by a discussion of the validation results (see section 5.1.2.3).

3.3.1.5.1 IMD 2018 scatterplots & regression analysis

A scatterplot is a way of displaying data against Cartesian coordinates to show and compare values for two variables within a dataset. The data is displayed as a series of points, where the x and y locations relate two variables assigned to a particular recording instance, in this case a PSU. The available measurements for each PSU are the original reference data (called REFERENCE in each figure) and the mapped value from the product (called MAP on the figures). For this validation exercise the position / value on the horizontal axis represented the reference information and the position / value on the vertical axis represents product (MAP) information. In this way the relation of the reference and product information for a point can be compared to a 1:1 line which runs diagonally across the scatter plot. The closeness of a point to the point to the 1:1 line is an indication of the similarity between the reference and mapped results. The points that lie exactly on the x and y axes are related to omission and commission rather than the calibration of the IMD values themselves.

The scatterplots presented in Figure 3-38 show very limited scatter of up to 20 % each side of the best fit line. There are a few numbers of point on the x and y axes showing commission, where sealing is mapped that is not present in reality, and omission, where sealed areas are missed which seems to indicate that the commission and omission errors are also limited. It can also be seen that the distribution is almost centred on the 1:1 line.

The scatterplots and the resulting best fit lines are controlled by the actual geography of the test sites. The results for the South-West and South-East show that there are limited number of sealed areas, the areas present tend to have low imperviousness values. There are also a few omissions and commissions which is highly likely for Mediterranean regions where vegetation is limited and there may be extensive areas of bare soils. The Central test site is more representative of Europe as a whole and contains significant sealed areas of varying imperviousness.

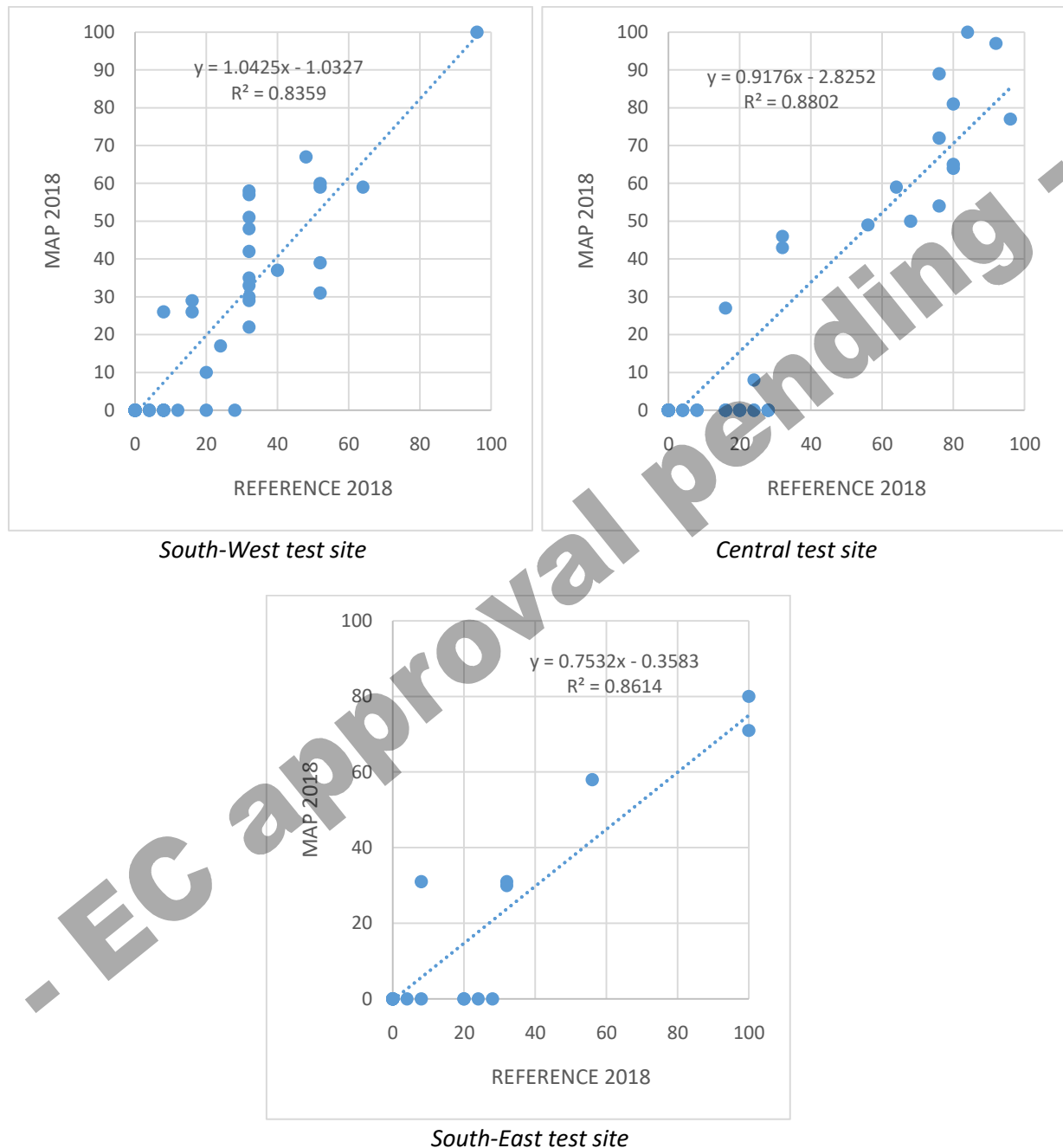


Figure 3-38: Scatterplots for the continuous IMD layers 2018 validation

To quantitatively summarise the results displayed in the scatterplots above a linear regression analysis is performed to estimate the relationships between the reference and mapped product information. The analysis produces a coefficient of determination (R^2) which gives information about the goodness of fit of the estimated regression model. Coefficients of determination closer to 1 represent a better fit. In this case as the reference and map information are meant to represent the same information then it is useful to also consider the slope and intercept of the estimated

regression model. The slope should therefore approach 1 and the intercept should be close to 0 for the required relationships. Deviations from the expected values give an indication of the correspondence of the reference and mapped imperviousness data.

All regression coefficients are greater than 0.80 which indicate a very good relationship between the reference and map density values. The lowest value is related to the South-West test site with complex landscapes with a R^2 close to 0.8. Nevertheless, all regression coefficients are greater than 0.80 which prove the good relationship.

Regression slopes are consistently close to 1 with very few variabilities except for the South-East demonstration site (close to 0.75). All slope values are mostly smaller than 1,0 which is mainly a result of commission errors (IMD detected in non-sealed areas). The intercept parameter shows good values close to 0 but slightly under 0 which is a result of the omission errors (negative intercept).

3.3.1.5.2 Thematic accuracy

The below confusion matrices give a summary of the internal accuracy assessment of the HRL Imperviousness 2018 for the demonstration sites, see Table 3-20, Table 3-21 and Table 3-22, respectively for the South-West, the Central and then the South-East test sites.

Table 3-20: Confusion matrix of the internal validation of the IMD 2018 in test site South-West (area-weighted)

IMD_2018_Testsite_SouthWest_0303_5_10m		REFERENCE			User Accuracy	Confidence Interval
PRODUCT		Non-Sealed	Sealed	Total		
	Non-Sealed	53.64	0.58	54.22	98.92 %	0.37 %
	Sealed	0.88	6.63	7.51	88.32 %	1.84 %
	Total	54.52	7.22	61.73		
	Producer Accuracy	98.39 %	91.90 %		97.63 %	Overall Accuracy
	Confidence Interval	0.45 %	1.52 %		1.01 %	Confidence Interval
					0.99	F-Score Non IMD
					0.90	F-Score IMD
					0.89	Kappa

Table 3-21: Confusion matrix of the internal validation of the IMD 2018 in test site Central (area-weighted)

IMD_2018_Testsite_Central_03035_1_0m		REFERENCE			User Accuracy	Confidence Interval
PRODUCT		Non-Sealed	Sealed	Total		
	Non-Sealed	60.87	0.58	61.45	99.05 %	0.10 %
	Sealed	0.29	4.68	4.97	94.12 %	1.11 %
	Total	61.16	5.26	66.42		
	Producer Accuracy	99,52 %	88,89 %		98,68 %	Overall Accuracy
	Confidence Interval	0.09 %	1.77 %		0.18%	Confidence Interval
					0.99	F-Score Non

	IMD
0.91	F-Score IMD
0.91	Kappa

Table 3-22: Confusion matrix of the internal validation of the IMD 2018 in test site South-East (area-weighted)

IMD_2018_Testsite_SouthEast_03035_10m		REFERENCE			User Accuracy	Confidence Interval
PRODUCT		Non-Sealed	Sealed	Total		
	Non-Sealed	71,92	0,29	72,21	99,60 %	0.04 %
	Sealed	0,29	1,46	1,75	83,33 %	2.28 %
	Total	72,21	1,75	73,97		
	Producer Accuracy	99,60 %	83,33 %		99,21 %	Overall Accuracy
	Confidence Interval	0.05 %	2.39 %		0.11 %	Confidence Interval
					1.00	F-Score Non IMD
					0.83	F-Score IMD
					0.83	Kappa

The below confusion matrices give a summary of the internal accuracy assessment of the HRL built-up 2018 for the test sites, see Table 3-23, Table 3-24 and Table 3-25 respectively for the South-West, the Central and then the South-East demonstration sites.

Table 3-23: Confusion matrix of the internal validation of the BU 2018 in test site South-West (area-weighted)

BU_2018_Testsite_SouthWest_03035_10m		REFERENCE			User Accuracy	Confidence Interval
PRODUCT		Non-built-up	Built-up	Total		
	Non-built-up	56,76	0,29	57,06	99,49 %	0.13 %
	Built-up	0,88	3,80	4,68	81,25 %	2.44 %
	Total	57,64	4,09	61,73		
	Producer Accuracy	98,48 %	92,86 %		98,11 %	Overall Accuracy
		0.48 %	0.22 %		0.51 %	Confidence Interval
					0.99	F-Score Non-built-up
					0.87	F-Score Built-up
					0.86	Kappa

Table 3-24: Confusion matrix of the internal validation of the BU 2018 in test site Central (area-weighted)

BU_2018_Testsite_Central_03035_10m		REFERENCE			User Accuracy	Confidence Interval
		Non-built-up	Built-up	Total		
PRODUCT	Non-built-up	62,35	0,30	62,65	99,53 %	0.09 %
	Built-up	0,59	5,19	5,78	89,78 %	1.66 %
	Total	62,94	5,49	68,43		
	Producer Accuracy	99,06 %	94,61 %		98.70 %	Overall Accuracy
	Confidence Interval	0.15 %	1.11 %		0.23 %	Confidence Interval
					0.99	F-Score Non-built-up
					0.92	F-Score Built-up
					0.91	Kappa

Table 3-25: Confusion matrix of the internal validation of the BU 2018 in test site South-East (area-weighted)

BU_2018_Testsite_South-East_03035_10m		REFERENCE			User Accuracy	Confidence Interval
		Non-built-up	Built-up	Total		
PRODUCT	Non-built-up	72,21	0,29	72,50	99,60 %	0.07 %
	Built-up	0,29	1,17	1,46	80,00 %	2.49 %
	Total	72,50	1,46	73,97		
	Producer Accuracy	99,60 %	80,00 %		99,21 %	Overall Accuracy
	Confidence Interval	0.08 %	2.36 %		0.10 %	Confidence Interval
					1.00	F-Score Non-built-up
					0.80	F-Score Built-up
					0.80	Kappa

3.3.1.6 Summary and conclusions

The analysis performed as part of the phase 2 shows better results for the following set of parameters:

- a mono-temporal approach, image-by-image;
- the use of an active learning;
- the input being a 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 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. In phase 1, 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, Sentinel-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.

Firstly envisaged, the multi-sourcing approach, with not only one sensor, Sentinel-2, but also other sensors including Sentinel-3 or a substitute such as PROBA-V, was not explored due the low spatial resolution of these kinds of image. Different studies (Pesaresi, et al., 2013), (Hansen, et al., 2013) exploit this multi-source approach to create global built-up maps with remarkable success. The realization of those prototypes results has a major impact on the following activities (WP 34 and 35).

Regarding the Built-up product, it should be noticing the high adding value of the Pantex Index for this layer. Actually, test was undertaken based on HR data (Sentinel) but the use of VHR data at larger scale for the BU detection should be envisaged since better results are expected as shown in Table 3-26 and Figure 3-39 based on a benchmarking performed on a spatial subset in the South-West test site.

Table 3-26: Comparison of Built-up Layers based on VHR vs HR data compared

Test Site	User Accuracy	Producer Accuracy
VHR Based	88.89%	99.00%
HR Based	68.86%	93.59%

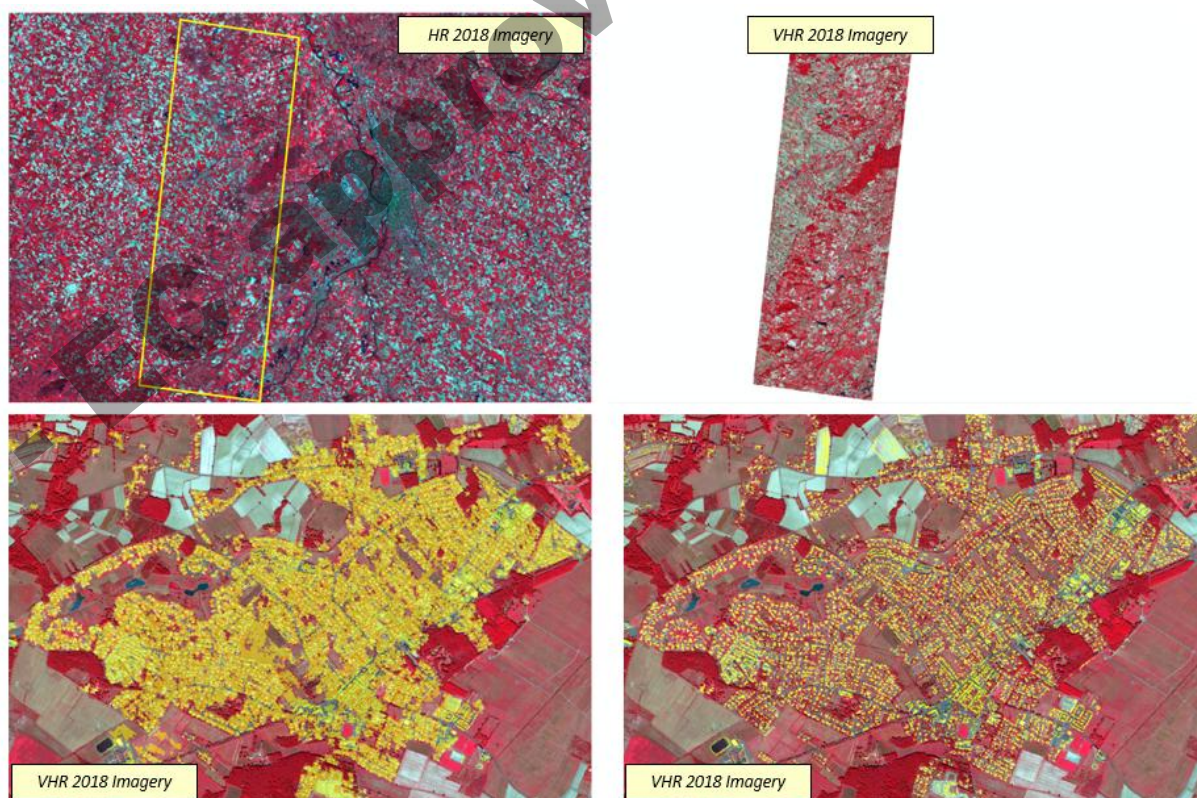


Figure 3-39: Comparison of Built-up Layers based on VHR vs HR data compared with VHR imagery (Toulouse, France)

3.3.2 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 the reference year 2018 is currently under production. 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 an improved Tree Cover Mask (TCM) and the status layer Dominant Leaf Type (DLT), whereas the first one is used to derive the incremental update Tree Cover Change (TCC) for 2017/2018. In that regard, the use of temporal-spectral metrics as classification input features is assessed and compared between different sensor data scenarios.

Furthermore, the possibilities and limitations to generate the continuous Tree Cover Density (TCD) layer using optical time series data are thoroughly assessed. Tests have been performed in the ECoLaSS North test site in Sweden, Central region in Austria/Germany and South-East test site in Greece/Bulgaria (see test sites distribution in Figure 1-1).

3.3.2.1 Description of candidate methods

According to previous tests and the first prototypic implementation in phase 1, the Random Forest classifier was chosen as the best-rated classification algorithm for testing and implementation in phase 2. Time features can capture the intensity of significant change information and statistical time series properties (section 3.1.4) and have been used as basic input data for thematic classification. The Random Forest classifier was applied in a number of experiments using different combinations of sensor data (representing different data scenarios) and time periods to benchmark their respective feasibility, effort and accuracy in view of the FOR product generation. For generation of the continuous-scale Tree Cover Density product in project phase 2, a multiple linear regression estimator has been used and different time features and periods have been tested. Testing in project phase 2 was characterized by an extension of the observation period with an increased cloud cover threshold (60% compared to 50% in phase 1). Table 3-27 provides an overview of the relevant parameters in both project phases.

Table 3-27: Sentinel data scenarios and time periods for Forest classification.

Project Phase 1			Project Phase 2		
Sensor	Time Period	Cloud Cover	Sensor	Time Period	Cloud Cover
S-2	01.01.-31.12.	50%	S-2	15.03.-15.09.	60%
			S-1	15.03.-15.09.	N/A
S-2	15.03.-15.06.	50%	S-2	01.06.-30.06.	60%
S-1	15.03.-15.06.	N/A	S-2	01.07.-31.07.	60%
S-2	01.01.-31.12.	50%	S-2	01.08.-31.08.	60%
S-1	15.03.-15.06.	N/A			
S-2	15.03.-15.06.	50%	S-2	01.06.-31.08.	60%
S-1	15.03.-15.06.	N/A			

According to the TCM and DLT classification tests carried out in WP 33, the use of Sentinel-2 data from the spring period was expected to provide the best ratio of high classification accuracy and lowest processing cost. On the other hand, the combined use of Sentinel-2 and Sentinel-1 for the

same period was expected to provide the highest classification accuracy by concurrent highest processing cost.

In summary, the following aspects have been investigated and implemented in the second project phase:

- Including of Sentinel-1 SAR time series data (timely consistent with S-2 features)
- Integration of additional Sentinel-2 time features, based on spectral bands
- Improved sampling based on a HRL2015 Sampling Layer
- Analysis and implementation of different feature selection methods

3.3.2.2 Benchmarking criteria

In addition to a traditional classification accuracy assessment (Overall accuracy, class specific producer and user accuracy, Kappa, F-score) 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.

3.3.2.3 Implementation and results of benchmarking

The following section focuses on the implementation of the benchmarking process, starting with the classification input data (section 3.3.2.3.1), followed by explaining the class separability analysis (section 3.3.2.3.2), the results of the classification (section 3.3.2.3.3) and the outcome of the benchmarking process.

3.3.2.3.1 Classification input data

The ECoLaSS FOR test sites comprise: North test site in Sweden covering two adjacent Sentinel-2 tiles (33VVF and 33VWF), South-East test site covering adjacent Sentinel-2 tiles (34TFM and 35TFL) and Central test site covering adjacent Sentinel-2 tiles 32UNU and 32TNT. For all test sites Sentinel-2 and Sentinel-1 data were processed.

In project phase 1 tests have been performed in the North test site in Sweden only. Sentinel-2 imagery was atmospherically corrected and topographically normalized using the ESA Sen2Cor software (Louis et al. 2016). Only scenes with a cloud cover lower than 50% were used for the classification and analysis. The cloud cover metrics do not rely on the official metadata cloud score provided by the original Sentinel-2 Level 1C product, but were calculated as part of the pre-processing chain using Sen2Cor to derive Level-2A data. Figure 3-40 shows the Sentinel-2 scene cloud cover distribution in the test site. Details on the pre-processing of EO data are described in the final issue of WP 32 [AD07]. Figure 3-41 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.

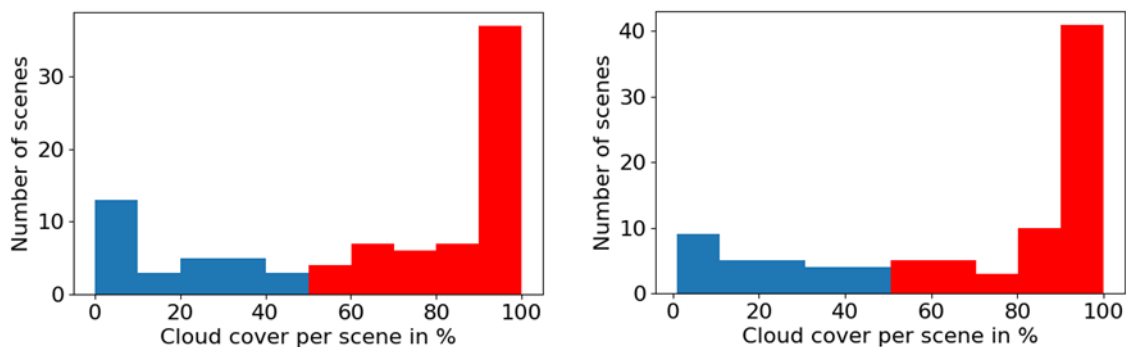


Figure 3-40: Cloud coverage of Sentinel-2 tile VVF (left) and VWF (right) of the test site North in Sweden.
 Blue: Scenes with < 50% cloud cover.

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 pre-processed to gamma naught (radar backscatter coefficient for an assumed ellipsoidal ground surface) and a multi-temporal filter was applied on the time series [AD07].

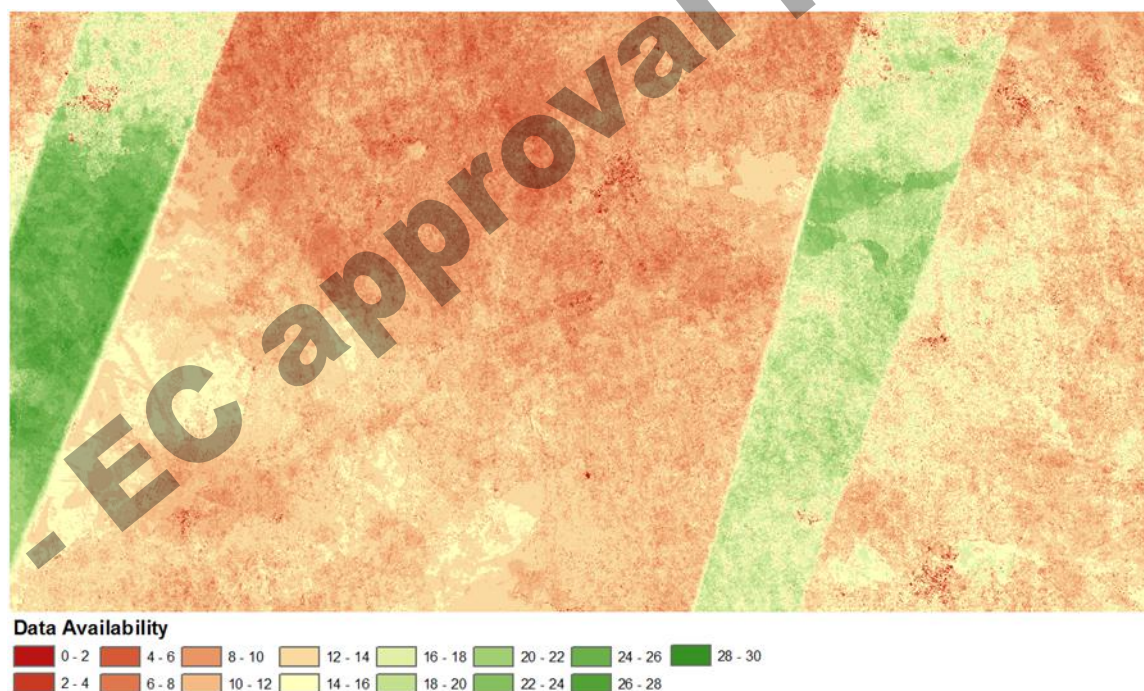


Figure 3-41: Sentinel-2 data score (number of cloud-free images) of scenes with average cloud cover < 50% for ECoLaSS north test site (VWF/VVF tiles), within the full year 2017.

Time features were calculated for the NDVI, NDWI, Brightness and IRECI indices using Sentinel-2 data, once for the full year 2017 and once for the spring period (15. March - 15. June 2017) for the Northern test site. For Sentinel-1, 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 for an analysis of the leaf type discrimination.

Two independent sample datasets were used for classification and validation. The training samples for coniferous and broadleaf forest were extracted from the combined HRL2015 Dominant Leaf Type product and the HRL2015 Grassland product. 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 DLT product classes (broadleaved, coniferous, and no tree cover). 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 broadleaved class to match the number of coniferous 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 DLT 2015 product and the sample enhancement process. For that, the DLT 2017 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 3-28 shows the sample and response design for the creation of the validation dataset, Table 3-29 the distribution of sample points of the training and validation dataset.

Table 3-28: 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 Sentinel-2 data pairs (spring/summer(in leaf)) as secondary source

Table 3-29: Sample distribution of training and validation dataset.

Class ID	Class name	Training data # polygons	Validation data # points
0	no tree cover	500	127
1	broadleaved	200	62
2	coniferous	200	141

In phase 2, new time features (see section 3.1.4) derived from the Sentinel-2 spectral bands (B02 to B12) have been calculated and analysed towards their importance in the Random Forest classification of both, the TCM and the DLT. In total, 234 features (compared to 160 features in phase 1) were available to feed the machine learning algorithm: 182 features for the Sentinel-2 indices and bands, and 52 features from the Sentinel-1 single bands and indices. For the TCD, features derived from the spectral bands have been assessed towards their fitness towards a seamless production on larger scale.

Besides the extension of the time feature portfolio, an improved sample base for the automated reference sampling has been generated, based on all available HRL2015 20m status layers (Dominant Leaf Type, Imperviousness, Grassland and Water) as exemplarily shown in Figure 3-42.

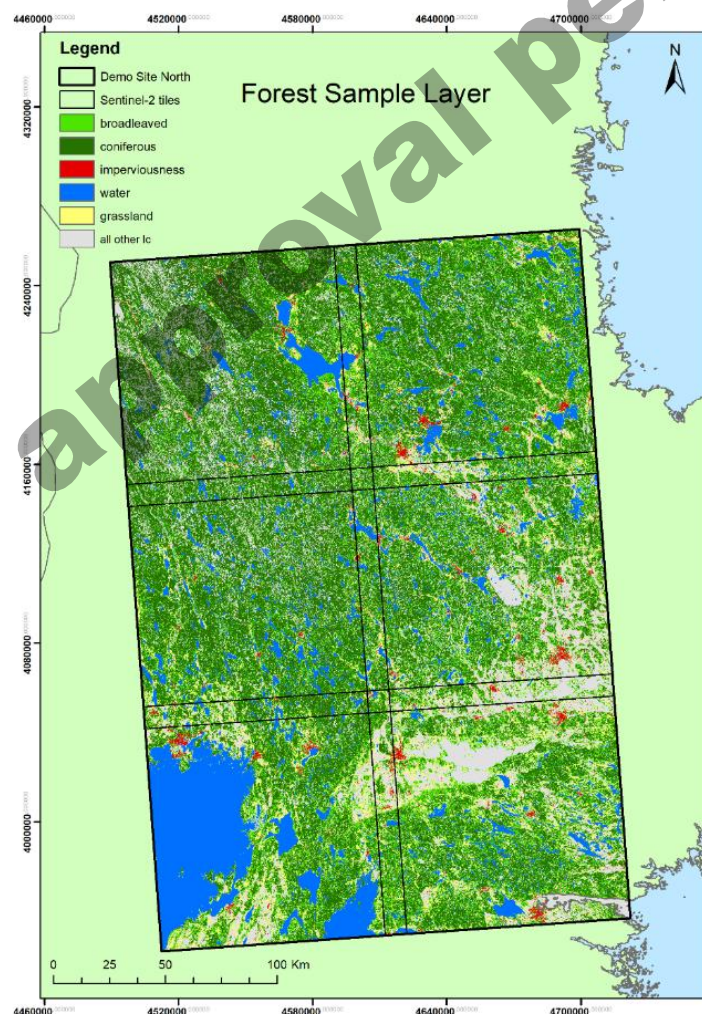


Figure 3-42: Example for the Sample Layer 2015 derived from the Copernicus High-Resolution Layers 2015.
(© EuroGeographics for the administrative boundaries)

The thereof derived Sample Layer (SLA) 2015 consists of 6 thematic classes (broadleaved trees, coniferous trees, imperviousness, grassland, water, all other land cover) from which training samples were automatically extracted (250 per tile) and provides a more sophisticated representation of the “no tree cover” class, compared to the phase 1 approach. The sampling is oriented towards the proportional coverage of tree cover (broadleaved & coniferous) and no tree cover (all other classes) within the test sites. Similar measures as in phase 1 were undertaken to reduce the number of outliers and errors:

1. Reduction of edge effects and mixed pixels through negative buffering (20m) of the HRL2015 SLA
2. Removal of patches smaller than 1ha
3. Systematic stratified random polygon (30m x 30m) sampling within the six SLA 2015 classes, following the proportional distribution of classes within the test site, considering the general Sentinel-2 data availability by incorporation of the Sentinel-2 Data Score Layer
4. Removal of sampling errors through scatter-plot analysis based on time features
5. Iterative resampling and outlier detection

This approach pursues a generally wider representation of forest and non-forest samples for generation of the tree cover status maps as input for all FOR prototypes.

3.3.2.3.2 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 3-43 and Figure 3-44 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 and brightness indices, whereas the multiple simple features for spectral bands 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.

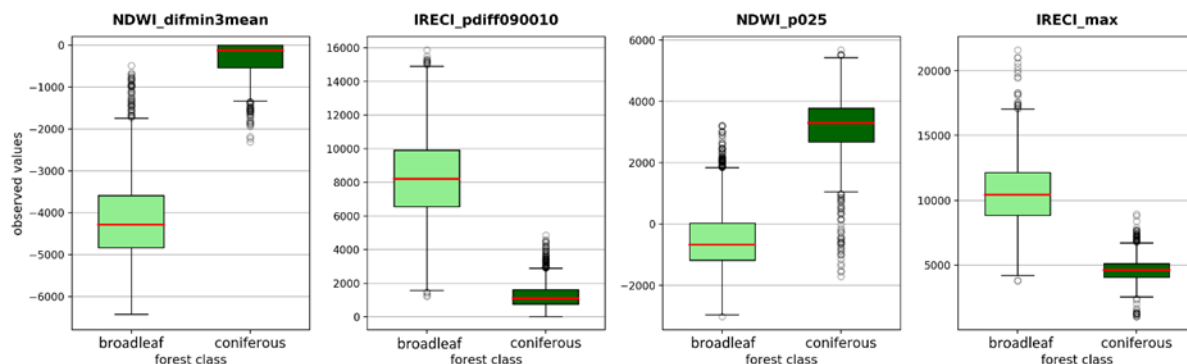


Figure 3-43: Forest class separability box plots for selected Sentinel-2 time features.

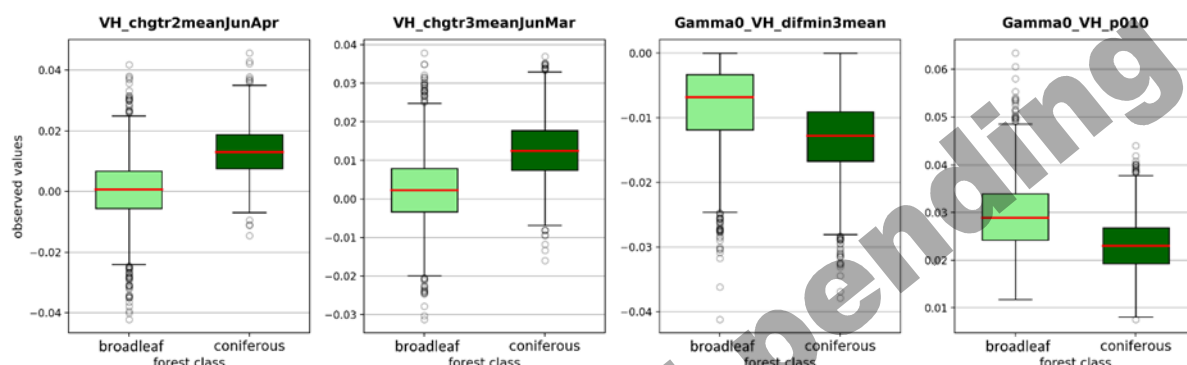


Figure 3-44: Forest class separability box plots for selected Sentinel-1 time features.

In project phase 2, further analyses have been performed towards the forest class separability using NDVI profiles from consecutive years (2017, 2018) and additional band-specific time features derived from Sentinel-2 data. NDVI profiles have been collected in all test sites for selected coniferous and broadleaf trees stands and are presented in the figures below.

In the northern site, the NDVI values of broadleaf and coniferous forest can be clearly separated in the spring period (March to June) as well as in the winter months (see Figure 3-45). For the summer period, NDVI values are slightly different in a short time window ranging from July to September. This might be challenging when the leaf type classification relies on NDVI features as input only.

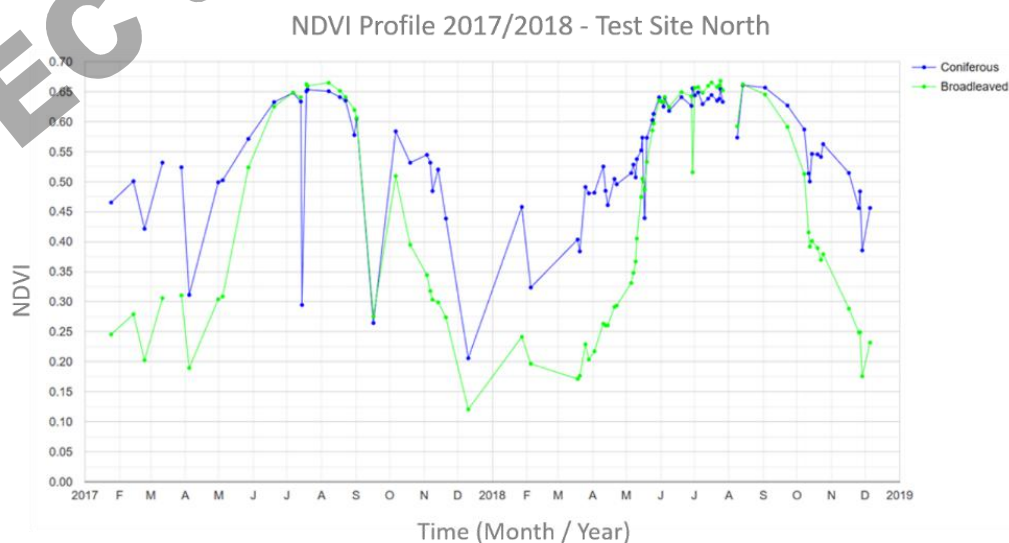


Figure 3-45: NDVI profiles 2017/2018 for broadleaf and coniferous tree stands in the North test site.

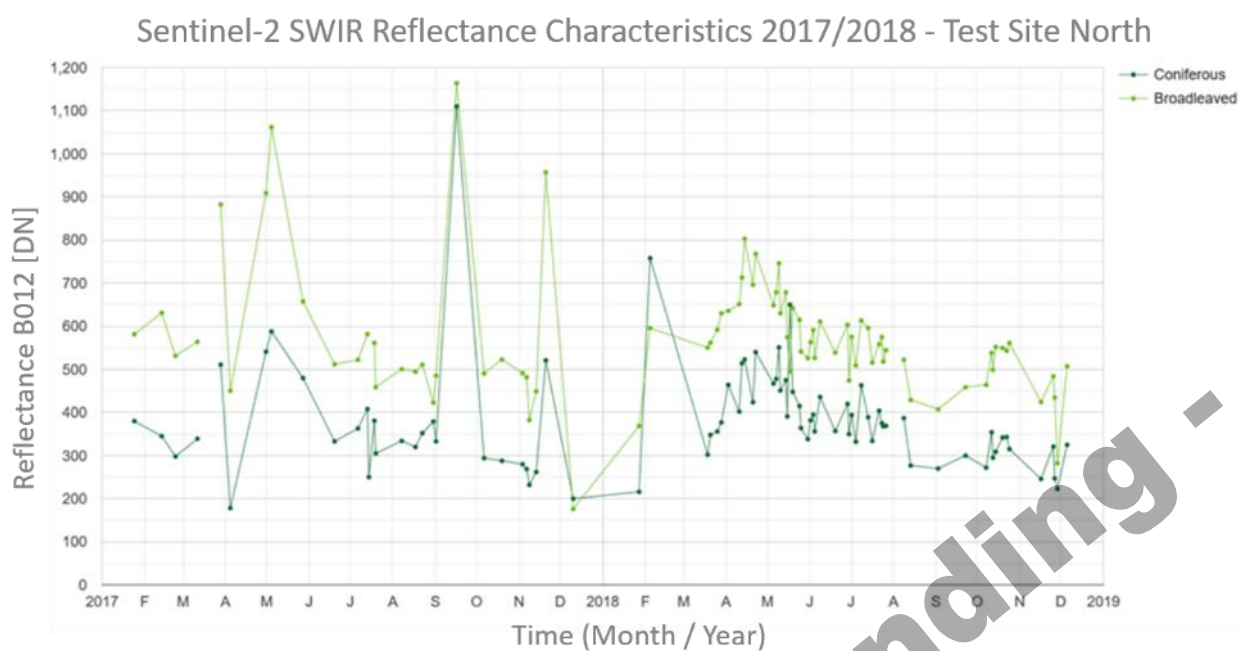


Figure 3-46: Sentinel-2 SWIR Band (B12) reflectance characteristics for broadleaf and coniferous tree stands.

Unlike the NDVI values for the northern test site, there is a clear difference in the spectral response of broadleaf and coniferous trees in the SWIR band (B12) of Sentinel-2 (see Figure 3-46), making it well suited for a leaf type discrimination within the selected spring/summer period. This is also underpinned by the feature importance of the SWIR band in the random forest classification across all test sites (see section 3.3.2.3.3).

In the Central test site, one can notice strong differences towards summer months (Figure 3-47). This behaviour in the NDVI profile was key for the selection of the final time period for the time feature calculation.

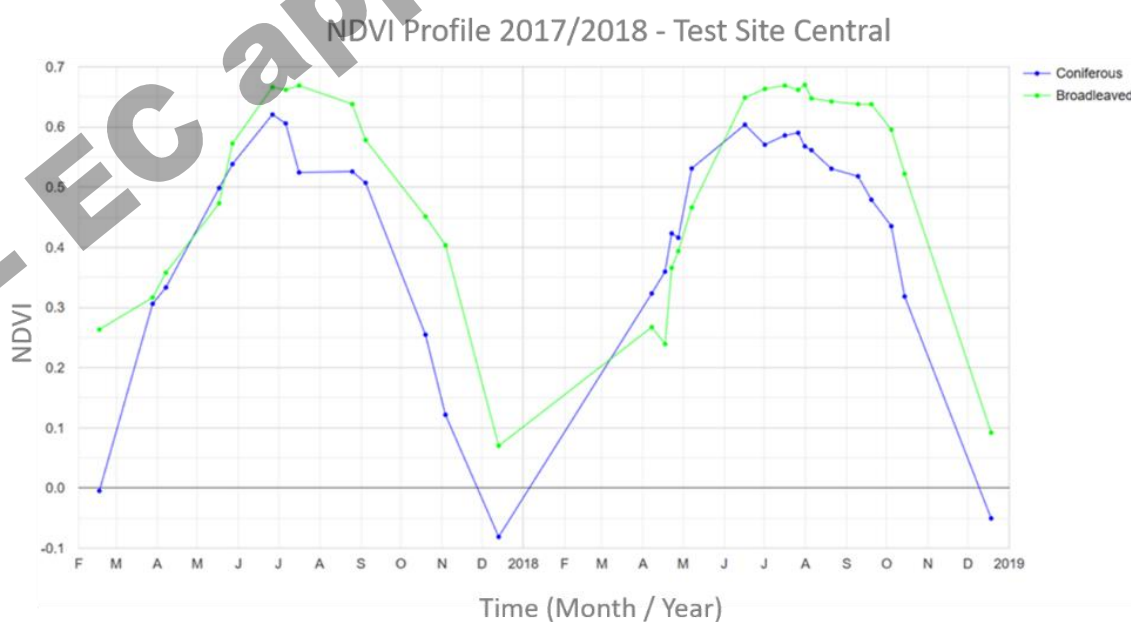


Figure 3-47: NDVI profiles 2017/2018 for broadleaf and coniferous tree stands in the Central test site.

A similar pattern in the NDVI values for the two forest classes is also evident in the South-East site, with higher values for broadleaf beginning in the spring (see Figure 3-48). This highlights a different phenology compared to the test sites in Sweden (North) and Austria/Germany (Central).

From the figures above and TCM & DLT tests results, the class separability analysis shows the highest importance correspond to the time features with strongest delta value in the time series (e.g. broadleaf seasonal patterns versus more stable coniferous). NIR (B08) and SWIR (B11, B12) band derived features from Sentinel-2 are of high importance for the leaf type discrimination and consequently highly recommended for the thematic classification.

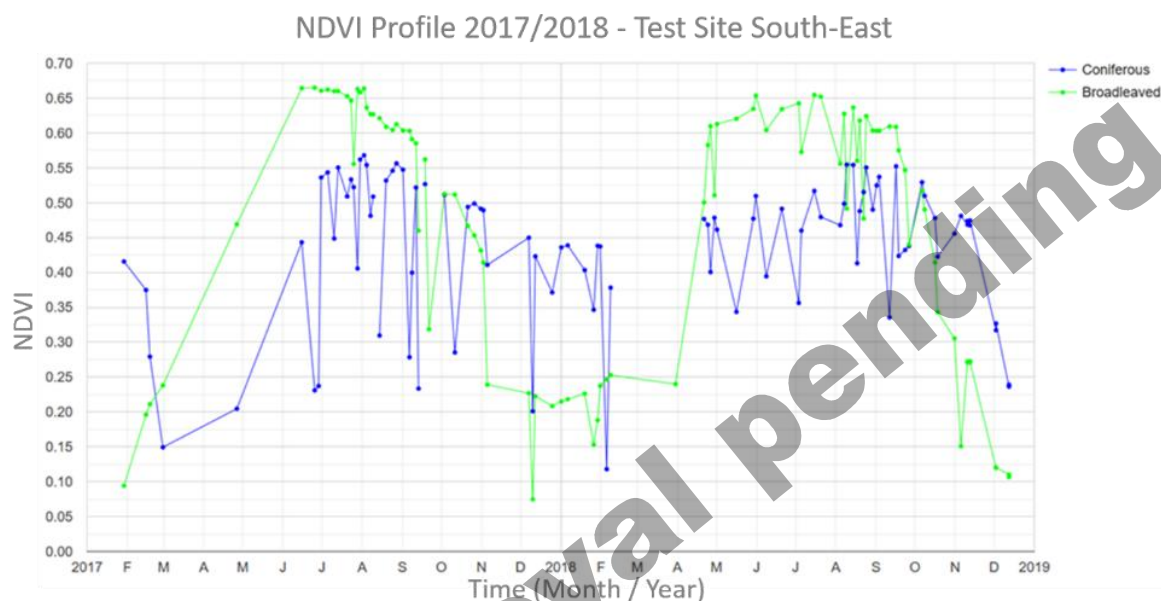


Figure 3-48: NDVI profiles 2017/2018 for broadleaf and coniferous tree stands in the South-East test site.

In conclusion, the analyses performed confirm the importance of the spring period for the Dominant Leaf Type classification and a strong added value towards the integration of features derived from the NIR and SWIR bands of Sentinel-2.

3.3.2.3.3 Classification results

In phase 1, the classification was carried out using a random forest classifier with preceding recursive feature elimination. The general accuracy metrics of the different input data configurations can be seen in Table 3-30 and Figure 3-49, while Table 3-31 shows the class specific results.

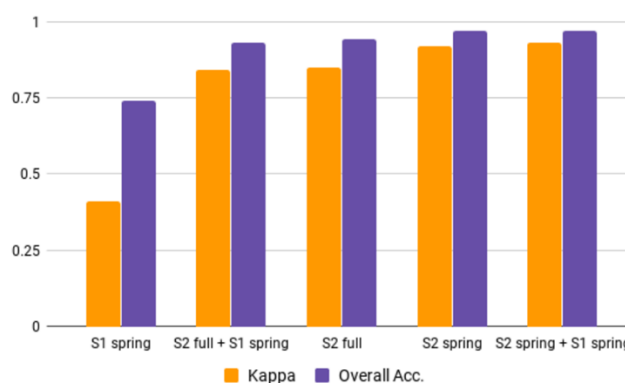


Figure 3-49: Kappa and overall accuracy for the five DLT input data configurations.

Table 3-30: Accuracy metrics for the five DLT input data configurations.

	User's Acc.	Producer's 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 3-31: User and producer accuracy 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% in the Northern test site. 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. As mentioned above, Sentinel-2 only (spring period) provides the best results with an overall accuracy of 97%. The combination of Sentinel-2 and Sentinel-1 features does not add any significant gain in accuracy. Figure 3-50 shows a detailed view of the Sentinel-2 and Sentinel-1 spring period classification map.



Figure 3-50: 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). *Modified Copernicus Sentinel data [2017].*

In project phase 2, classification tests have been extended to all FOR test sites. Based on the experiences and lessons learned from Task 4 within the first project phase, several analyses and measures have been undertaken to improve the results. This concerns the generation of the Tree Cover Mask (TCM) as basic product for all FOR prototypes, as well as the Dominant Leaf Type (DLT) product and the continuous-scale Tree Cover Density (TCD) product.

TREE COVER & DOMINANT LEAF TYPE MAPPING

After some sobering results of the phase 1 tree cover classification in the demonstration site North due to an insufficient data situation within the selected spring period for most of the tiles, the utilization of Sentinel-1 SAR data has been reconsidered in phase 2 to improve the tree cover detection. An improved Tree Cover Mask at 10 m spatial resolution is calculated from time features derived from Sentinel-2 and Sentinel-1 SAR data of the spring/summer period. Compared to the results of project phase 1, the integration of SAR data in the TCM classification has reduced the number of omission errors over cloudy areas significantly. Simultaneously, commission errors could be drastically reduced. This is especially the case for agricultural areas (hops, vineyards, maize fields), moors and wetlands. This observation could be made in all test sites and thus supporting the integration of SAR data in the classification process. In this context, SAR data contributes to a reliable tree cover detection, which is mandatory for generation of the Dominant Leaf Type and Tree Cover Density products (by masking) as well as a reliable map-to-map change approach in form of the Incremental Update Layer 2017-2018.

SAR time features derived from the VH polarisation (e.g. VH_p025, VH_p010) turned out to have a high importance in the tree cover detection followed by features derived from the Sentinel-2 bands B02 and B03 as well as NDVI and NDWI features. With respect to the DLT classification, SAR features show no benefit for the leaf type discrimination. Here, band-specific features from Sentinel-2 dominate clearly over all other derived features. Worth mentioning is the dominance of features derived from the SWIR bands (B11, B12). Figure 3-51 presents the top 20 ranking of the time feature importance for the TCM and DLT classifications of the Central test site for the reference year 2018.

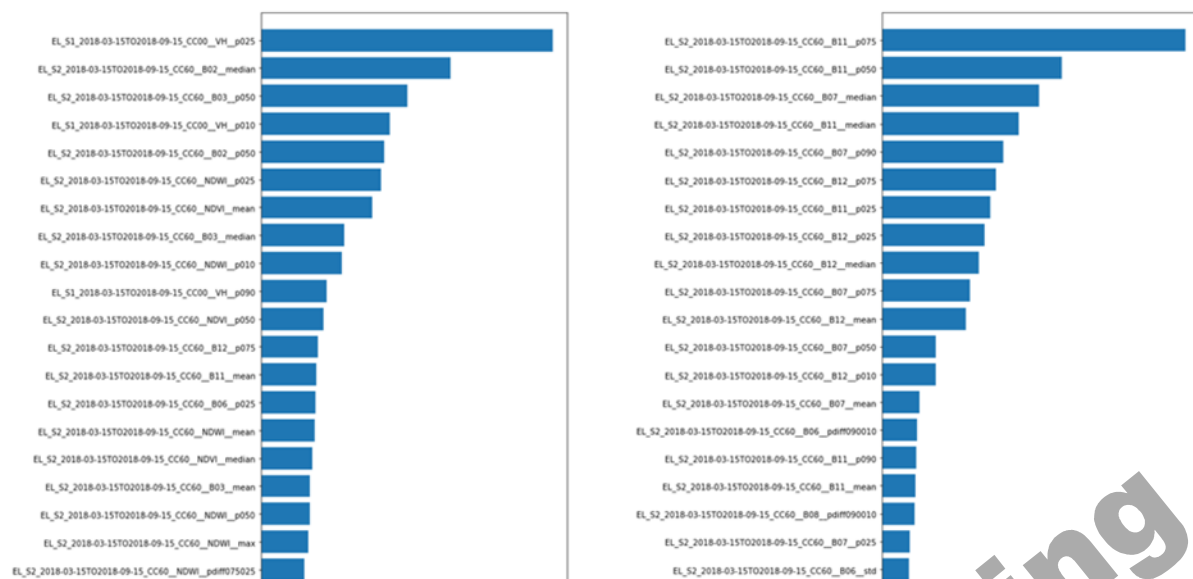


Figure 3-51: Top 20 time feature ranking for combined S1/S2 TCM 2018 (left) and DLT 2018 classification (right).

In contrast, and with respect to the DLT classification, the influence of SAR time features within a combined Sentinel-2 and Sentinel-1 SAR approach plays a minor role only. Sentinel-1 time features do not appear in the top 20 ranking, which underpins the importance of optical data for the classification of the dominant leaf types. In general, the integration of SAR data in the classification process show high processing costs in terms of processing time and additional storage costs. Consequently, a well-balanced usage of SAR data for status layer generation is recommended for an operational service on continental or global scale. This can be addressed by an elaborated stratification approach of the production area.

In parallel to the new feature calculation and analysis, two different feature selection methods have been tested. The first one is based on the statistical analysis of variance of the features and the target class to be classified. The process comprises the comparison of the mean values in every feature for the land cover class “tree cover” vs other land cover classes (urban, water, grassland, and cropland), using an Analysis of Variance and a post-hoc analysis (Tukey test). The features selected under this feature selection scheme are those with a significant statistical difference between forest and other land cover classes. In the case of the Dominant Leaf Type the statistical analysis was made on the mean values of two classes (broadleaved and coniferous), using a Student’s t-test.

The second feature selection method is the K-Fold Cross Validation. This is based on a stratified k fold sampling integrated in the machine learning package. This sampling method splits the training and test dataset into a number of k-folds. Subsequently, it clones the classifier by every iteration and produces accuracy figures and a new training and test set. The algorithm finally yields a combination of the features with the highest accuracy. This subset of features is used for the classification process.

In terms of processing costs (processing time), the variance analysis seems to be superior to the K-Fold feature selection, as it provides similar model accuracy figures by using significantly less features. A significant influence of the feature selection method on the overall accuracy figures (retrieved by LUCAS 2018 points) could not be observed. Figure 3-52 shows an example of the number of time features versus accuracy performance for the TCM 2018 test classification in the Central site (time feature selection method K-fold cross-validation):

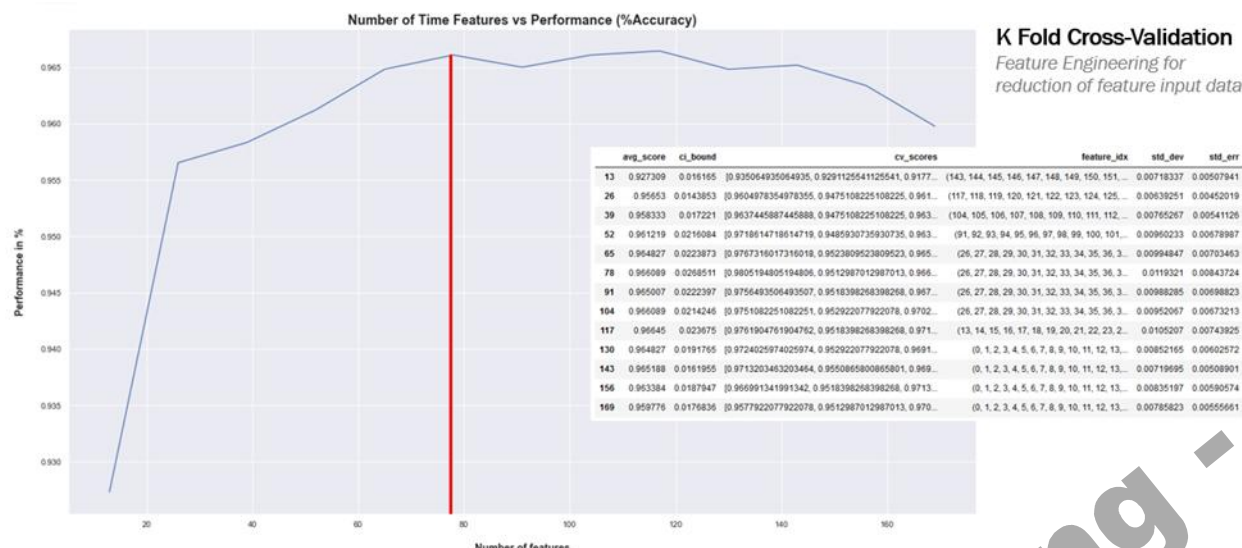


Figure 3-52: Number of time features versus TCM 2018 accuracy performance for test site Central.

The following figures show respectively the final settings (blue square) for generation of the Tree Cover Masks and Dominant Leaf Type products, considering the performance cost and accuracy for selected time windows, based on tests carried out for the number of input time features, features election method and sensors.

Given that the overall data situation is good or even very good, the spring period is sufficient to retrieve a high quality Tree Cover Mask using optical Sentinel-2 data only (Figure 3-53). The feature selection method has a high impact on the number of features to be classified and consequently a significant influence on the processing costs while retaining high accuracy. From an economic point of view, a features selection method is highly recommended for large scale applications. The highest accuracy could be achieved using a combined Sentinel-1/Sentinel-2 classification approach with extended observation period (spring/summer) and application of a feature selection method.

Experiment	TCM-01	TCM-02	TCM-03	TCM-04	TCM-05	TCM-06	TCM-07	TCM-08	TCM-09	TCM-10
Time Window	15-03-2018 TO 15-06-2018	15-03-2018 TO 15-06-2018	15-03-2018 TO 15-06-2018	15-03-2018 TO 15-06-2018	15-03-2018 TO 15-06-2018	15-03-2018 TO 15-06-2018	15-03-2018 TO 15-06-2018	15-03-2018 TO 15-06-2018	15-03-2018 TO 15-09-2018	15-03-2018 TO 15-09-2018
Cloud Cover	60%	60%	60%	60%	60%	60%	60%	60%	60%	60%
Number of Features	182	234	52	52	234	117	63	56	64	234
Satellites	Sentinel-2	Sentinel-2	Sentinel-1 Sentinel-2	Sentinel-2	Sentinel-1 Sentinel-2	Sentinel-1 Sentinel-2	Sentinel-1 Sentinel-2	Sentinel-1 Sentinel-2	Sentinel-1 Sentinel-2	Sentinel-1 Sentinel-2
Feature Selection	N/A	N/A	KFold	KFold	N/A	KFold	Phase 1 Features + S1 integration	Statistics - Variance Analysis	Statistics - Variance Analysis	N/A
Overall Accuracy LUCAS 2018	0.94	0.96	0.95	0.96	0.94	0.94	0.90	0.91	0.98	0.96

Final Settings

Figure 3-53: Parameter set evaluation for Tree Cover Mapping in the Central test site.

Experiment	DLT-01	DLT-02	DLT-03	DLT-04	DLT-05	DLT-06
Time Window	Mar - Jun	Mar - Jun	Mar - Sep	Mar - Sep	Mar - Sep	Mar - Sep
Cloud Cover	60%	60%	60%	60%	60%	60%
Number of Features	95	65	147	104	91	145
Sensors	Sentinel-1 Sentinel-2	Sentinel-1 Sentinel-2	Sentinel-1 Sentinel-2	Sentinel-1 Sentinel-2	Sentinel-1 Sentinel-2	Sentinel-1 Sentinel-2
O. Accuracy – Machine Learning Model	0.95	0.95	0.98	0.99	0.96	0.98
Overall Accuracy LUCAS 2018	0.88	0.89	0.84	0.84	0.93	0.92

Final Settings

Figure 3-54: Parameter set evaluation for Dominant Leaf Type Mapping in the Central test site.

The highlighted settings proved to provide overall high accuracies and could be successfully transferred to all test sites. However, the strongest influence on classification accuracies can be observed by the quality of the samples. In ECOLaSS, a multi-stage sampling approach has been applied: automatic reference sampling based on a sample layer generated from HRL2015 products, outlier detection with visual validation of the samples and split of the sample dataset into training and validation dataset, initial classification and re-sampling based on omission and commission errors with subsequent iteration loops. A subsequently performed validation with LUCAS 2018 points confirmed very high overall accuracies for all TCM and DLT classifications. The following tables provide the error matrices for all TCM and DLT test products within the three test sites.

Table 3-32: Error matrix for the improved TCM 2018 of the test site Sweden

TCM_2018_010m Sweden		REFERENCE			User Accuracy	Confidence Interval
		No Tree Cover	Tree Cover	Total		
PRODUCT	No Tree Cover	514	36	550	93.45%	91.30 – 95.61%
	Tree Cover	12	361	373	94.68%	95.77 – 98.71%
	Total	526	397	923		
	Producer Accuracy	97.72%	90.93%		94.80%	Overall Accuracy
	Confidence Interval	96.35 – 99.09%	87.98 – 93.88%		93.31 – 96.21%	Confidence Interval
					0.955	F-Score No Tree Cover
					0.937	F-Score Tree Cover
					0.893	Kappa

Table 3-33: Error matrix for the improved TCM 2018 of the test site Austria/Germany

TCM_2018_010m Austria/Germany		REFERENCE			User Accuracy	Confidence Interval
		No Tree Cover	Tree Cover	Total		
PRODUCT	No Tree Cover	664	8	672	98.81%	97.92 – 99.70%
	Tree Cover	10	150	160	93.75%	89.69 – 97.81%
Total		674	158	832		
Producer Accuracy		98.65%	94.94%		97.84%	Overall Accuracy
Confidence Interval		97.53 – 99.50%	91.20 – 98.67%		96.79 – 98.89%	Confidence Interval
					0.986	F-Score No Tree Cover
					0.943	F-Score Tree Cover
					0.930	Kappa

Table 3-34: Error matrix for the improved TCM 2018 of the test site Bulgaria/Greece

TCM_2018_010m Bulgaria/Greece		REFERENCE			User Accuracy	Confidence Interval
		No Tree Cover	Tree Cover	Total		
PRODUCT	No Tree Cover	557	24	581	95.87%	94.16 – 97.57%
	Tree Cover	14	70	84	83.33%	74.77 – 91.0%
Total		571	94	665		
Producer Accuracy		97.55%	74.47%		94.29%	Overall Accuracy
Confidence Interval		96.19 – 98.90%	65.12 – 83.81%		92.45 – 96.13%	Confidence Interval
					0.967	F-Score No Tree Cover
					0.786	F-Score Tree Cover
					0.753	Kappa

Table 3-35: Error matrix for the improved DLT 2018 status layer of the test site Sweden

DLT_2018_010m Sweden		REFERENCE				User Accuracy	Confidence Interval
		No Tree Cover	Broadleaved	Coniferous	Total		
PRODUCT	No Tree Cover	514	27	9	550	93.45%	91.30 – 95.61%
	Broadleaved	6	125	17	148	84.46%	79.50 – 89.42%
Coniferous		6	16	203	225	90.22%	86.12 – 94.33%
Total		526	168	229	923		
Producer Accuracy		97.72%	74.40%	88.65%		91.22%	Overall Accuracy
Confidence Interval		96.35 – 99.09%	67.51 – 80.28%	84.32 – 92.97%		89.34 – 93.10%	Confidence Interval
						0.955	F-Score No Tree Cover
						0.791	F-Score Broadleaved
						0.894	F-Score Coniferous
						0.846	Kappa

Table 3-36: Error matrix for the improved DLT 2018 status layer of the test site Austria/Germany

DLT_2018_10m Austria/Germany		REFERENCE				User Accuracy	Confidence Interval
		No Tree Cover	Broadleaved	Coniferous	Total		
PRODUCT	No Tree Cover	664	6	2	672	98.81%	97.92 – 99.70%
	Broadleaved	6	88	7	101	87.13%	77.74 – 96.52%
	Coniferous	4	5	50	319	84.75%	74.72 – 94.77%
	Total	674	99	62	832		
	Producer Accuracy	98.52%	88.89%	80.65%		96.03%	Overall Accuracy
	Confidence Interval	97.53 – 99.50%	82.19 – 97.76%	70.00 – 91.29%		95.07 – 97.72%	Confidence Interval
						0.986	F-Score No Tree Cover
						0.880	F-Score Broadleaved
						0.847	F-Score Coniferous
						0.889	Kappa

Table 3-37: Error matrix for the improved DLT 2018 status layer of the test site Bulgaria/Greece

DLT_2018_10m Bulgaria/Greece		REFERENCE				User Accuracy	Confidence Interval
		No Tree Cover	Broadleaved	Coniferous	Total		
PRODUCT	No Tree Cover	557	21	3	581	95.87%	94.16 – 97.57%
	Broadleaved	12	32	3	47	68.09%	51.71 – 84.46%
	Coniferous	2	7	28	37	75.68%	60.50 – 90.85%
	Total	571	60	34	665		
	Producer Accuracy	97.55%	53.33%	82.35%		92.78%	Overall Accuracy
	Confidence Interval	96.19 – 98.90%	39.88 – 71.57%	68.07 – 96.64%		90.74 – 94.82%	Confidence Interval
						0.967	F-Score No Tree Cover
						0.598	F-Score Broadleaved
						0.788	F-Score Coniferous
						0.699	Kappa

Although the target threshold of 90 % overall accuracy could be exceeded within each test site for all products, results for the test site Bulgaria/Greece should be treated with caution. Here, no full LUCAS 2018 coverage was available, and data situation was the worst compared to the other sites. Besides already mentioned artefacts coming from the inadequate cloud-masking, dry related-effects hindered a proper tree cover detection. The thereof resulting omission errors are negatively influencing the producer accuracies of both, TCM and DLT products. However, looking at the results on demonstration site level (see “D42.1b - Prototype Report: Consistent HR Layer Time Series/Incremental Updates”, Issue 2), results are quite better thanks to a more balanced distribution of LUCAS 2018 points.

TREE COVER DENSITY

In addition to the aspects listed above, project phase 2 has also concentrated on the generation of an improved continuous-scale Tree Cover Density (TCD) product at 10 m spatial resolution using optical Sentinel-2 data. The pixel-based TCD product provides information on the proportional crown coverage per pixel in percent (0-100%), whereas tree cover density is defined as the „vertical projection of tree crowns to a horizontal earth’s surface“. It is well suited to map spatial patterns, but is phenological and radiometrically sensitive.

Two different approaches have been used to generate the status layer Tree Cover Density 2018 using a linear regression estimator: a mono-temporal classification using a “best-of” scene approach and a multi-temporal classification using band-specific time features for defined time windows. Samples (300 per tile) have been automatically collected from the TCD 2015 product. Outliers have been removed in frame of the scatter plot analysis based on a threshold approach. Finally, 268 samples entered in the classification of each input data stack. Table 3-38 provides the parameters for the TCD time feature testing.

Table 3-38: Parameter testing for time series TCD classification.

Parameter sets for TCD Time Feature Testing				
Bands	Feature	Time Period	Cloud Cover	No. of Samples
B02, B03, B04, B05, B06, B07, B08, B08A, B11, B12	<ul style="list-style-type: none">• Mean• Median• p010• p025	01.06.-30.06.	60%	300
B02, B03, B04, B08, B11, B12		01.07.-31.07.		
		01.08.-31.08.		
		01.06.-31.08.		

Band-specific spatio-temporal features (each 10 m & 20 m band) have been tested for a multi-temporal TCD classification. Whereas most of the features show no suitability for the classification, results of the median features provide very promising results. Figure 3-55 shows certain time periods for band-specific median time feature stacks from within the summer months. From experience, these windows show the highest potential for good weather conditions with low cloud cover rates. Whereas single month show divergent cloud conditions with remaining clouds and nodata gaps (highlighted in blue), the time period 01.06.to 31.08. can be rated the best in terms of completeness and overall data quality. Remaining nodata areas are related to snow and ice cover, which is not relevant for the HRL Forest. Compared to the monthly feature stacks and stacks derived from the “mean”, “p010” and “p025” time features, the median time feature stack shows only few artefacts in the landscape and has been finally selected for the Tree Cover Density classification within ECoLaSS.

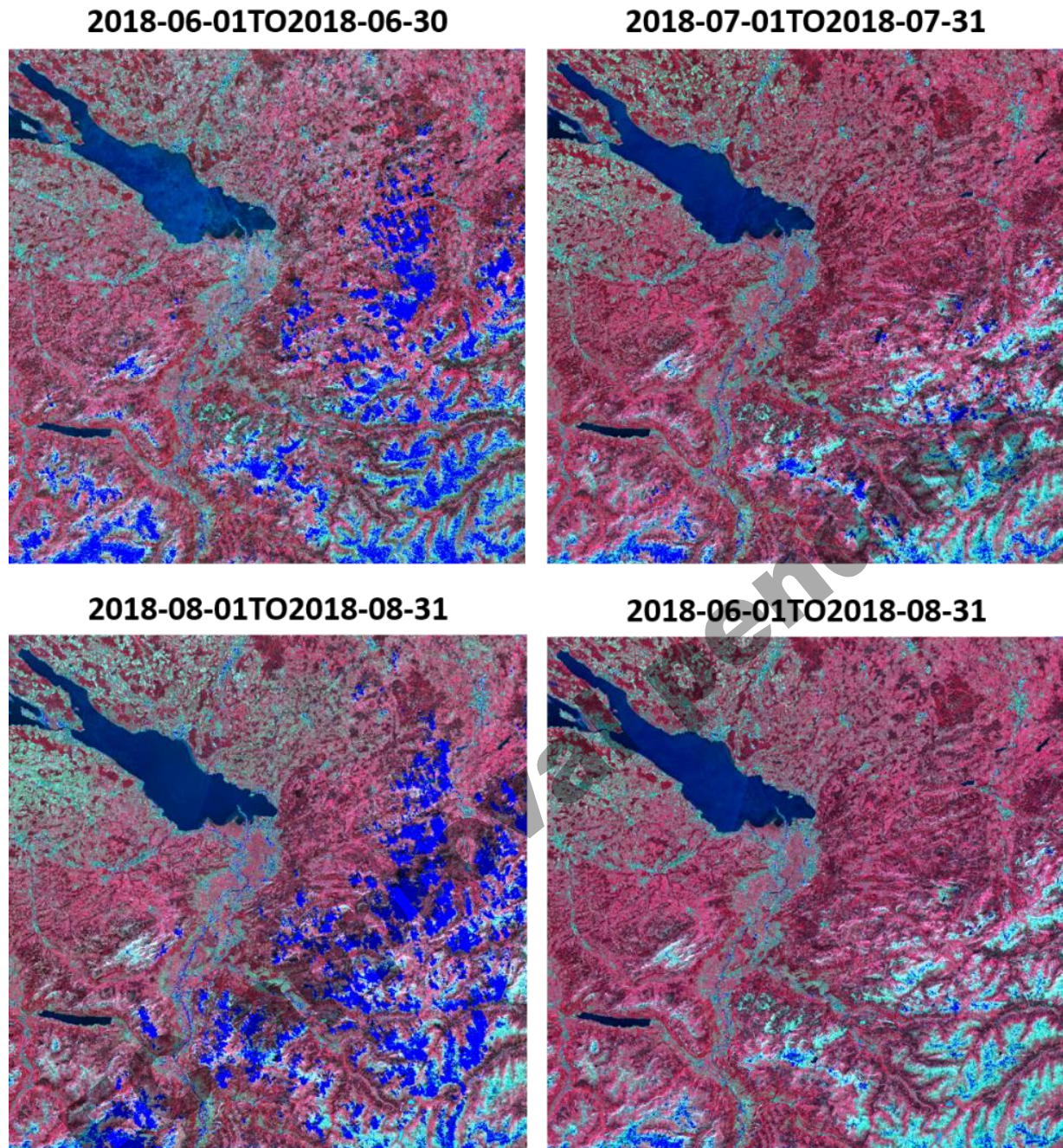
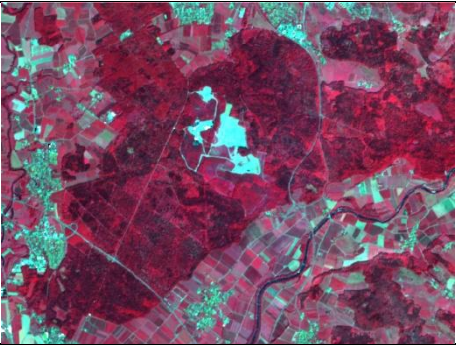
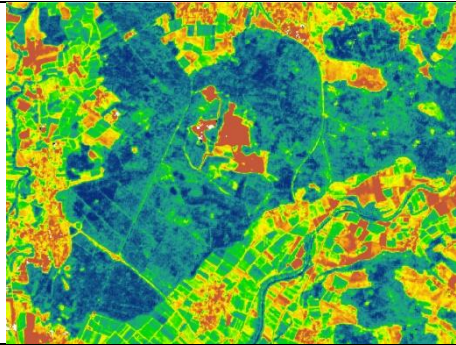

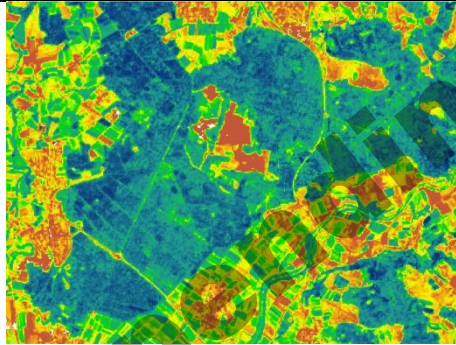


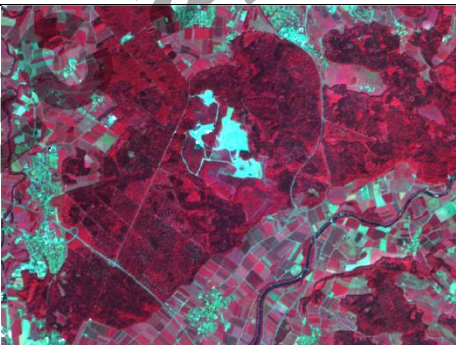
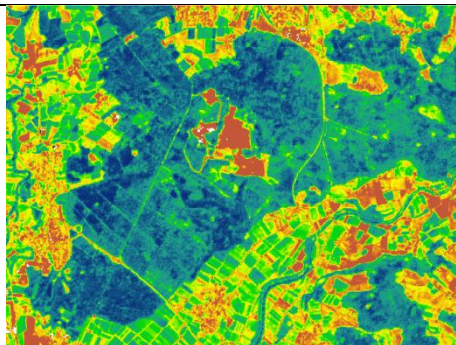


Figure 3-55: Sentinel-2 median time feature stacks for Tree Cover Density classification. Blue areas represent remaining clouds, nodata gaps and/or snow and ice cover. Modified Copernicus Sentinel data [2018].

Classification results have been compared with results derived from the Copernicus Sentinel-2 Global Mosaic (S2GM - <https://land.copernicus.eu/imagery-in-situ/global-image-mosaics/>) and a cloud-free, mono-temporal “best-of” scene acquired in June 2018. For this purpose, the same sample dataset as for the TCD time feature testing exercise has been used. Figure 3-56 shows the results of the TCD time feature testing and benchmarking exercise.

Input Data	Imagery	TCD 2018	Accuracy Figures
Mean time features			$R^2 = 0.9341$ $MAE = 7.43$ $RMS = 9.43$
p010 time features			$R^2 = 0.89825$ $MAE = 9.30$ $RMS = 12.42$
p025 time features			$R^2 = 0.91652$ $MAE = 8.49$ $RMS = 11.29$
Median time features			$R^2 = 0.93109$ $MAE = 7.66$ $RMS = 9.67$

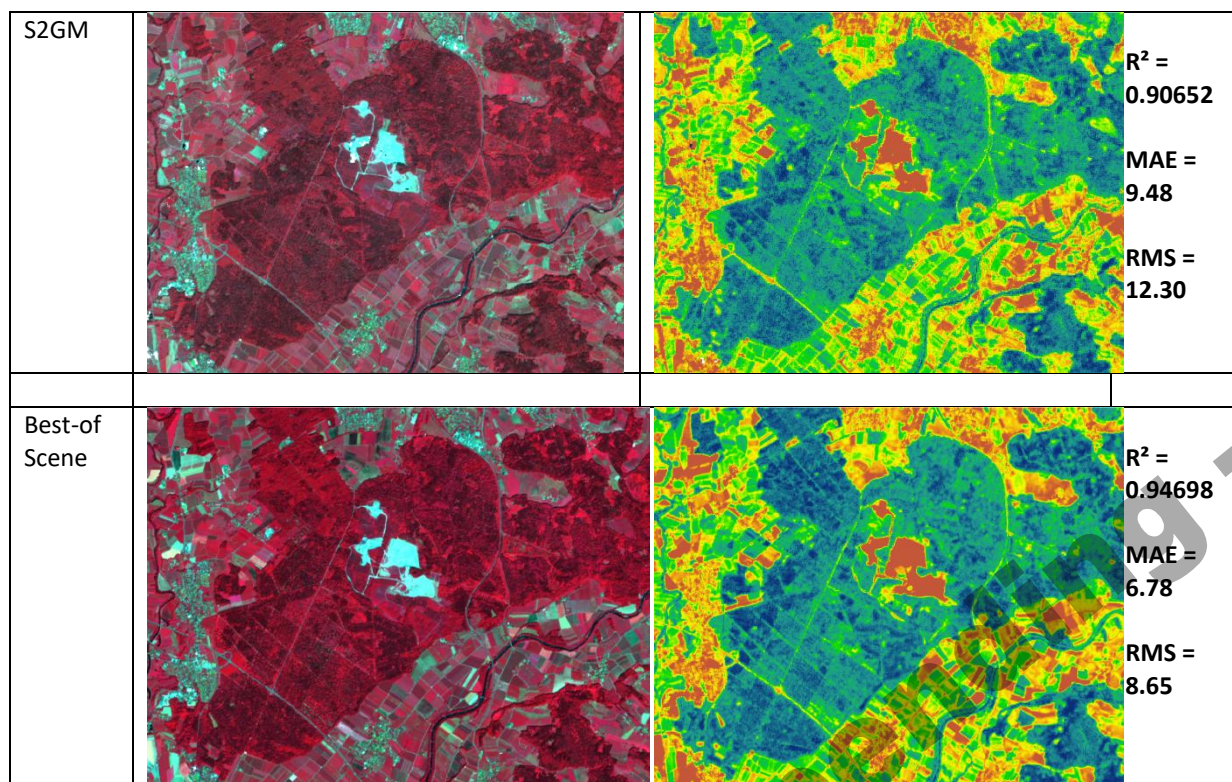


Figure 3-56: Comparison of Tree Cover Density classification results based on different input data.
Modified Copernicus Sentinel data [2018].

The following observations have been made in the evaluation process of the feature-based TCD:

- Length of the selected time window for feature computation is crucial to avoid cloud gaps
- Inadequate cloud masks lead to artefacts in the TCD classification
- Overcorrections in the topographic normalization (performed by Sen2Cor) lead to significantly lower TCD values than may be realistically the case
- High agreement with results obtained from the mono-temporal Sentinel-2 scene
- Much less artefacts compared to the classification derived from the S2GM

As expected, the single scene classification based on a cloud-free “best-of” scene provides the best results with a high level of detail and an $R^2 = 0.95$, followed the results obtained by mean features ($R^2 = 0.93$) and Median time features ($R^2 = 0.93$). The percentile time features and the S2GM provide significantly lower R^2 values and show lots of artefacts, which are negatively influencing the “look & feel” of the product. This is especially truth for the S2GM data. Main issues are referred to the topographic normalization of the input data and the quality of the derived cloud masks (see Figure 3-57) as reported in WP 32 [AD07]. Some further research and improvements are necessary to compensate these effects in the time series TCD classification.

As for the TCD, band-specific time features (mean and median) of Sentinel-2 are suitable to derive the continuous-scale TCD at 10 m resolution. Even though median features provide slightly poorer results in the regression model, they provide a much better “look & feel” thanks to overall less artefacts. For this reason, median features have been finally selected for the TCD classifications in all test sites.

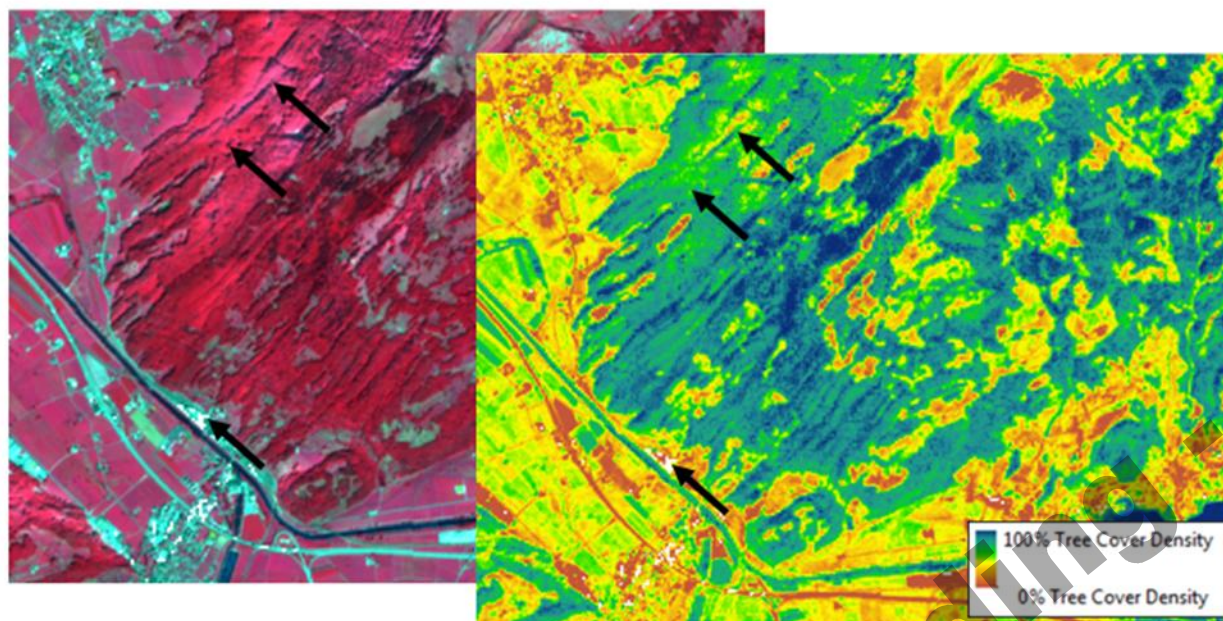


Figure 3-57: Sentinel-2 median time feature stack (01.06.-31.08.2018) and unmasked Tree Cover Density. Arrows point to issues caused by the terrain correction and cloud masking. Modified Copernicus Sentinel data [2018].

The ECOLaSS prototype in 10m resolution shows much more details than the TCD 2015 based on 20m satellite imagery (see Figure 3-58). Notwithstanding, a direct comparison of the two products is not recommended. This is mainly due to scaling issues (20 m vs 10 m native resolutions) and radiometric inconsistencies (productions based on varying input spectral bands e.g., Landsat 8, Sentinel-2A, IRS LISS-III, SPOT-5 in 2015 vs Sentinel-2A+B in 2018). Furthermore, a phenology mismatch due to varying acquisition dates in the 2015 production (2014-2016 acquisitions with different sun angles) has been reported and varying acquisition angles coming from different satellites/sensors led to shadow-related overestimations of tree cover at the fringe of forest borders.

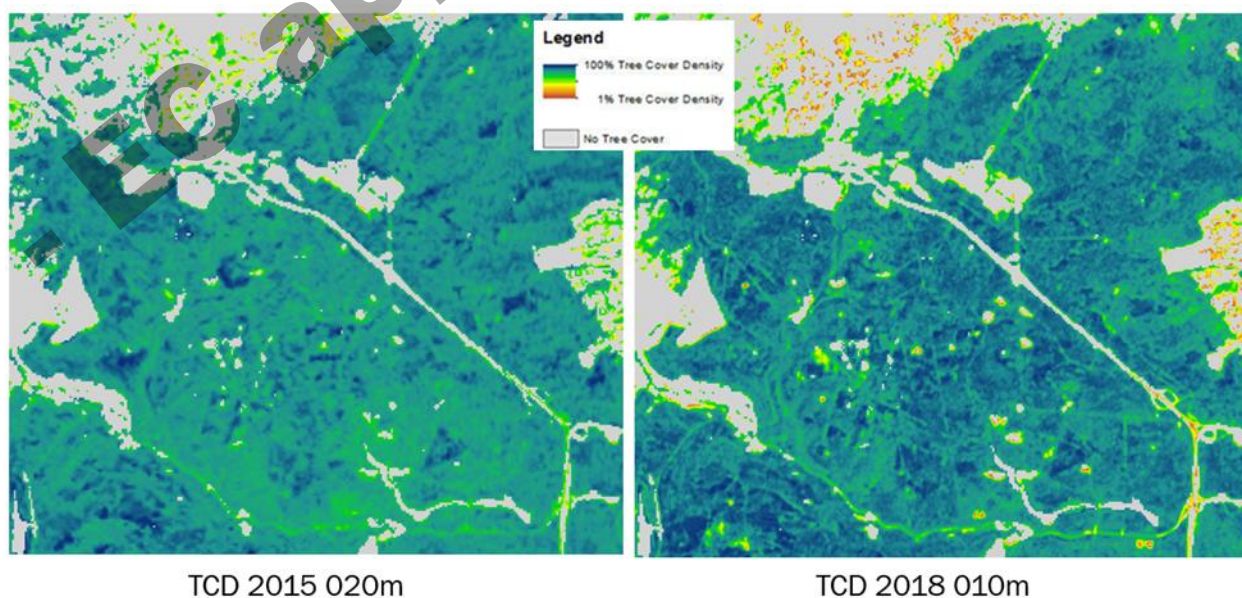


Figure 3-58: Comparison of the TCD 2015 (20 m) and the improved TCD 2018 at 10 m spatial resolution. Produced using modified Copernicus Sentinel data [2016/2018].

In the tests carried out in phase 2, and compared to the TCD production in 2015, there has been no time-consuming scene selection process from scenes acquired within the vegetation period. In general, seamline creation between tiles/scenes are being dropped and there is no need for a cloud gap-filling, which can be often exhaustive. Instead, a high performance is achieved through an automated sampling approach (outlier detection through scatter plot analysis) and a rapid calculation thanks to a multiple linear regression estimator. These points argue for a cost-efficient utilization of median time features for large-scale productions.

However, as with the other prototypes, it must be taken into account that inadequate cloud masks lead to artefacts and nodata gaps, and that the topographic normalization in Sen2Cor as reported in WP 32 is tending to overcorrections, which are causing artefacts in the TCD classification. In addition, resampling in Sen2Cor (20 m bands to 10 m) may partially lead to unwanted geometry effects (visible 20 m pixel borders). These issues are traced over the improved Tree Cover Density product, as it is very sensitive towards radiometrically distortions. Nevertheless, the TCD product convinces through its high level of detail and will definitively benefit from future improvements in the Sentinel-2 cloud masking algorithm and terrain correction.

3.3.2.3.4 Benchmarking

Table 3-39 gives a summary of how leaf type classification results (accuracy) relate to processing costs. Furthermore, scenario-specific chances and issues are listed. This type of benchmarking has been firstly made in phase 1 within Task 3 (compare Issue 1 of the document at hand) but had to be revised after the experiences and lessons learned from the first prototypic implementation in Task 4. Due to high cloud cover rates in several Sentinel-2 tiles, the recommended input data parameters from WP33 (Sentinel-2 only; spring period) did not deliver the desired results. This was the main reason for an extension of the observation period and the general integration of SAR data in project phase 2.

Furthermore, processing costs have been generally underestimated. Therefore, the benchmark criteria have been extended to the item storage cost. Regarding the overall processing costs, the following principles are valid:

- a) The longer the observation period the more data is needed and the higher the processing costs.
- b) The more time features are calculated, the higher the storage costs.

Consequently, the overall processing costs are dependent on the length of the time series (observation period) and the overall number of time features to be calculated. The latter significantly increases the storage costs, being part of the overall processing costs.

The input-data scenarios with the highest achieved accuracies are "S2 spring", "S2/S1 spring" and "S2/S1 spring/summer". All scenarios reach very high Kappa values greater than 0.9. While accuracies are comparable, the overall processing cost (processing + storage costs) for "S2 spring" is significantly lower than for all other scenarios, except the "S1 spring" scenario, which is not recommended for a leaf type classification. It therefore can be concluded that "S2 spring" offers a very good balance between cost and benefit. The pre-processing of 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, multi-temporal filtering, ratio calculation and time feature calculation). These empirical values may vary in dependence of the given infrastructure, but have a high influence on the budget planning in operational production projects as long as relevant input data is not provided in the desired format from the very beginning.

However, issues caused by clouds and cloud shadows in the input imagery are mitigated by the use of time features to a certain degree. This is always dependent on cloud cover situations specific to a region and the particular year, and finally also from the quality of the provided/derived cloud masks. Therefore, the addition of Sentinel-1 data to the data scenario might be required when the cloud cover/data availability situation is difficult or if the focus is set on a reliable tree cover detection.

Consequently, the addition of Sentinel-1 data to the data scenario is recommended when cloud cover is an issue in the area of interest. This solution also provides an additional benefit: The combination of Sentinel-1 and Sentinel-2 input data further increases the accuracy of the results. On the other hand, increasing costs are a direct consequence.

Table 3-39: Benchmarking criteria, chances, and issues of the different input data scenarios

Satellite / Period	Accuracy (Kappa)	Processing cost	Storage cost	Chances	Issues
S1 spring	0.41	+	+	Independent from cloud cover	SAR inherent properties (foreshortening, layover in strong relief, speckle)
S2 full + S1 spring	0.84	+++++	+++++	Partially dependent on cloud cover, but SAR and Sentinel-2 time features mitigate problematic areas	Clouds/cloud shadows, artefacts, nodata gaps, SAR inherent properties (foreshortening, layover in strong relief, speckle)
S2 full	0.85	++++	++++	Partially dependent on cloud cover, but time features mitigate problematic areas	Clouds/cloud shadows, artefacts, nodata gaps
S2 spring	0.92	+	++	Dependent on cloud cover, but time features mitigate problematic areas	Clouds/cloud shadows, artefacts, nodata gaps
S2 spring + S1 spring	0.93	++	+++	Partially dependent on cloud cover, but SAR and Sentinel-2 time features mitigate problematic areas	Clouds/cloud shadows, artefacts, nodata gaps, SAR inherent properties (foreshortening, layover in strong relief, speckle)
S2 spring/summer + S1 spring/summer	0.90 0.97	+++	++++	Partially dependent on cloud cover, but SAR and Sentinel-2 time features mitigate problematic areas	Clouds/cloud shadows, artefacts, nodata gaps, SAR inherent properties (foreshortening, layover in strong relief, speckle)

3.3.2.4 Summary and conclusions

This work investigates the potential of combining Sentinel-2 and Sentinel-1 time series data for generation of specific forest products in selected ECOLaSS test sites. Considering the limited availability of cloud-free optical satellite scenes and heterogeneous character of the analysed forest types in the areas of interest, the results are very promising for future applications on larger areas.

On basis of the tests performed in ECOLaSS, it is possible to conclude about that time features describe distinct spectral, temporal and phenological properties, mapping phenological transition points and phases while mitigating cloud cover issues and thus being well suited for status layer

classifications, monitoring and change analysis. However, time features are computationally intensive and storage intensive, relevant drawbacks to be considered when up-scaling to continental and global scales.

Time features represent the basic input data for the thematic forest classifications in ECoLaSS. Their quality heavily depends on the length of the observation period (time window selection), the time series density (scene count) and the quality of the provided (derived) cloud masks. Specific pre-processing steps such as a topographic normalization can have a negative influence on the input data, as frequently observed in image data produced by Sen2Cor (e.g. within test site Central). The applied terrain correction tends to overcorrect slopes exposed to the North and North-West, resulting in very bright surfaces, artefacts or even nodata gaps. This turned out to be problematic for the improved Tree Cover Density product, being radiometrically sensitive. However, this topic has been partially addressed by ESA in April 2019 by implementing an improved terrain correction algorithm in the Processing Baseline 02.12 for generation of Sentinel-2 Level-2A products. Further improvements can be expected with integration of a better DEM in the terrain correction. However, the above mentioned improvement could not be assessed by the ECoLaSS team.

Next to the spring period (15th March to 15th June), which has been rated as the observation period providing the best ratio of classification accuracy and lowest processing cost in project phase 1, an extension of the time window ranging from 15th March to 15th September has been successfully tested and finally applied in all FOR sites within phase 2. Increasing the time window is drastically increasing the data volume and processing time (and logically also the production costs) but has a positive effect on the achieved overall accuracy figures (retrieved by LUCAS 2018), which is in the magnitude of 2 to 4 percentage points. Cloud cover issues are generally better mitigated by considering a longer observation period. This is mainly due to the increased data availability and a potentially higher rate of data acquisitions without or with less cloud cover, additionally reducing the number and pattern of artefacts in the derived input features. From this perspective, and due to the ever-growing requirements and expectations on user side, an extended time window is recommended for operational use.

As already reported in WP 32, inadequate cloud masks represent the source for artefacts in the derived time features, which may have a negative impact in the thematic classification in case of frequent congruent artefacts, resulting in a disturbed pattern of the feature to be classified. This effect could be observed in all three sites, independently from the selected processor (Sen2Cor or MACCS). However, when looking into the achieved results, one can state that Sen2Cor Level-2A products lead to significantly less artefacts than products generated by the MACCS processor. The latter one is strongly overestimating the cloud/shadow cover, resulting in artefacts in a typical block structure, negatively influencing the “look and feel” of the products. From this point of view, the MACCS processor is less favorable than Sen2Cor. Overall, the improvement of cloud masks coming from various processors (e.g. Sen2Cor) is an asset for Copernicus Sentinel-2 applications, or even for combined Sentinel-1/-2 products and services.

In view of the achieved thematic accuracies (mainly obtained through LUCAS 2018), it has to be stated that the highest accuracy for the Tree Cover Mask is achieved by the combined use of Sentinel-1 and Sentinel-2 time features, however at highest cost. SAR features proved to be very valuable in (frequently) cloudy regions and in agricultural areas, but show some weaknesses in rugged terrain. Steep terrains lead often to commission errors of tree cover with characteristic patterns in the Tree Cover Mask.

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. In

such cases, it would not justify the enormous overhead for the pre-processing, time features calculation and data handling of additional Sentinel-1 data. However, Sentinel-1 data on its own shows only moderate predictive performance for both, TCM and DLT, and would be a viable input data complement if the area is even stronger affected by cloud cover in optical satellite imagery. Sentinel-2 time features clearly dominate the feature ranking in the DLT classifications across all test sites, whereby NIR and SWIR band features showed the highest importance for the leaf type discrimination.

In project phase 2 an improved Tree Cover Density product at 10 m spatial resolution has been generated. The TCD fully relies on optical spectral bands and is phenological and radiometrically sensitive. Any distortions in the input data (e.g. artefacts within time features) are traced over the derived product. The TCD results obtained from Sentinel-2 median time features are very promising. Median features offer the advantage to be more robust towards radiometric disturbances (e.g. artefacts, haze cover) and allow a spatially more consistent production compared to a scene-based classification approach as performed in the HRL2015 production. Thanks to the increased resolution, results provide much more detail and pattern in the forest structure as the HRL2015 predecessor. The selected time window from within the vegetation period (summer period) has been rated as suitable to cover the periods of full foliage and could be successfully transferred to all sites, covering different geographic areas and conditions. Nevertheless, an extension of the selected time window might be appropriate in case of frequent cloud cover or phenological issues.

A high quality of reference samples for training and validation is key for generation of high-quality Forest products. Indeed, samples proved to have the highest influence on the classification accuracy. The presented automated reference sampling methods (including outlier detections) based on existing HRL2015 products (e.g. Forest Sample Layer 2015, Tree Cover Density 2015) provide efficient results and contribute to a more automated workflow with shorter production times for TCM, DLT and TCD products. Iterations of the sampling process (after outlier detection) have the potential to further improve the sampling basis.

Finally, the examined and most promising methodologies could be successfully transferred to all FOR demonstration sites, especially in areas strongly affected by cloud cover, e.g. Sweden. The achieved results are largely produced by automatic routines and consistently of high quality. No manual enhancement steps have been performed, but can be easily applied on regional or local scale in order to further improve the results. Compared to previous HRL Forest production approaches (2012: single scene classification; 2015: multi-temporal classification), time features offer a more streamlined workflow as lots of manual processing steps are being dropped. They allow a consistent production over large areas and are superior to previous production approaches.

The ongoing HRL2018 Forest production could already benefit from some findings and conclusions made in ECOLaSS. This is especially the case for the continuous-scale Tree Cover Density product by utilizing median time features and the integration of Sentinel-1 SAR time features for the tree cover mapping.

3.3.3 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 implemented in ECoLaSS test sites and prototypes. 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: Definition of Grassland according to the HRL Grassland 2015). 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 developed have 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 are 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.

The grassland mapping workflow presented in Figure 3-59 shows a general overview of the workflow applied. In the first step the Sentinel-1 and Sentinel-2 time series, which are used as input data are downloaded and pre-processed. The pre-processing of S-1 time series is 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 [AD07]. Additionally, temporal statistics have been calculated which are used as input data for the time series classification processing chains. The temporal features generated are the minimum, maximum, mean, standard deviation, coefficient of variation and percentiles. The Sentinel-2 Level-1C products are automatically downloaded from the CopHub and processed including following steps: atmospheric correction, topographic normalization and cloud masking. In turn, the Sentinel-2 sensor system has an overall number of 12 bands from 10 m to 60 m spatial resolution, thus only the 10 m and 20 m bands are used. Further vegetation indices are derived from the Sentinel-2 data sets for each image and are used for the generation of multitemporal features. Based on the reflectance bands and the vegetation indices, annual features like median, mean, maximum, minimum and standard derivation and percentiles are derived and used as input for the classification method. Several classification algorithms can be applied. Since the Random Forest algorithm showed good results in phase 1 in the WEST demonstration site it was applied in phase 2 in other sites. LUCAS 2018 has been used as reference data for training. The output of the machine learning classification approach are the thematic grassland layer and probability for each pixel for belonging to a certain class.

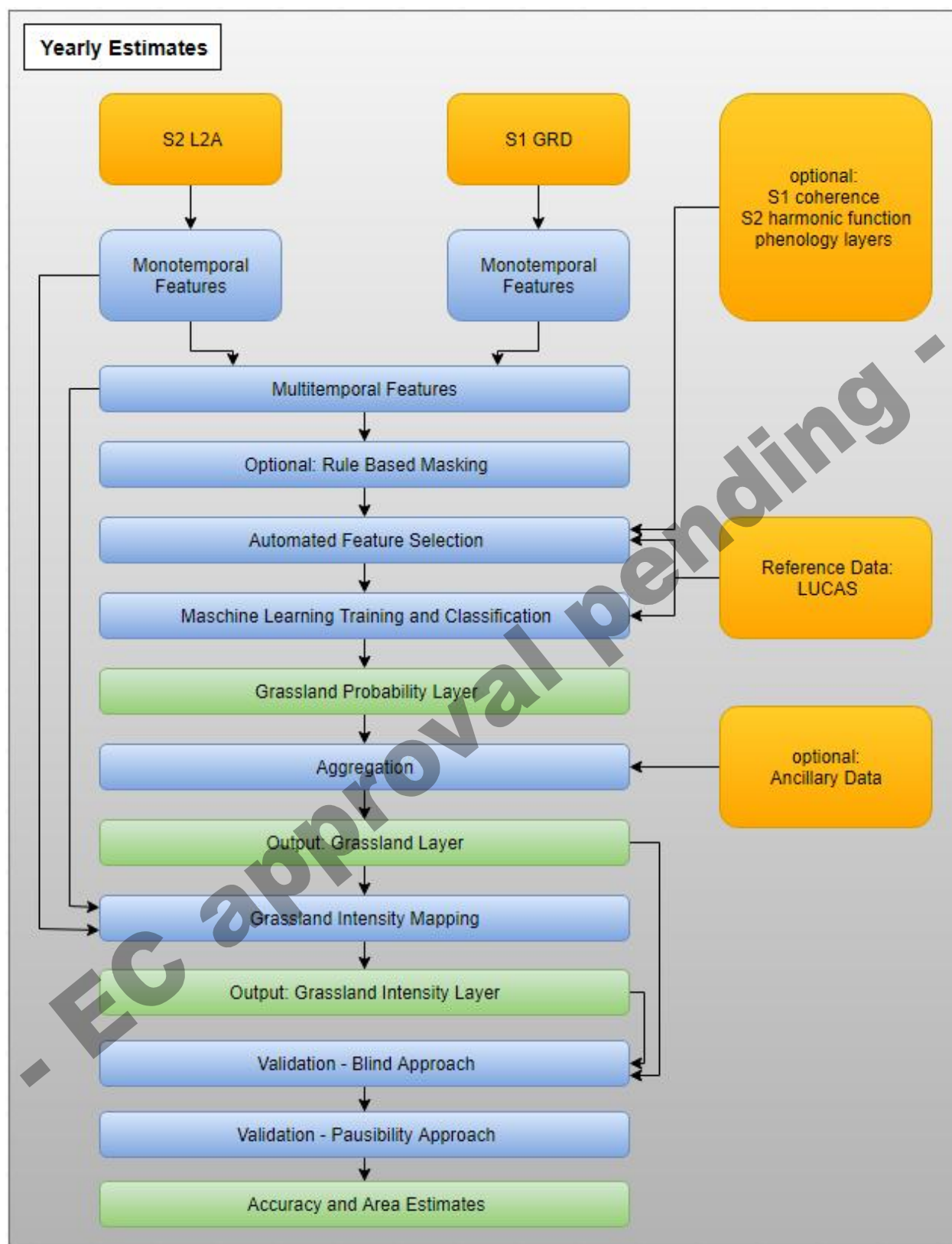


Figure 3-59: Grassland mapping workflow overview. Sentinel-2 (S2); Sentinel-1 (S1); Land Use and Land Cover Survey (LUCAS).

The aggregation is done by taking into account the probabilities, the MMU. Further, the use intensity is derived only within the grassland mask. The final products are statistically assessed. A blind

approach is applied to start with and then a plausibility analysis is implemented for all layers considering the information of the map layer.

3.3.3.1 Description of candidate methods

CLASSIFICATION METHODS

THRESHOLD SCHEMES

In order to classify grasslands out of S1 time features a set of training areas were ascertained and from these samples, the signatures were extracted. The spectral signatures from S1 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 WEST demonstration 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. This classifier has been selected also for the other thematic topics in ECOLaSS due to its satisfactory performance in phase 1 implementations.

West test site

In phase 1 for training and validation, 3408 LUCAS points covering the Belgium site are visually interpreted based on the Sentinel-2 time series data from 2016 until 2017. 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. In phase the LUCAS 2018 dataset is used, where some points are filtered based on database queries. 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.

Central test site

Likewise, in the Central test site (tiles 32TNT and 32UNU), respectively 744 and 685 samples were derived from the HRLs 2015 reference maps and LUCAS 2018, filtering by the Observation Type attribute to be sure the point had been interpreted from a reasonable distance limit. As explained in section 2.1 referring to automatic sampling, outlier detection is a key step in this process. Accordingly, an outlier detection based on spectral signals is carried out by means of applying zonal statistics for 30x30 m samples. In addition, some manual sampling has been applied in the test sites. In this case, the interpreted points were randomly split into training and validation data sets at a ratio of 50% training to 50% validation. 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 250 and from the training models the grouped forward feature selection was used.

South-East test site

In the South-East site, the LUCAS 2018 points were used as trainings samples. In total 3871 LUCAS samples were available in the demonstration site, of which 743 belonged to the grassland classes. LUCAS data were filtered by identical criteria to the demonstration site West lined out in Table 3-41. After filtering, 2168 LUCAS samples remained (482 grassland samples), of which 25% were set-aside for internal validation. The LUCAS land-cover classes were then converted into binary form, that is, “grassland” and “non-grassland”.

GRASSLAND INTENSITY MAPPING METHODS

TRACKING WITH KALMAN FILTER:

Previous work on the mapping of grassland use intensity has shown that mowing events are related to abrupt drops of the NDVI level. Therefore, the initial approach was characterized by the attempt to track the NDVI level of a given pixel through time with a Kalman filter and to label abrupt and statistically significant drops as mowing event. The filter should be able to follow gradual NDVI changes, also in the presence of larger gaps in the time series, without signaling a mowing event. While early tests indicated the validity of the approach in good conditions, they also showed problems caused by sparse observation availability, the inevitable presence of un-masked clouds and haze, or increased soil moisture, which can affect the NDVI level in a similar way as a mowing event does. It is expected that the detection of mowing events based on a single spectral index is not reliable enough. An approach based on multiple observables is deemed favorable, because it offers the possibility to search for specific multivariate change vectors associated to mowing events.

The harmonic regression tests with time series of the Tasseled Cap components Brightness, Greenness, and Wetness, which have been conducted in an earlier phase of the project, provided the basis for further development. The Greenness signal is highly correlated with the NDVI, therefore mowing events also correspond to abrupt drops of the respective signal level. In addition, the Wetness signal is systematically affected by mowing events in a similar way, because the removal of healthy biomass after a mowing event causes an increase of the soil reflectance. Hence, the idea is to apply the signal-tracking approach to the Tasseled Cap components and thus take information from six input bands (blue, green, red, NIR, SWIR1, SWIR2) into account. The implemented algorithm signals a mowing event if a statistically significant change vector in the two-dimensional feature space created by Greenness and Wetness is detected and its direction corresponds to a drop of both variables. The statistical significance is influenced not only by the change vectors' magnitude, but also by the length of the time gap between consecutive observations. Large gaps in the time series will result in a lower sensitivity of the detection method,

because the algorithm has not enough information to distinguish between abrupt and gradual signal changes.

The following figures illustrate time series corresponding to single grassland pixels with varying number of mowing events. For each case, the observed six input bands are plotted together with the Kalman filter predictions. A second plot shows the estimated Tasseled Cap components and the associated 95% confidence intervals. Mowing events signalled by the algorithm are marked as well and the corresponding changes of the signals can be observed.

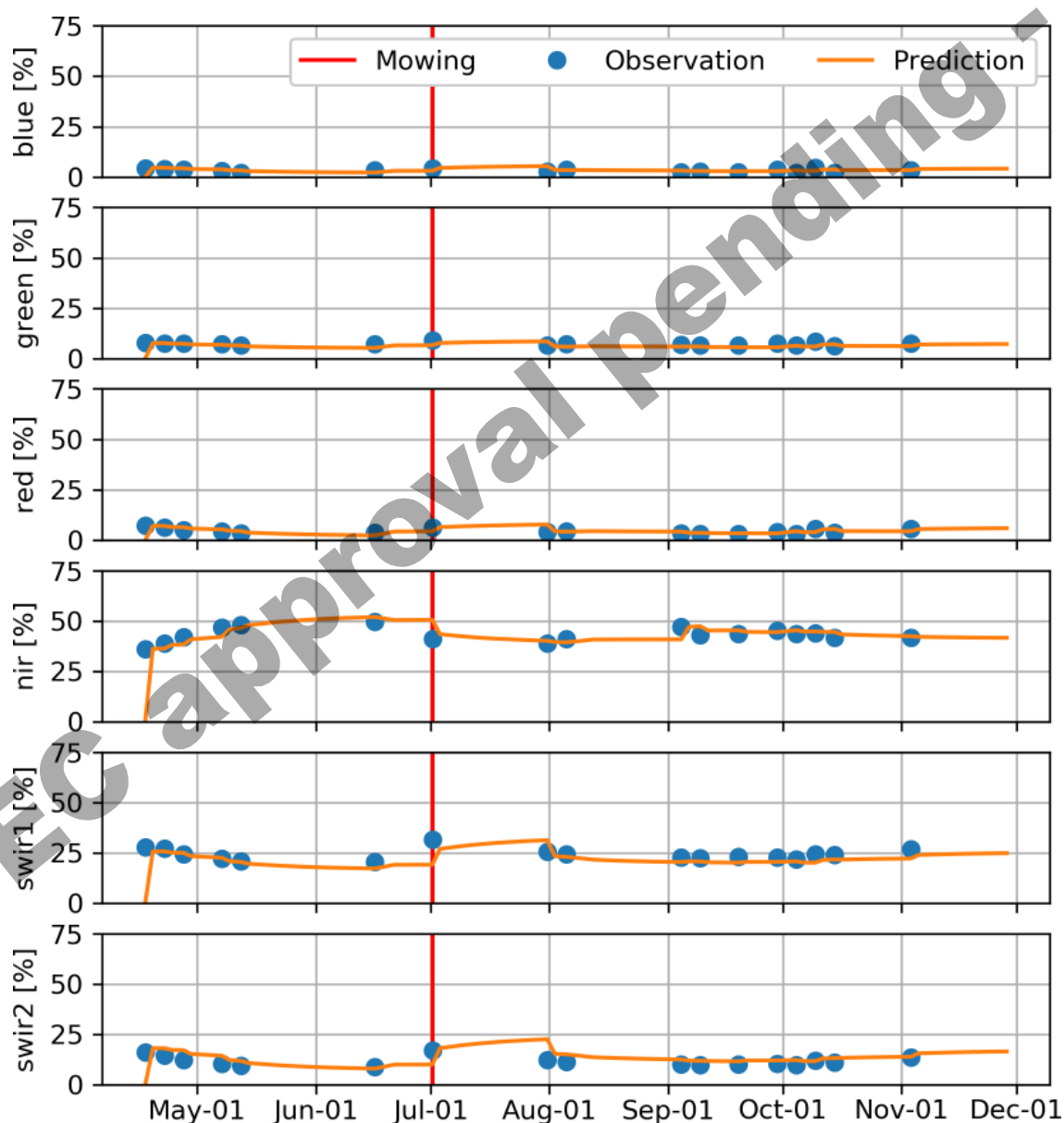


Figure 3-60: Multispectral time series of a single grassland pixel, one mowing event according to INVEKOS.

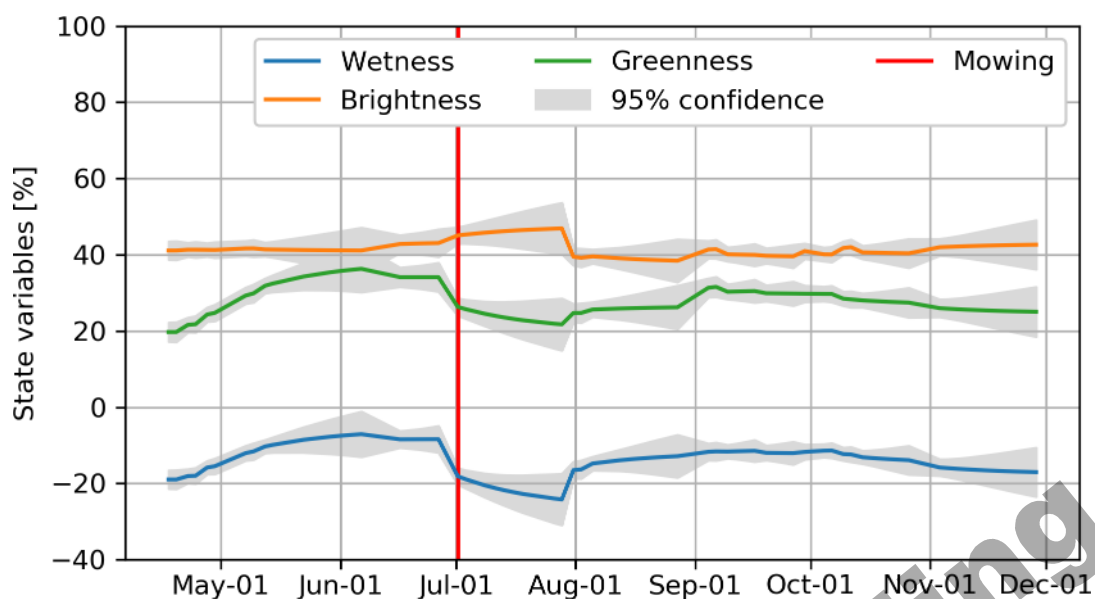


Figure 3-61: Variables estimated by the Kalman filter from the observations plotted in Figure 3-60.

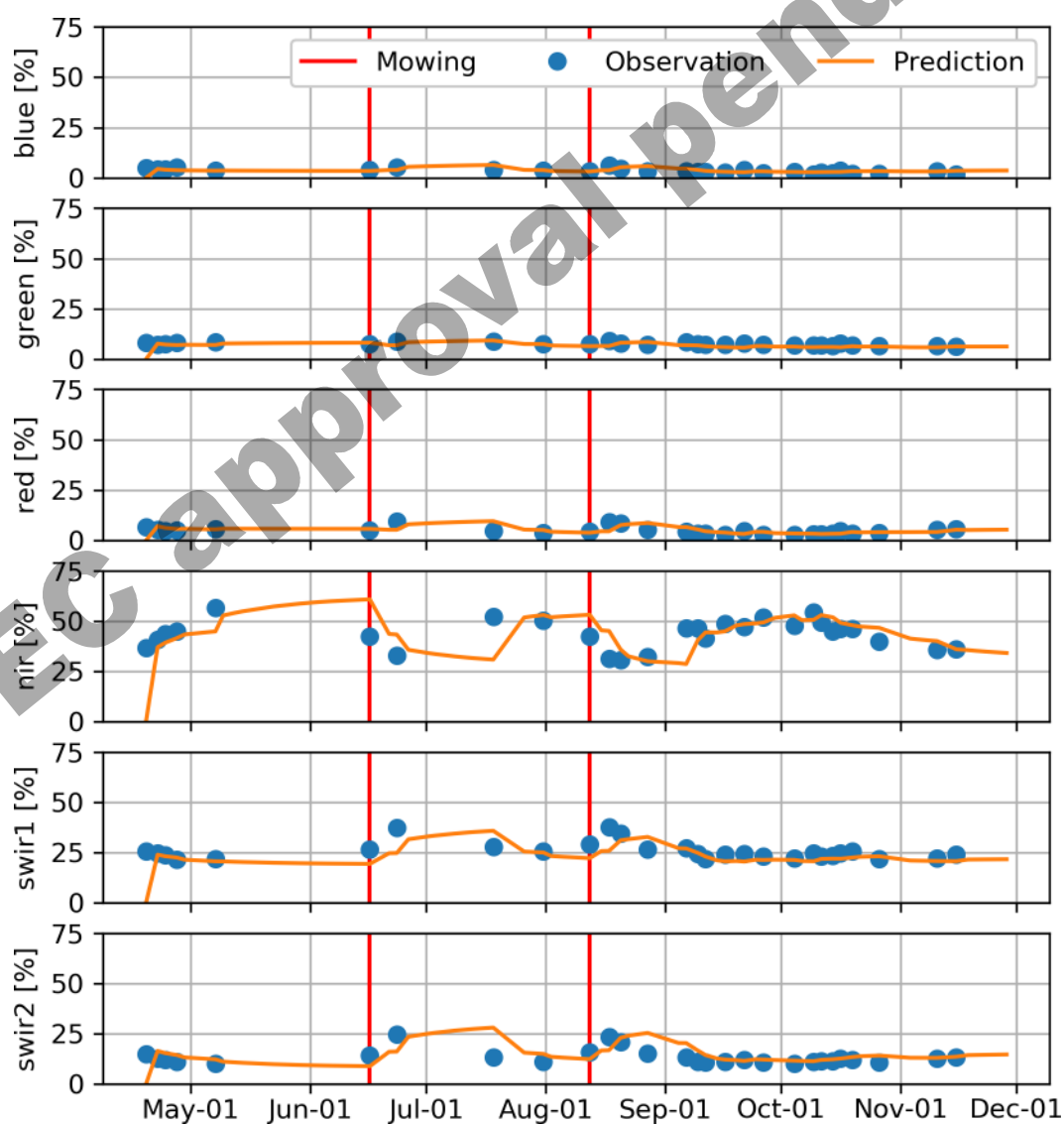


Figure 3-62: Multispectral time series of a single grassland pixel, two mowing events according to INVEKOS.

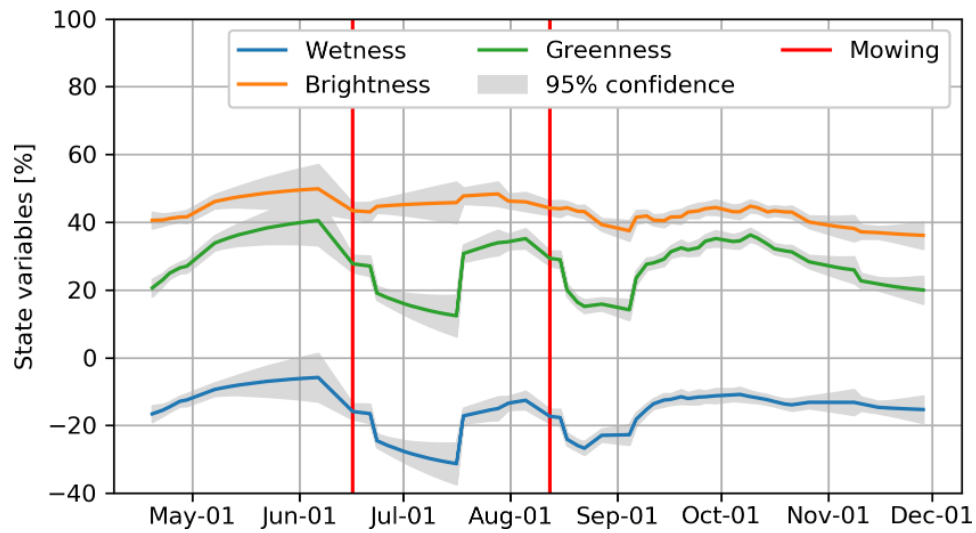


Figure 3-63: Variables estimated by the Kalman filter from the observations plotted in Figure 3-62.

- EC approval pending -

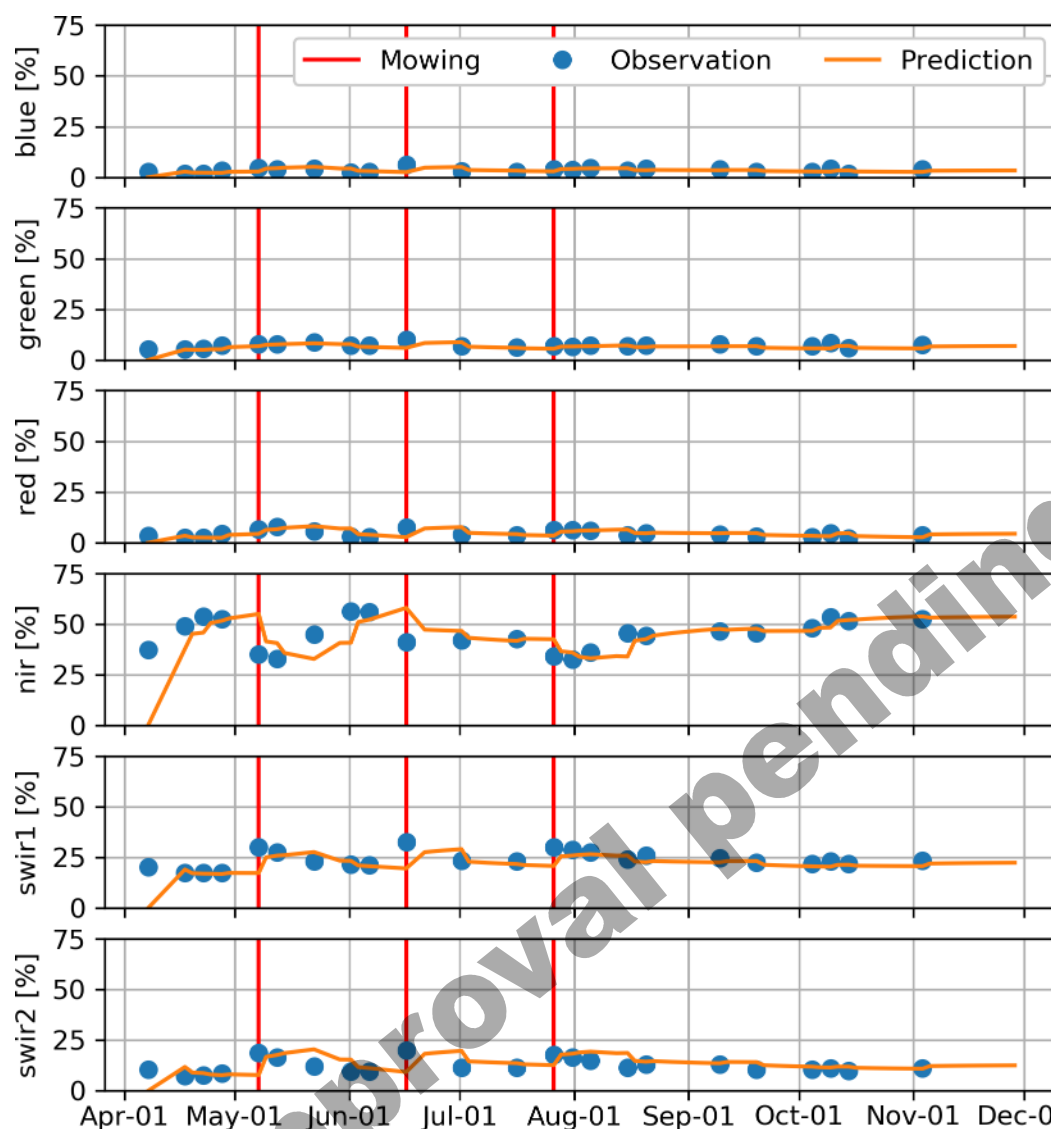


Figure 3-64: Multispectral time series of a single grassland pixel, three mowing events according to INVEKOS.

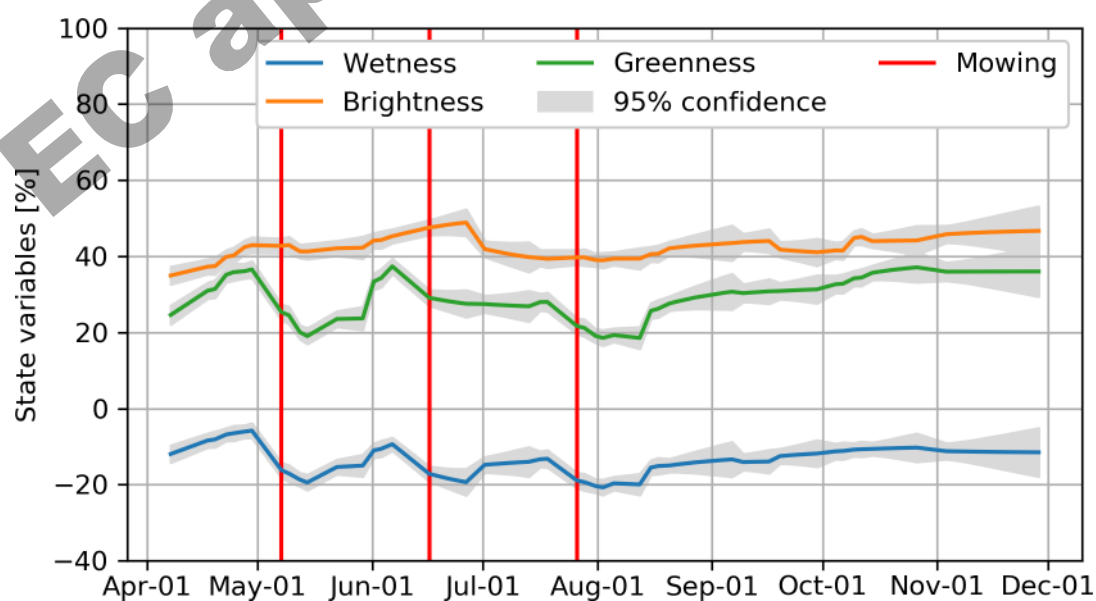


Figure 3-65: Variables estimated by the Kalman filter from the observations plotted in Figure 3-64.

Within Task 3, tests in the Central site on a new product were carried out which provided more details for the detected grassland areas, regarding the mowing intensity. The resulting mowing intensity layer at 10m spatial resolution is based on the number of mowing events detected. For this purpose, NDVI time series are used to derive a layer showing the number of mowing events. This layer is then clipped to the grassland mask and re-classified by defining the extensive use category when less than or equal to two mowing events are detected during the year.

The intermediate product generated is the number of mowing events. At first, the coherence features from SAR data were tested, although according to the accuracy versus performance benchmarking (e.g., computation costs, product quality and timeliness, etc.) and taking into account the upscaling of the products to larger scales in a cost-efficient manner, another approach was selected instead for implementation in the demonstration site. Coherences are highly sensible to changes, even on micro-level, and therefore, events like heavy rainfall are likely to make coherence images unusable for intensity analysis. This is highly risky, besides the expense of the processing of SAR coherences, when considering automation and large scale products. Consequently, in Central the approach based on NDVI time series has been applied.

NDVIs were computed for all scenes available in 2018 to detect mowing events all throughout the year by subtracting consecutive NDVI acquisitions and a rule based classification, defining that all pixels with ≥ 3 mowing events are intensively used and 0-2 mowing events means extensively used, in terms of mowing events. The following Figure sketches the workflow applied in the mowing intensity layer generation:

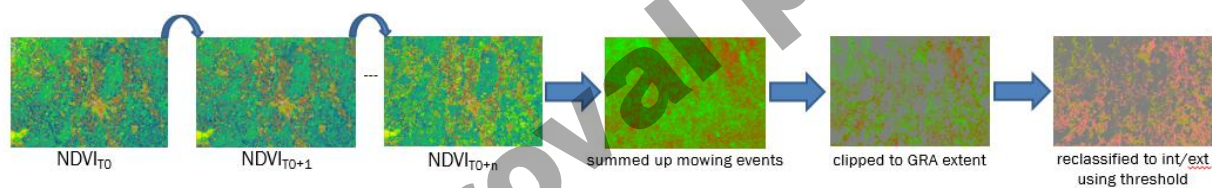


Figure 3-66: Intensive/Extensive grassland use layer workflow.

As for the other grassland layers, filtering improves significantly the look and feel by reducing noise. For the mowing intensity layer, a filter of 4 pixels in size was applied. All areas within the grassland mask were filtered, so that there is no patch for one of the two intensity classes smaller than 5 pixels in the end. Within small grassland patches, it might happen that e.g. 3 pixels are classified as extensive and 2 are classified as intensive. In such cases, the filter would cause the class values to jump between classes with each filter iteration without getting a patch of 5 unique values. If so, it was filtered in favour of intensive use because most of the areas are used intensively in the region. This layer is also useful to check for natural grasslands if it is assumed that the latter are present when no mowing events are detected at all. For this assumption to be more reliable, a longer time series (e.g., several years) should be considered.

The Figure below shows the test in Central for the grasslands mowing intensity in 2018.

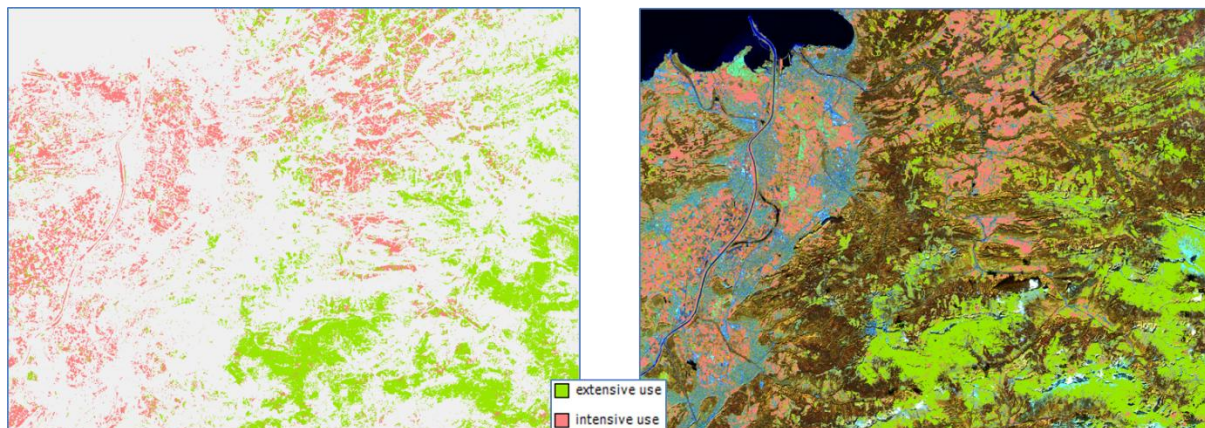


Figure 3-67: Grassland mowing intensity in Central 2018.

In Figure above, and in Figure 3-68 it can be observed that grasslands are extensively managed in Alpine regions whereas more intensively in valleys around settlements.

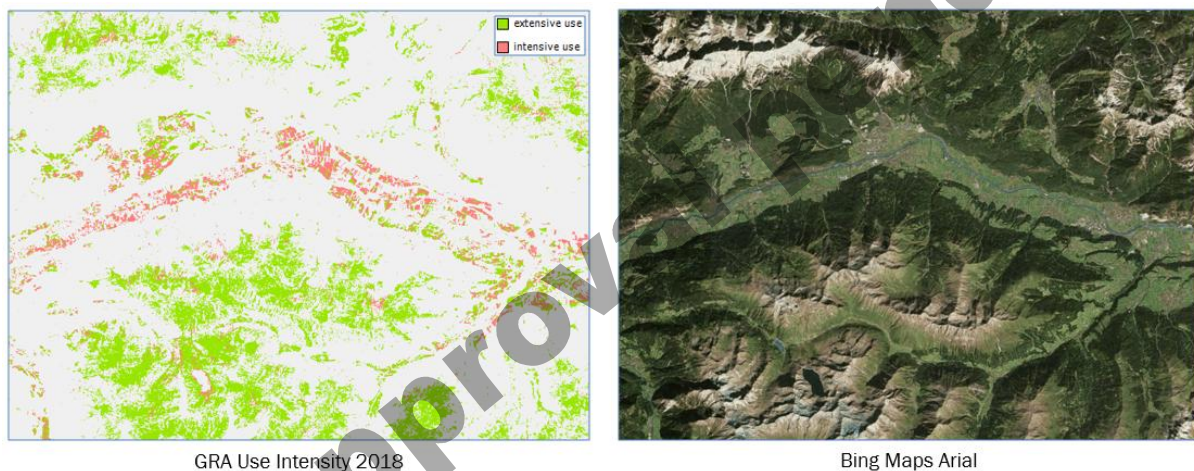


Figure 3-68: Detailed view of the grassland mowing intensity layer compared to Bing Maps Aerial of the same region, showing more intense management is concentrated in valleys around human settlements.

3.3.3.2 Benchmarking criteria

In the following chapter, the benchmarking criteria applied for the grassland status mapping are described, including visual inspections with reference data, thematic accuracy assessments, separability analysis and cost efficiency evaluations.

FEATURE IMPORTANCE/SELECTION:

Different Features selection methods are applied and tested. Seasonal and annual statistical features are investigated regarding the grassland status and intensity mapping to achieve the optimal set of features/indices per biographic region/elevation stratum. 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). Further, the grouped forward feature selection method based on decision trees is also tested.

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).

GROUPED FORWARD FEATURE SELECTION

The generation of suitable time series time features, especially consider upscaling and in particular the pan-European or global roll-outs, is challenging and require large computation capacities. Indeed, many reasonable combinations of time series metrics, sensor bands, indices and suitable temporal windows are conceivable, leading to a potentially quite large number of potential features. The testing experiences in Task 3 for grasslands in Central contributed to the definition of the classification parameters and proved the temporal windows and features that were performing best. In this regard, the Random Forest classification algorithm provides information about feature importance.

In parallel to the new feature calculation and analysis, the grouped forward feature selection method was applied in the tests in Central. The grouped forward feature selection method adapted and embedded in the Random Forest classification process is based on the sequential feature selector integrated in the machine learning package (python module scikit-learn in the machine learning extension MLxtend).

This sampling method is used to reduce an initial d -dimensional feature space to a k -dimensional feature subspace where $k < d$. The goal of feature selection is two-fold: improve the computational efficiency and reduce the generalization error of the model by removing irrelevant features or noise. This wrapper removes or adds one feature at the time based on the classifier performance until a feature subset of the desired size k is reached. The difference with other methods like the Recursive Feature Elimination, is that the latter is computationally less complex using the feature weight coefficients (e.g., linear models) or feature importance (tree-based algorithms) to eliminate features recursively whereas the forward feature selection eliminates or adds features based on a user-defined classifier/regression performance metric. The algorithm finally yields a combination of the features with the highest accuracy. This subset of features is used for the classification process.

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 is assessed with an error matrix and following accuracy metrics: Overall Accuracy and Error, User's accuracy, Producer's accuracy, Kappa Coefficient and Confidence Intervals, following the principles from section 2.4.

3.3.3.3 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 applied in the West test site in Belgium and Random Forest classifier applied in all test and demo sites.

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.

3.3.3.3.1 Demonstration site WEST

In the demonstration site, WEST testing and benchmarking has been performed in both phases. In phase, 1 threshold schemes and the Random forest classifier are tested. Since the Random forest provided promising results, it has been further tested and applied in phase 2. Additionally, S1 coherence and S2 harmonic region parameter features are tested regarding the grassland non-grassland discrimination.

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). In this sense, in the development of the ECoLaSS prototypes it was decided that for an automated approach, and homogeneity purposes for a potential larger scale roll out at a Pan-European level, the grasslands nomenclature should remain general, without differentiation between the categories below.

- 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

In further developments, as is the case in the forest with the dominant leaf type and tree cover density, and agriculture products with the crop types maps described in this document, the grassland binary mask can be enriched with more detailed classes like the ones listed above. In ECoLaSS the use intensity layer is designed as such: a higher detail layer on the grassland areas defined in the grassland/non-grassland mask.

A further reference dataset has been created by Joanneum Research through visual interpretation. The dataset is based on the LUCAS 2012 points located on the demonstration site West. The reference for the interpretation is the Sentinel-2 and Landsat-8 time series 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) and VHR data ordered from the DWH have been used.

It is necessary to compare both reference datasets to assess their quality. Therefore, only VIRP points located within the LGP polygons are compared with each other as shown in Table 3-40: Reference data comparison LPG2016 vs VIRP2016. Reference data comparison (LGP2016 vs VIRP2016).

Table 3-40: Reference data comparison LPG2016 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.

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 (900m²) 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.

Additionally, **LUCAS 2018** data is used within the Grassland benchmarking and prototype generation. It is recommended to use only homogenous LUCAS points where the distance from the interpreter to the point location is less than 100m. Following attributes from the LUCAS 2018 table can be used for the selection of homogenous points.

Table 3-41: LUCAS 2018 inclusion rules.

Inclusion Rules	Comments
"OBS_TYPE"= 1 or "OBS_TYPE"= 3	1: In situ < 100 meter 3: In situ PI

Inclusion Rules	Comments
"CPRN_LC1N" >= 20 and "CPRNC_LC1E" >= 20 and "CPRNC_LC1S" >= 20 and "CPRNC_LC1W" >= 20	If the point is a Copernicus point it will be excluded if the homogenous LC is smaller than the circle with 20m radius.
"PARCEL_AREA_HA" >= 2	2: 0.1ha <= area < 1ha
"LC1_PERC" = 100	Land cover percentage

S1 COHERENCE:

The InSar coherence products are derived from Sentinel-1 SLC data [AD07]. Regarding the processing, 6-day and 12-day and 18-day coherence products were generated based on Sentinel-1A and Sentinel-1B SLC imagery for different time periods in 2018, e.g. March/April, April/May, April/June and for the entire period from February to November 2018. The coherence estimation has also been performed for different output resolutions, i.e. 20m and 40m. Based on the outcomes of WP 32 6-day coherence in 20m resolution images from different time intervals seemed the most promising features and are tested and analysed regarding their potential for grassland and grassland mowing intensity mapping.

6-day Coherence: March-October, RGB: Cov, Mean, Min

WV02: 18.06.2017, RGB: NIR, Red, Green



Figure 3-69: InSar Coherence 6-days (March-October).

Figure 3-69 shows an RGB image of the 6-day coherence product (2017 March - October) including the Coefficient of Variation, Mean and Minimum statistics. In comparison to the World View 2 reference image from 2017, the delineation of agricultural fields are blurred compared to the WV2 VHR image. The exact field structures are not displayed in a detailed manner, i.e. border lines or ratio of field sizes. Further urban areas are largely overrepresented. The reason for this is that the coherence is calculated from SLC (Single Look Complex; <https://sentinel.esa.int/web/sentinel/user-guides/sentinel-1-sar/resolutions/level-1-single-look-complex>) data with 40m spatial resolution. This spatial resolution does not support better resolution image quality, although the data have been resampled to 20m for better calculation. Consequently, a pixel-based approach cannot be recommended for data types with such high difference of spatial resolutions. This deviation in resolutions can lead to inaccurate results due to geometric problems if a pixel wise classification approach is performed.

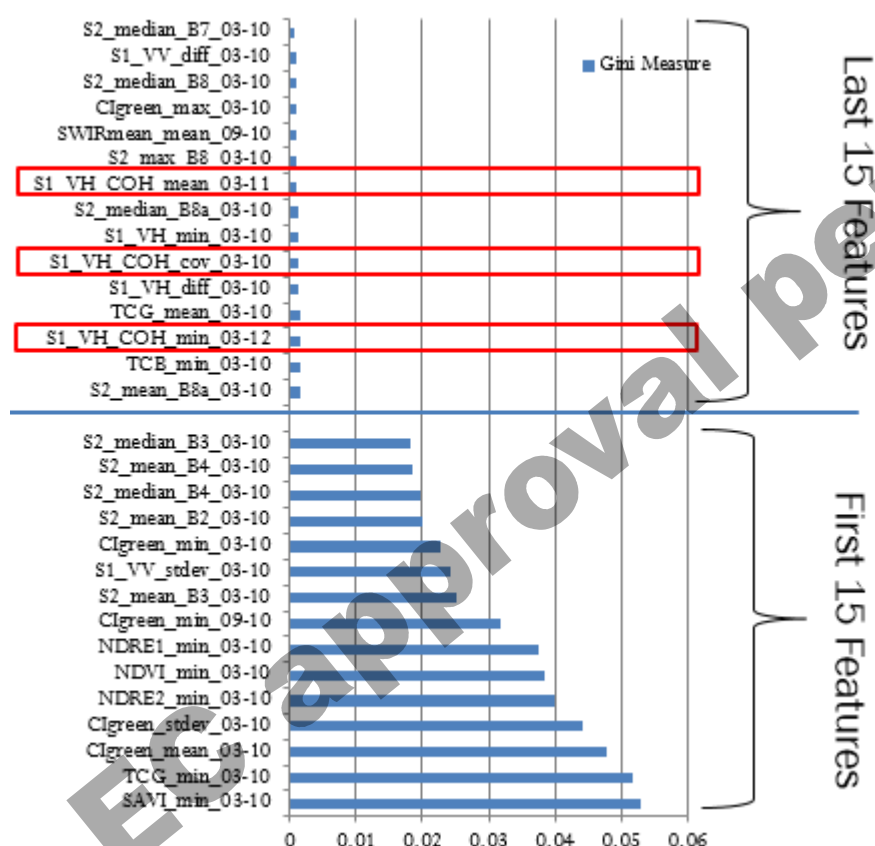


Figure 3-70: Feature importance S1 coherence, S1 backscatter time features and S2 time features (reflectance + indices)

Further tests have been conducted using the 6 day coherence products in the Random Forest feature selection approach. Figure 3-70 shows the results of the feature selection using S1 and S2 time series features in combination with the coherence features. The LUCAS 2018 points which serve as training dataset for the Random Forest classification are used to derive the feature importance. The results show that the coherence features are not considered as important by the analysis method. The aforementioned geometric issues have a negative impact on the value of the coherence product.

S2 HARMONIC REGRESSION PARAMETERS

Dense Sentinel-1 SAR and Sentinel-2 optical time series data are used to derive temporal parameters with the aim to distinguish between land cover type or even intensively and extensively managed grasslands. A signature analysis showed the potential for this separation, especially during the early stages of the phenological cycle in spring. Several differences between the fitted curves can be observed, for example the value of the trend parameter c_0 , the composition of the seasonal pattern with respect to the amplitudes of the different frequencies, and the minimal value and range (difference max, min) of the fitted curve.

The regression model, which has been fitted to the time series of each pixel of the test site, is given below. It features a constant trend as well as a seasonal component of 3 frequencies, thus there are 7 parameters to estimate. To introduce a certain degree of over-determination, the minimum number of available unmasked observations required to carry out the IRLS procedure is set to 14.

$$z(t) = c_0 + \sum_{j=1}^3 \alpha_j \cos(\omega_j t) + \beta_j \sin(\omega_j t)$$

The first step of the implemented test set-up is to eliminate remaining gross outliers in the data set caused by unmasked clouds, snow, and cloud shadows. It is assumed that these outliers are visible more prominently in the time series of Tasseled Cap Brightness, subsequently referred to simply as Brightness. The coefficients for the Tasseled Cap transformation were taken from (Crist, 1985). Clouds and snow are expected to cause unusually high Brightness values, whereas cloud shadows should correspond to low magnitudes of Brightness. An example of a Brightness time series for a single grassland pixel is given in Figure 3-71. Additional to the observations, the results of an OLS fit on the one hand and an IRLS fit on the other are plotted. The OLS solution is influenced by the outlying values in the series, whereas the robust IRLS fit is not. The weights assigned to each observation by the IRLS procedure are plotted in Figure 3-71 and it can be seen that 4 weights are zero. Using Figure 3-71, it can be verified that weights of zero in Figure 3-72 correspond to anomalous observations. Therefore, all observations with an assigned weight of zero are excluded from further processing steps.

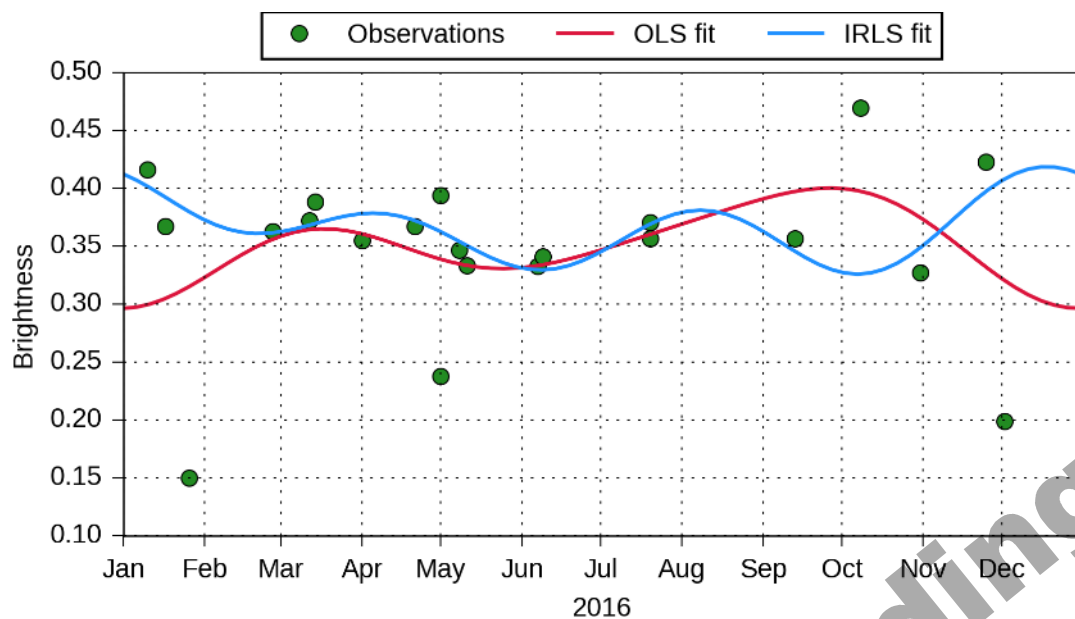


Figure 3-71: Example of a Tasseled Cap Brightness time series of a grassland pixel and fitted regression models using OLS and IRLS.

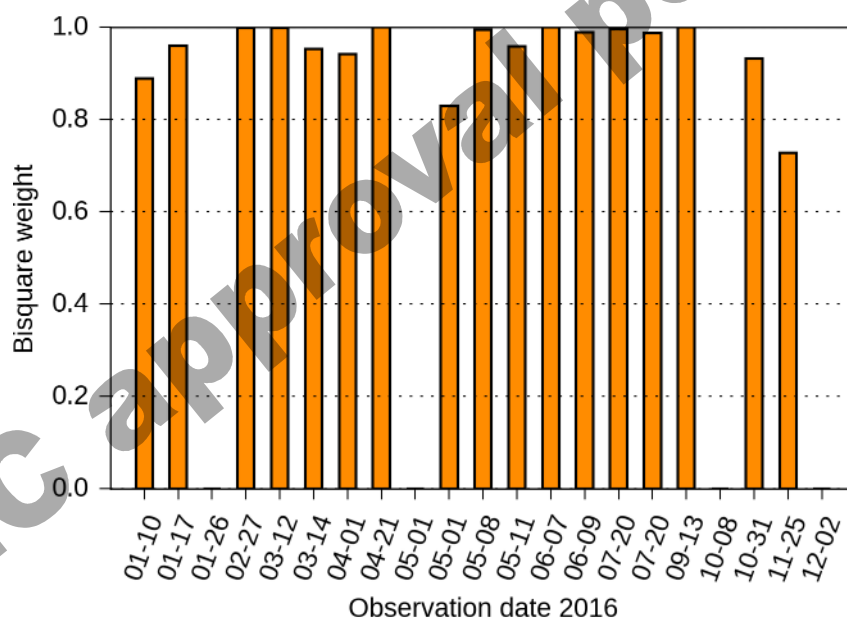


Figure 3-72: Computed observation weights of the IRLS fit.

The second step of the implemented test set-up is to fit the regression model to the time series of Tasseled Cap Greenness, subsequently referred to simply as Greenness, which is assumed to be an appropriate spectral index to capture vegetation dynamics. Since most outliers should have been eliminated in the previous step, OLS is used to estimate the parameters. Taking the same grassland pixel as discussed above, the associated Greenness time series is plotted in Figure 3-73.

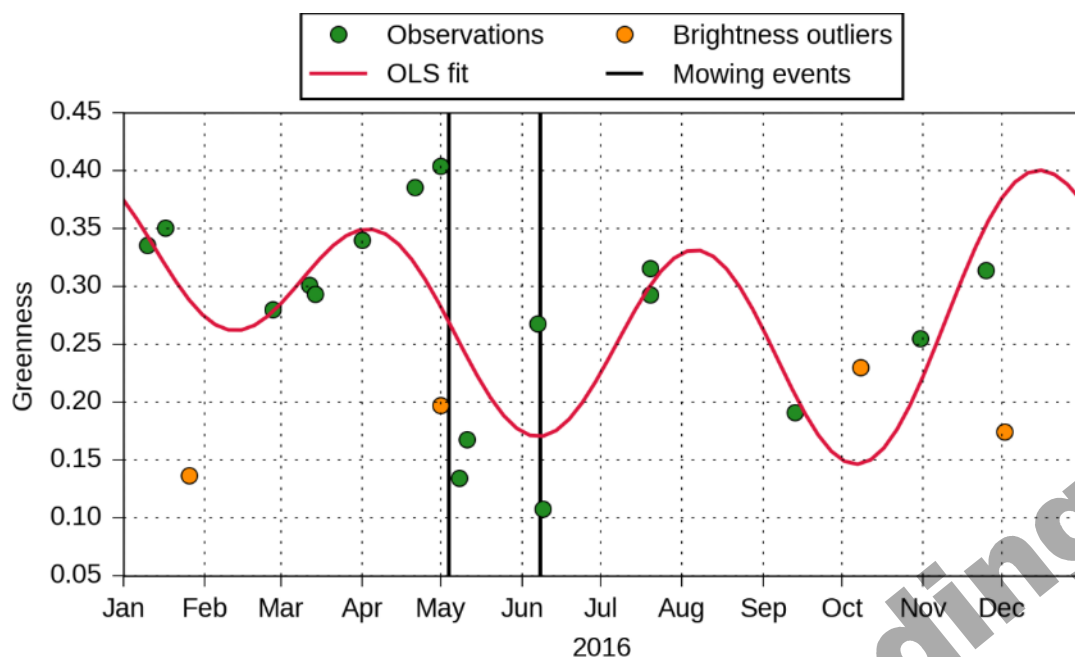


Figure 3-73: Example of Greenness time series of a grassland pixel and fitted regression models using OLS.

Inspecting Figure 3-73 several observations can be made:

- Two mowing events designated by abrupt jumps in the Greenness level can be clearly identified, with a possible third one indicated somewhere at the end of August.
- The time series model cannot capture the high temporal dynamics of the mowing events since it does neither account for abrupt jumps nor short growing periods of only one month.
- The Greenness level of the outliers identified as cloud shadows is similar to the level after a mowing event, thus a confusion of the two conditions is possible. This emphasizes the necessity of outlier detection.

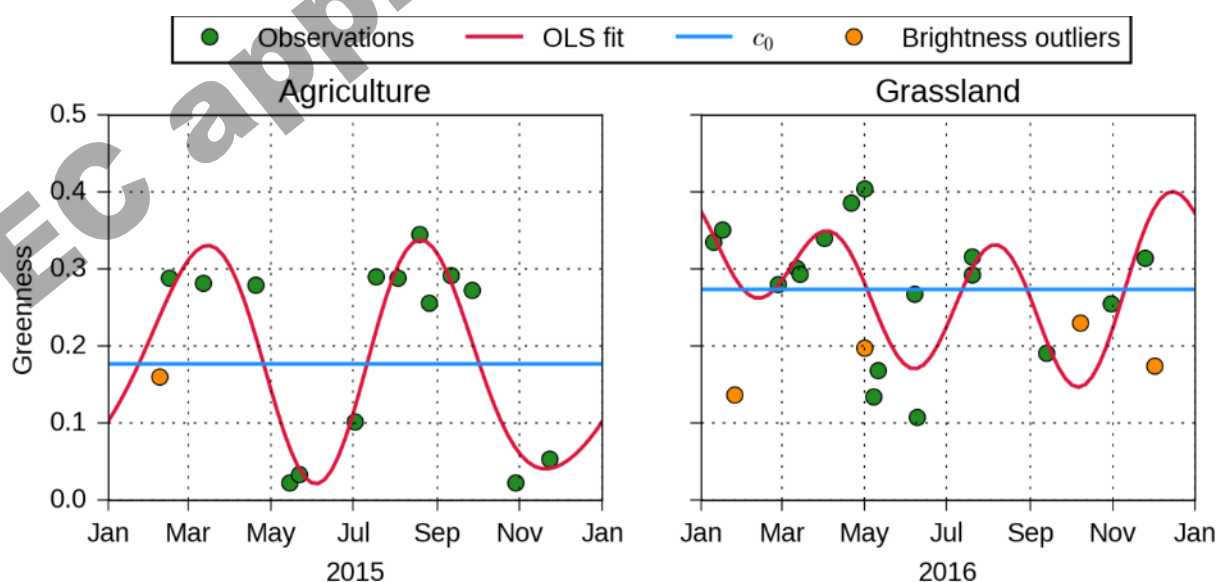


Figure 3-74: Comparison of an agricultural pixel to a grassland pixel.

Although the time series model cannot capture abrupt jumps, the set of estimated parameters may still be interpreted as a compressed form of the multi-temporal information contained in the image stack. While the parameters do not reflect all the variation in a time series, they might hold enough information to separate different land cover/land use classes. The idea is illustrated in Figure 3-74, where the Greenness time series of an agriculture pixel is contrasted with the time series of the previously used grassland pixel. The same workflow has been applied to obtain the OLS fit. Several differences between the fitted curves can be observed, for example:

- the value of the trend parameter c_0 ,
- the composition of the seasonal pattern with respect to the amplitudes of the different frequencies, and
- the minimal value and range (difference max, min) of the fitted curve.

The parameters α_j and β_j can be converted to amplitude values A_j using

$$A_j = \sqrt{\alpha_j^2 + \beta_j^2}$$

Additional to the parameter values, the OLS method also yields the covariance matrix of the estimates, which can be further used to derive the uncertainty of the amplitude values. Figure 3-75 contrasts the trend and amplitude parameters of the agriculture pixel with the corresponding values of the grassland pixel. The 90% confidence interval of each estimate is also illustrated, suggesting that there is a statistically significant difference in the c_0 and A_2 parameters. A possible approach to detect anomalies and indicate change is to carry out hypotheses tests to determine if one or more parameters have significantly changed from one year to another.

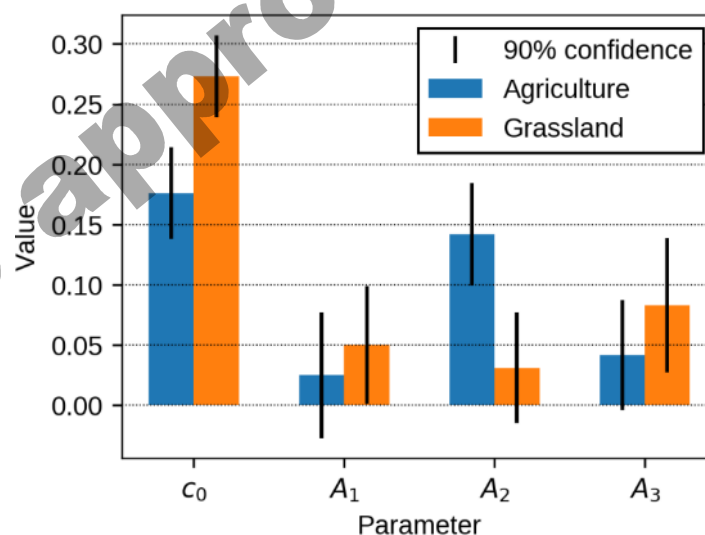


Figure 3-75: Greenness trend and amplitudes of an agriculture and grassland pixel.

A different example illustrates the behaviour of a grassland pixel in consecutive years. Figure 3-74 shows the time series of Greenness and Figure 3-75 the corresponding values of the trend parameter as well as the amplitudes of the seasonal components. While the fitted curves look very different, the overall trend and the amplitudes stay roughly at the same level when their confidence

intervals are taken into account. The change in the appearance of the fitted curves can be explained by phase shifts in the seasonal components.

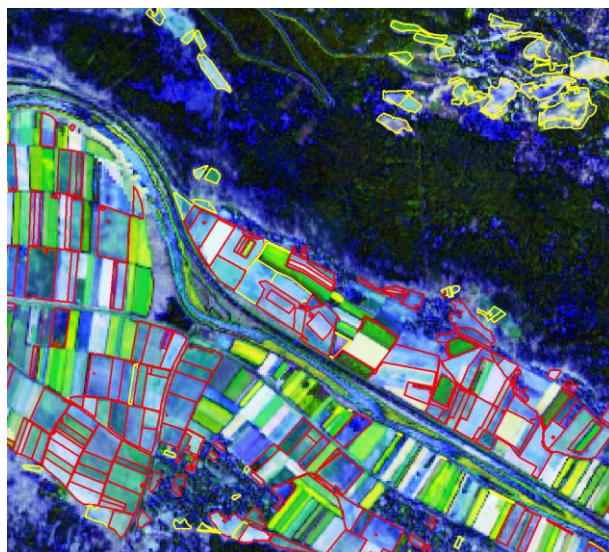


Figure 3-76. Regression function second order (R: A1 G: A2 B: P1).

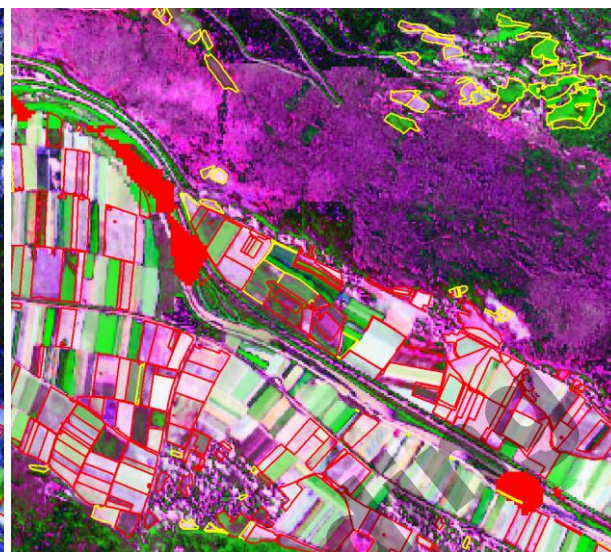


Figure 3-77: Regression function third order (R: A1 G: A2 B: A3).

Although this represents a promising approach also for discrimination of grassland management regimes it is necessary to have dense time series. Large gaps in the time series model as well if the first observation is very late and the last observation available is too early, result in unusable artefacts. Further using the 3 order frequency is better suited for grassland management regime but it needs at least 15 observations to generate valid results. Due to the high number of minimal observation needed the results is comprised with large Nodata areas. Therefore, the features are not considered useful in the random forest feature importance and are further not applicable in an operation large-scale approach.

THRESHOLD SCHEMES APPLIED ON S1 DATA

The SAR2017 image stack is derived over the year until 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.

The error matrices for the 2017 SAR classification for both reference datasets (VIRP-2017 and LPG2017) are depicted in Table 3-42 and Table 3-43.

Table 3-42: Confusion matrix using VIRP-LUCAS points and the threshold based S1 classification for 2017.

		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

Table 3-43: Confusion matrix using LPG2016 points and the threshold based S1 classification for 2017.

		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 3-78: SAR grassland threshold-based classification for 2017 (grassland in yellow).

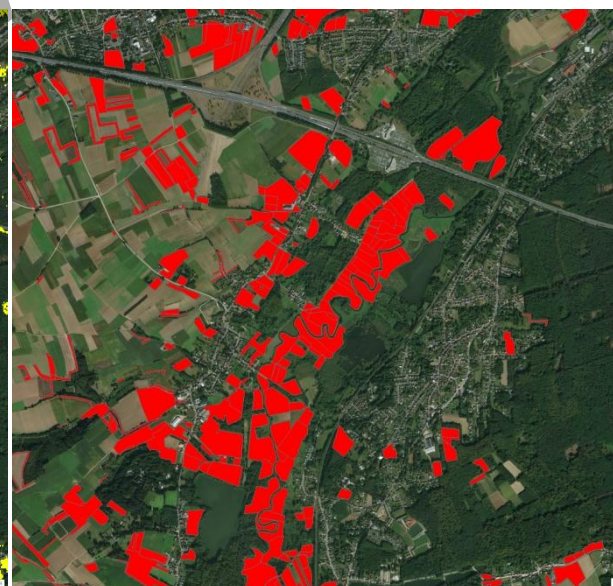


Figure 3-79: LGP grassland areas in red. Basis layer: ArcGIS Basemap.

Figure 3-78 and Figure 3-79 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.

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 points 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 3-44).

Table 3-44: 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). Since the threshold based classification results in lower accuracies in comparison to machine learning algorithms, analyzed in phase 1, this approach is not further applied and tested in phase 2.

RANDOM FOREST BASED GRASSLAND CLASSIFICATION WITH SAR DATA

The Random forest algorithm derived good results in phase 1, therefore further tests were applied in phase 2. First, the use of annual and seasonal statistical time series features are tested using only S1 data. In total 40 annual and seasonal S1 features are calculated as described in chapter 3.1. The feature importance for the 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. Providing all in chapter 3.1 mentioned features the feature selection dominantly includes seasonal percentile 90, median and the standard deviation statistics. It is recommended to exclude the seasonal percentile, CoV and standard deviation features, due to instability in short periods.

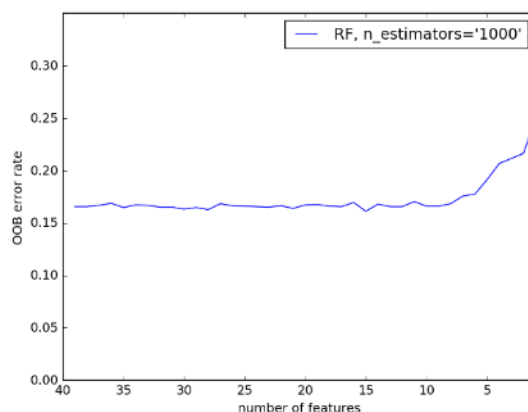


Figure 3-80: OOB error in relation the number of S1 input features.

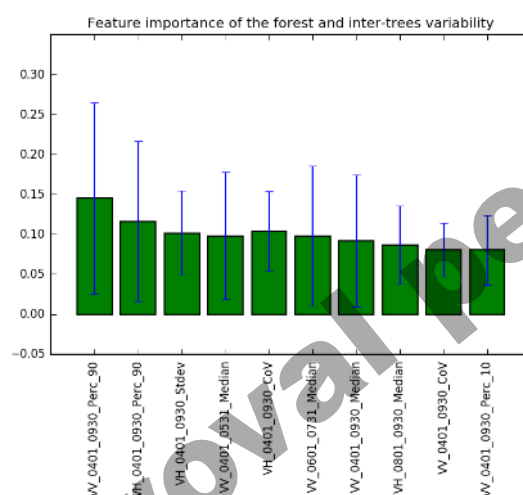


Figure 3-81: Top 10 S1 features for the grassland discrimination.

Figure 3-81 shows that both polarizations provide complementary information for the grassland identification. Both seasonal and annual features are considered useful for the grassland separation. All three mosaic over a two-month period are important for the discrimination of cropland and grasslands, but the spring period (April-Mai) and the summer period (June-July) are considered as more useful. Figure 3-81 also presents the feature selection results with the top 10 S1 features. The features are limited to 10 since it is shown in Figure 3-80 that the OOB (Out of Bag) error cannot be reduced using more than 10 features. 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.

Using the results of the feature analysis a grassland probability map is produced using the aggregated classes grassland and non-grassland. Based on the probability RF probabilities the classification results re derived with different thresholds. One third of the VIRP points are used for internal validation purposes and benchmarking. It has to be noted that using LUCAS points for validation purposes only allows to calculate count based accuracy metric, due to the fact that the inclusion probabilities of the points are not provided to the users.

Table 3-45: Count based accuracy metrics (in %) for random forest based classification for 2017 using S1 features.

	SAR 2017 >50%	SAR 2017 >55%	SAR 2017 >60%	SAR 2017 >65%	SAR 2017 >70%
Producer Accuracy	75.93	71.37	66.39	60.17	56.43
User Accuracy	58.84	62.32	67.23	70.39	75.98
Overall Accuracy	83.63	84.77	86.00	86.18	86.96

Table 3-46: Count based accuracy metrics (in %) for random forest based classification for 2018 using S1 features.

	SAR 2018 >50%	SAR 2018 >55%	SAR 2018 >60%	SAR 2018 >65%	SAR 2018 >70%
Producer Accuracy	67.63	63.9	59.34	52.28	45.64
User Accuracy	53.44	57.04	61.37	65.97	70.51
Overall Accuracy	80.63	82.13	83.45	84.14	84.42

Table 3-45 and Table 3-46 show the first results of the grassland classification based on the selected S1 features. The results show that with higher probability thresholds the producer accuracy decreases whereas the user accuracy increases. The overall accuracy does not change significantly. Using the threshold with 60% shows a balanced result. The random forest classification results confirm the conclusions derived from the threshold-based classification results based on the SAR data sets.

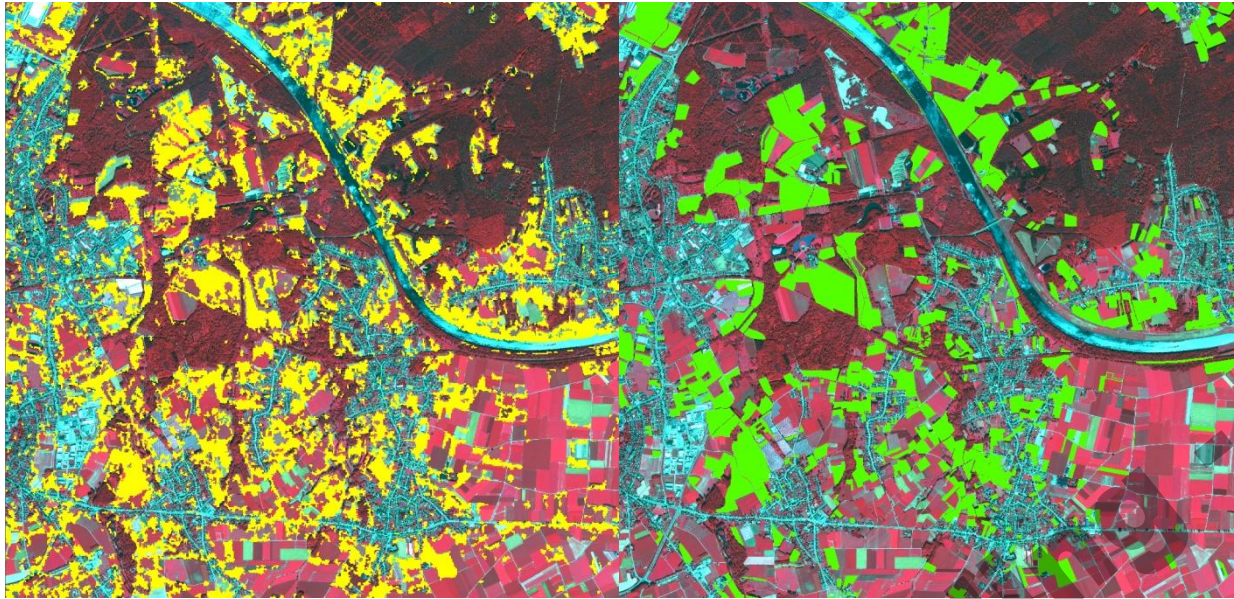


Figure 3-82: SAR grassland classification (grassland in yellow) with random forest and selected S1 features for 2017 ($p > 60\%$).

Figure 3-83: LGP grassland areas 2016 in green mapped on the World View 1 image from the 27. 06. 2018.

The classification results are filtered according to the defined MMU with 0.05 ha. Figure 3-82 and Figure 3-83 show that the filtered S1 classification tends 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.

RANDOM FOREST BASED GRASSLAND CLASSIFICATION WITH OPTICAL DATA

Further the use of only S1 data is also tested. In total 420 annual and seasonal S2 reflectance and indices based features are calculated as described in chapter 3.1. Figure 3-85 presents the feature importance of the listed features and shows that the importance slightly varies between the reflectance features.

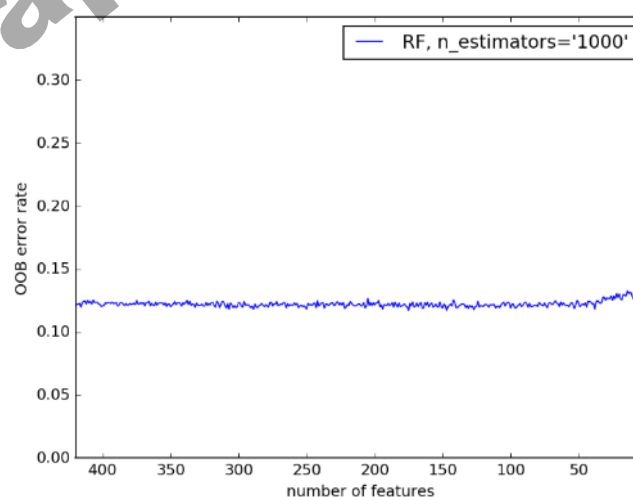


Figure 3-84: OOB error in relation the number of S2 input features.

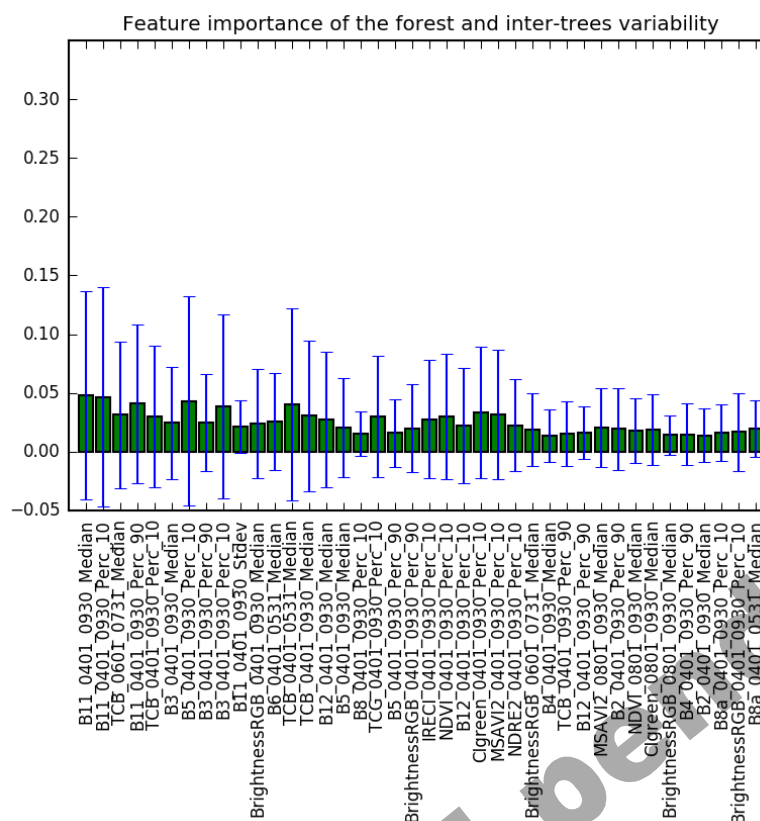


Figure 3-85: Top 40 S2 features for the grassland discrimination.

Figure 3-84 shows the feature selection results with the top 40 S2 features. The selection is mostly comprised by different annual features, which provide a more stable information in comparison to the seasonal features, which have to deal with a higher number of no data values. Seasonal features are included from summer and late summer capturing the phenology differences between grasslands and crop areas. Especially the differences between the growth period (spring/summer) and late summer are important. The features are limited to 40 since it is shown in Figure 3-85 that the OOB (Out of Bag) error cannot be reduced using more than 40 features.

According to the features selection analysis the most discriminative optical variables were then selected and the RF has been applied to derive a grassland probability map. Different thresholds are applied on the grassland probability maps to derive grassland/non grassland masks. To evaluate the classification performance, 1/3 of the VIRP points are used described in chapter 3.3.3.2. The results are presented in Table 3-47 and Table 3-48 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.

Table 3-47: Count based accuracy metrics (in %) for random forest based classification for 2017 using S2 features.

	OPT 2017 >50%	OPT 2017 >55%	OPT 2017 >60%	OPT 2017 >65%	OPT 2017 >70%
Producer Accuracy	85.89	83.4	79.67	76.35	70.54
User Accuracy	68.77	73.36	74.71	78.63	82.52
Overall Accuracy	88.73	90.05	89.96	90.57	90.58

Table 3-48: Count based accuracy metrics (in %) for random forest based classification for 2018 using S2 features.

	OPT 2018 >50%	OPT 2018 >55%	OPT 2018 >60%	OPT 2018 >65%	OPT 2018 >70%
Producer Accuracy	85.89	82.99	79.25	73.64	70.12
User Accuracy	71.13	73.26	77.33	80.73	82.44
Overall Accuracy	89.61	89.96	90.67	90.74	90.49

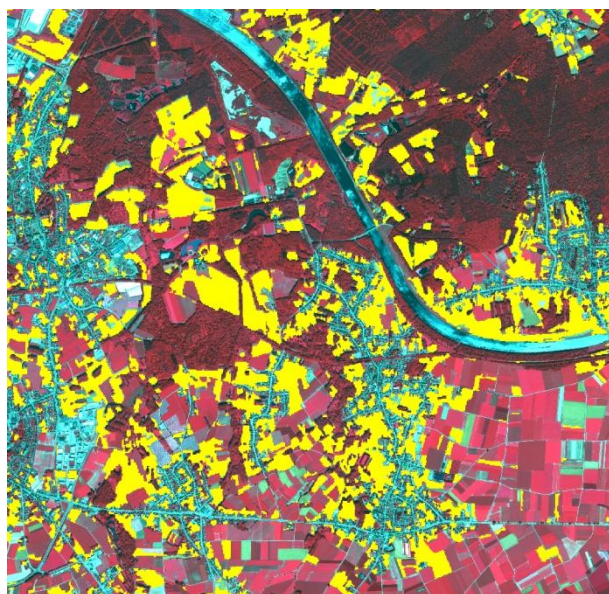


Figure 3-86: Optical grassland classification with random forest and selected features for 2017 (p>60%). (grassland in yellow)

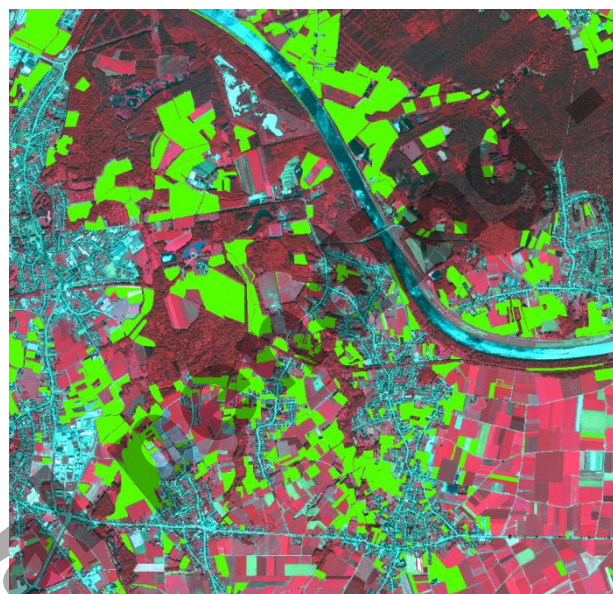


Figure 3-87: LGP grassland areas 2016 in green mapped on the World View 1 image from the 27. 06. 2018.

Figure Figure 3-86 and Figure 3-87 show 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. Compared to the SAR classification, the producer accuracy increased whereas the user accuracy decreased which leads to the conclusion that a combination of SAR and optical should improve the result.

RANDOM FOREST BASED GRASSLAND CLASSIFICATION WITH COMBINED S1 AND S2 DATA

Further, it is tested if the combination of both sensors might improve the grassland identification. The combined data set includes 460 features, 40 SAR features and 420 optical features. Figure 3-88 presents the feature importance of the listed features and shows that features of both sensors are included in the top 40 features. The features are limited to 40 since it is shown in Figure 3-89 that the OOB (Out of Bag) error cannot be reduced using more than 40 features. The selected variable set includes 13 SAR variables and 27 optical variables. It can be observed that the optical variables have a higher importance, indicating that the SAR variables are used to complement the optical variables.

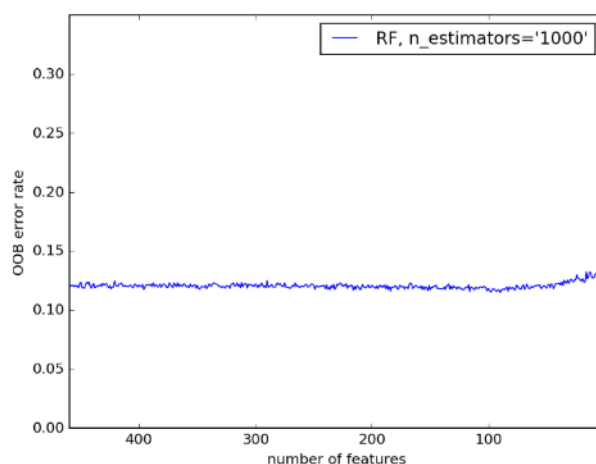


Figure 3-88: OOB error in relation the number of S1 and S2 input features.

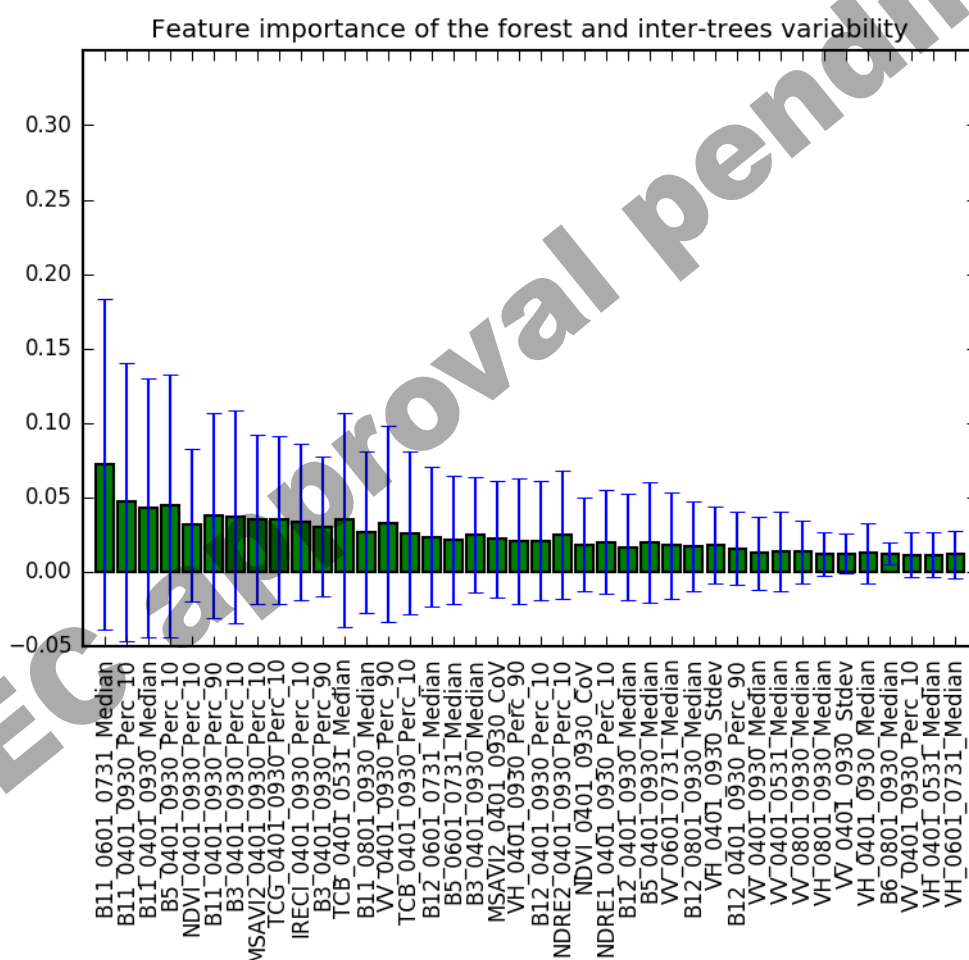


Figure 3-89: Top 40 S1 and S2 features for the grassland discrimination.

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 3.3.3.2. Table 3-49 and Table 3-50 show the results from the count based confusion matrices using optical and SAR variables. Globally the grassland mask is well separated. Using the 65% threshold produced the second highest overall accuracy and more balanced user and producer accuracies. Nevertheless,

it should be kept in mind that neither the classification maps nor the reference points are considering the MMU.

The results are presented in 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.

Table 3-49: Count based accuracy metrics (in %) for random forest based classification for 2017 using S1 and S2 features.

	SAR/OPT 2017 >50%	SAR/OPT 2017 >55%	SAR/OPT 2017 >60%	SAR/OPT 2017 >65%	SAR/OPT 2017 >70%
Producer Accuracy	86.31	85.89	79.67	79.67	72.5
User Accuracy	67.97	71.88	80	80	80.93
Overall Accuracy	88.47	89.88	91.11	91.46	90.57

Table 3-50: Count based accuracy metrics (in %) for random forest based classification for 2018 using S1 and S2 features.

	SAR/OPT 2018 >50%	SAR/OPT 2018 >55%	SAR/OPT 2018 >60%	SAR/OPT 2018 >65%	SAR/OPT 2018 >70%
Producer Accuracy	84.65	81.74	79.25	73.75	69.71
User Accuracy	65.38	67.93	72.62	77.97	80.77
Overall Accuracy	87.24	87.94	89.26	90.04	90.05

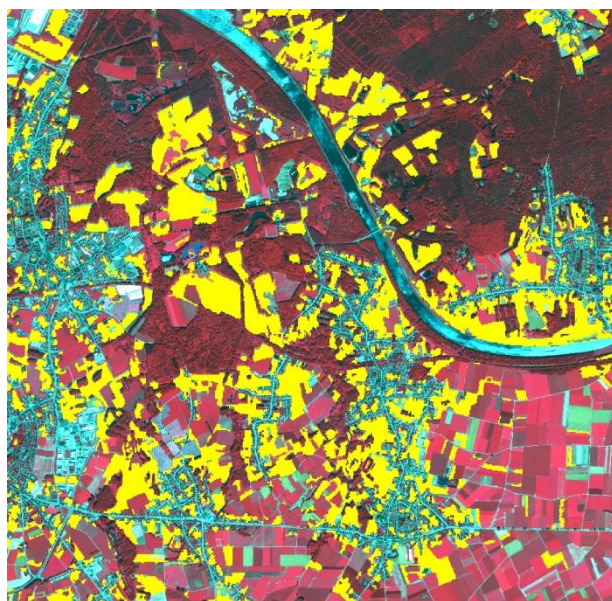


Figure 3-90: SAR + OPT grassland classification with random forest and selected features for 2017 ($p>50\%$). (grassland in yellow)

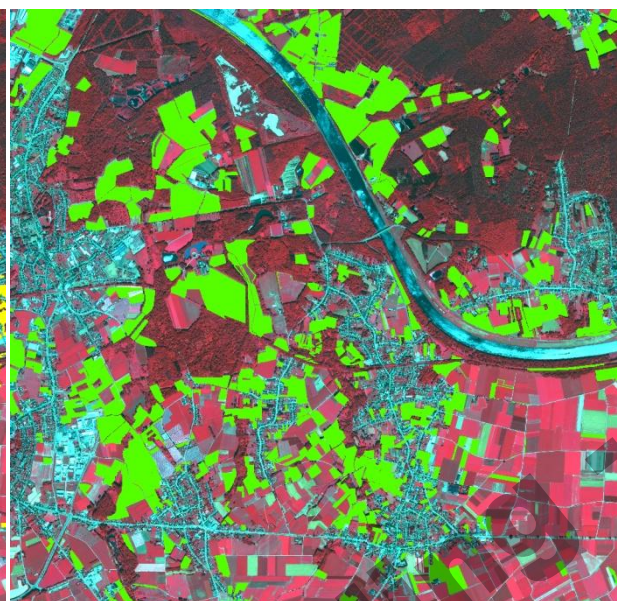


Figure 3-91: LGP grassland areas 2016 in green mapped on the World View 1 image from the 27. 06. 2018.

Further tests are applied using only 10m features to reduce the problems caused by geometric difference between the two resolutions and to be able capture small linear grass features more reliable. The resulting grassland probabilities are masked with a Vegetation/Non-Vegetation mask derived from the MSAVI2 over the vegetation period percentile 90 statistical feature and an empirically derived threshold. The result are shown in Table 3-51 and Table 3-52. Although no significant difference can be seen in the statistical evaluation, the visual evaluation showed that confusions between water areas and grasslands can be avoided using the rule based masking approach.

Table 3-51: Count based accuracy metrics (in %) for random forest based classification for 2017 using only S1 and 10m S2 features.

Only 10 m masked	SAR/OPT 2017 >50%	SAR/OPT 2017 >55%	SAR/OPT 2017 >60%	SAR/OPT 2017 >65%	SAR/OPT 2017 >70%
Producer Accuracy	85.89	84.23	80.08	75.93	70
User Accuracy	71.88	75.19	78.46	81.7	84.42
Overall Accuracy	89.88	90.76	91.11	91.29	90.92

Table 3-52: Count based accuracy metrics (in %) for random forest based classification for 2018 using S1 and 10m S2 features.

Only 10 m masked	SAR/OPT 2018 >50%	SAR/OPT 2018 >55%	SAR/OPT 2018 >60%	SAR/OPT 2018 >65%	SAR/OPT 2018 >70%
Producer Accuracy	85.89	83.4	80.91	78.42	73.64
User Accuracy	69.93	72.56	76.47	78.75	81.48
Overall Accuracy	89.17	89.79	90.67	90.93	90.92

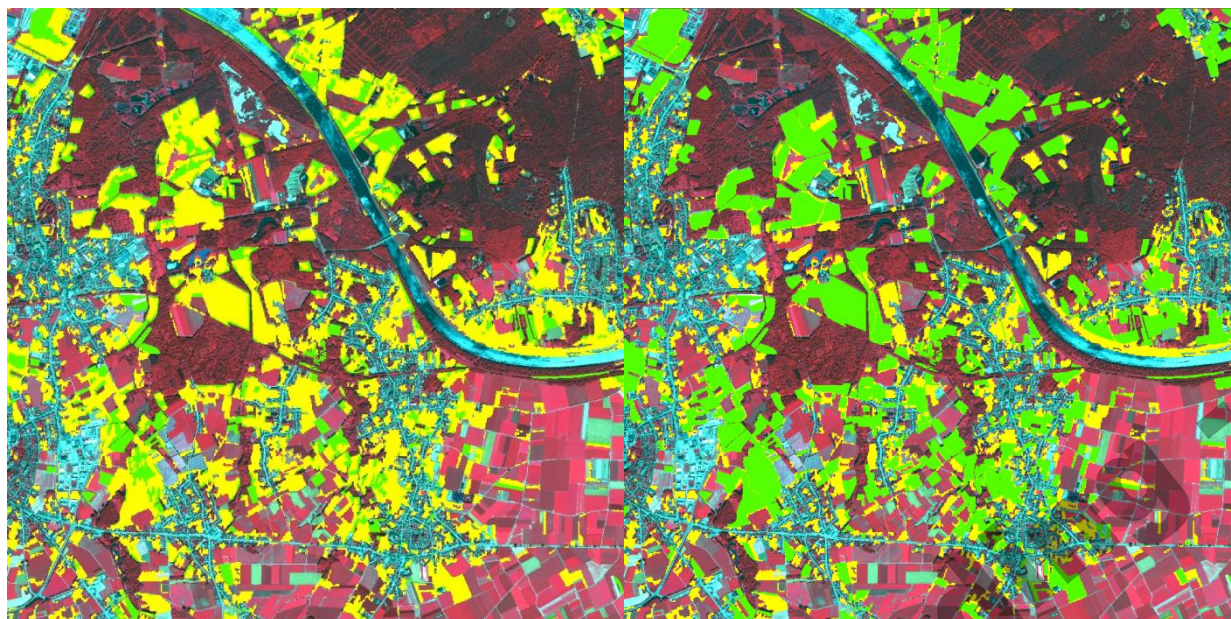


Figure 3-92: SAR + OPT classification 2017 10m aggregated: Omission errors (green). Classification (yellow) vs. LGP polygons 2016 (green) mapped on the World View 1 image from the 27. 06. 2018.

Figure 3-93: SAR + OPT classification 2017 10m aggregated: Commission errors (in yellow). Classification (yellow) vs. LGP polygons 2016 (green) mapped on the World View 1 image from the 27. 06. 2018.

As Figure 3-92 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 3-93 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.

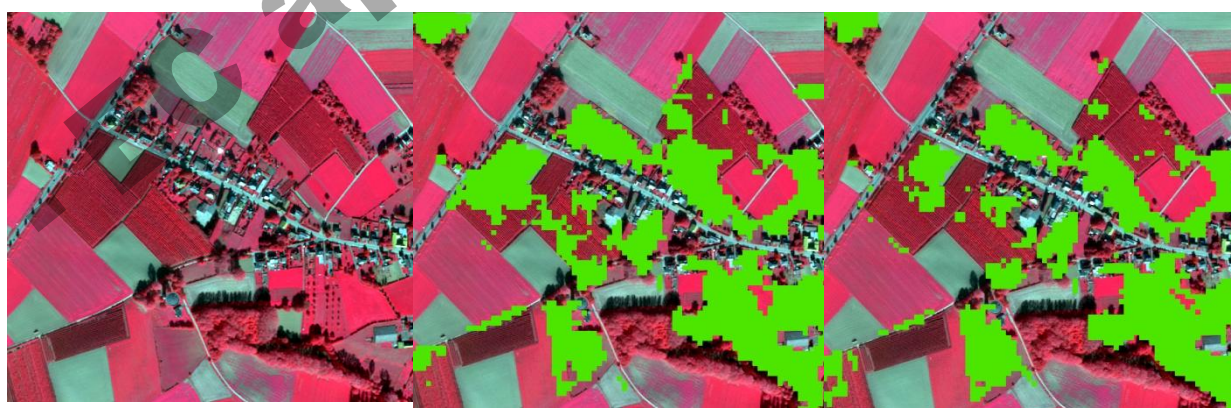


Figure 3-94: World View 1 image from the 05.10.2018.

Figure 3-95: S2 based classification mapped on the World View 1 image from the 05.10.2018.

Figure 3-96: S1/S2 based classification mapped on the World View 1 image from the 05.10.2018.

Table 3-53: Thematic accuracy (in %) comparison of different features.

	SAR 2017 p>60%	OPT 2017 p>65%	SAR/OPT 2017 p>65%	SAR/OPT 2017 10 m p>60%
Producer Accuracy	66.39	76.35	79.67	80.08
User Accuracy	67.23	78.63	80	78.46
Overall Accuracy	86.00	90.57	91.46	91.11

The classification result with SAR and OPTICAL combined datasets are quite encouraging. The combination of optical and SAR data showed significantly improved results with a producer accuracy of 79.67% (p>65%) and a user accuracy of 80% (p>65%) compared to using only S1 data and slightly better results compared with using only S2 data. The classification with combined datasets reduces SAR specific misclassification with roads and optical specific misclassifications with cropland and the optical data specific misclassification with tree orchards (see Figure 3-94, Figure 3-95, Figure 3-96).

Although the accuracies for the S1/S1 combined approach in comparison with the S1/S1 combined approach using only 10m features, show no significant improvement, in the visual interpretation it can be seen that the 10m features can better capture small linear features.

3.3.3.3.2 Demonstration site CENTRAL

REFERENCE DATA

For the grassland status layer in phase 2, the LUCAS 2018 points were used as additional trainings samples for test site central. In this case, the combination of the attribute observation type and the Copernicus module, was an effective way to select suitable samples. Additional samples were extracted from the HRL 2015, and manually selected to complete the reference datasets. As reported in the LUCAS 2018 feedback exchange with the EC and stakeholders [AD15], and other ECoLaSS deliverables, most issues with LUCAS points were found in the non-forest grassland class, which is explained by the fact that some points were collected in a transition area in between different land cover types (e.g., in the border between grasslands-forest areas). It was necessary to sample water bodies classes from additional datasets, other than LUCAS, as this minority class is underrepresented in the LUCAS dataset.

GRASSLAND MAPPING

The classification of grasslands in Central tests also proved the contribution of the SAR features to the improvement of the identification of grasslands. The combined approach, as can be observed in Figure 3-97, performs best. The visual inspection benchmarking criteria is key, as this enhancement is not totally clear from the accuracy statistics computed (i.e., from the confusion matrices, the overall accuracies and other quality metrics do not reflect this finding so evidently).

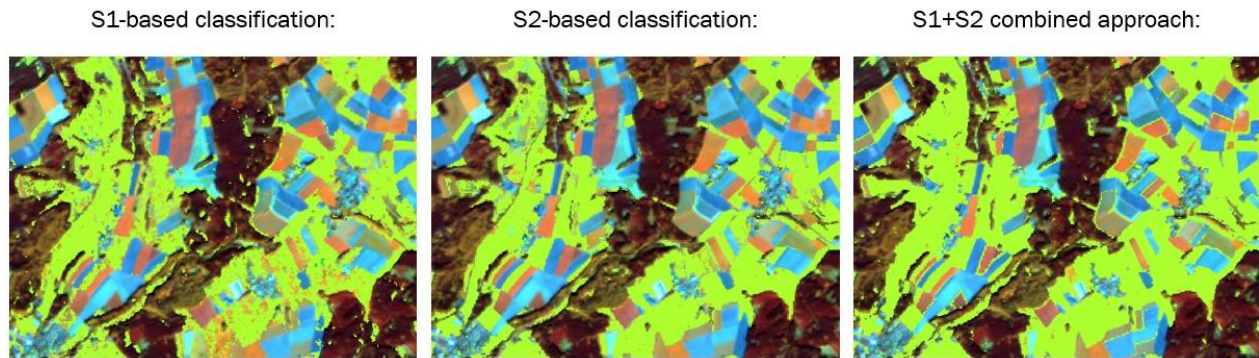


Figure 3-97: Comparison of multisensor grassland classifications: S1, S2 and combined in Central test site

The tests were carried out deriving random forest probability estimations based on annual optical and SAR features (minimum, maximum, mean, median, standard deviation and percentiles) over the spring period (seasonal time windows) and the whole year (annual time windows). The later winter to spring time windows was also tested in the 2018 grassland mask. In the case of the 2017 product, the late winter to spring could not be calculated because of data gap from January to March. This suggests that for recurrent updates, the data situation (i.e., cloud cover) in specific time windows might locally affect the production on a yearly/seasonal basis. In any case, in view of the performances, it is found that the time period should not be too short, as otherwise, the results are not meaningful due to a low number of scenes involved that might be not representative. It must be remarked that in different biogeographical regions, as confirmed in the other test sites, the shorter time window (spring in this case) is likely to be different. The cornerstone in this regard is to define a time window where grassland and cropland (responsible class for most miss-classifications) are best separable (e.g., grassland already greening when crops are not grown yet). In the Mediterranean areas, as described in section 2.3., this window may be shifted more towards the winter (e.g., December to March). In tests including complex time features, the added value was rather small. The most important features in the tests were percentiles.

Post-classification refinements consist of filtering based on a minimum pixel count of connected patches. All patches smaller than 5 pixels were removed to close wholes in grassland patches and remove very small. Filtering improves look and feel by reducing noise, caused by mixed pixels. Alternative approaches could imply a morphology filter which was not employed as it would change the shape of the classified patches. On the one hand, linear elements could have been removed by such a filter and, on the other hand, the patches nicely reflect the reality on the ground in most cases (i.e., the look and feel of the layer is satisfactory and the filter was not needed in the case of the Central tests). The improved grassland masks 2017 and 2018 at 10 m spatial resolution show more details when compared to the previous HRL 2015 at 20 m. When fulfilling task 4 scaling up to the larger demosite, limitations of the products to be taken into account were found in areas of higher elevation in the South, where snow cover is found for long periods of the year, hindering the classification accuracy locally (e.g., over- as well as underestimation of grassland was detected). In these areas, applying an elevation threshold enhances the classification results. Accordingly, a height threshold of 2800 m was applied for both status layers: all grassland above this height was removed. The refinements in the status layers clearly influence the performance in the change and incremental update products implemented in WP34 and WP35.

The visual inspection of the combined S1/S2 classification against LPIS polygons shows very few commission (green) and omission (red) errors, as can be appreciated respectively on the left and right detailed screenshots for the 2018 grassland mask in Figure 3-98.

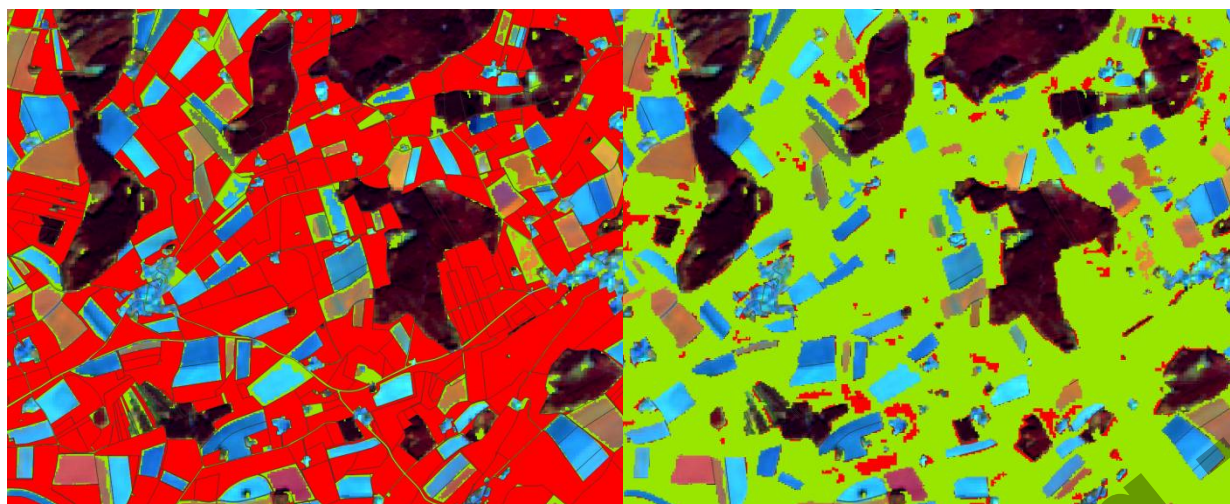


Figure 3-98: Omission and commission errors in the 2018 Grassland mask

The probability layers were also computed, considered an associated pixel based quality metrics by-product, mapping the areas where the classification is more reliable (Figure 3-99).



Figure 3-99: Probability layer for the grassland classification 2018 in the Central test site

The probability is dependent on the input data (which can be slightly different for each tile) and accordingly some tile border effects is present. As such, it is possible that the probability for grassland shows significant differences within a single patch that is located at a tile border. However, this is to be expected as the probability in the end shows how certain the classifier was in this specific case, with this specific data constellation, for a specific pixel to be grassland.

For computing the corresponding error matrices, the area-weighted accuracy calculation is applied as described in section 2.4. The statistical validation is performed on the basis of a LUCAS sample with count based accuracy calculation.

Once the grasslands status layers for 2017 and 2018 are produced, the change detection product is derived, as described in WP34 [AD08]. It must be remarked that the production of the change layer is in turn dependent on the status layers from which the change is derived.

GRASSLAND INTENSITY MAPPING

Different tests were also carried out to derive a new and more detailed product in the grassland areas within the grassland mask. The use intensity layer at 10 m spatial resolution is based on the number of mowing events detected. The definition in ECoLaSS is that the binary mask produced within the grassland masks considers the grasslands are intensively used when three or more mowing events are detected, and extensively managed otherwise. In this sense, natural grasslands can be derived by considering no mowing events (i.e., zero mowing events detected).

First, the coherence features from SAR data were tested, although in the end it proved not that satisfactory, and taking into account the upscaling of the products to larger scales in a cost-efficient manner, another approach was successfully tested instead and applied to the demosite scale. It was concluded that coherences are highly sensible to changes, even on micro-level, and therefore, events like heavy rainfall are likely to be messed up and make coherence unusable for intensity analysis, among other applications. This is highly risky, besides the expense of the processing of SAR coherences, when considering automation and large scale products. Consequently, in Central a more stable approach based on NDVI time series was elaborated and tested. Training samples are generated based on the IACS/InVeKoS samples through an outlier detection and an independent visual interpretation with Sentinel-2 time series data and additional VHR data if available. The use of temporal trajectories, seasonal statistical features and phenological features is investigated regarding the intensity to achieve the optimal set of features/indices per biographic region/elevation stratum. NDVIs were computed for all scenes available in 2018 to detect mowing events all throughout the year.

Mowing events were detected by comparing NDVI minimums of consecutive acquisitions and a rule based classification, defining that all pixels with ≥ 3 mowing events are intensively used and 0-2 mowing events means extensively used. A first validation based on INVEKOS data (where available) is quite promising. In view of the tests results, it is clear that the high cloud coverage in 2018 in Central Europe limits the method of comparing NDVIs of consecutive acquisition dates. Mowing events may not be detected in areas covered by clouds for a longer period of time.

For sampling and for the validation of the Grassland Use Intensity Layer the new grassland attributes in the LUCAS 2018 samples, available for some points of the LUCAS 2018 data in the demosite, would in theory be very helpful. However, as the number of points having these attributes is too low, and the sampling density might prove too low, the actual benefit could not be tested. The validation samples are extracted from INVEKOS Austria (i.e., German for LPIS/IACS), where available. INVEKOS Austria contains information about the mowing frequencies, which is perfect for validation. Unfortunately, this data is only covering Austria and therefore only parts of the demosite: two out of nine tiles in the case of Central. Consequently, the use intensity layer cannot be fully validated, nor at the demo scale nor for a potential Pan-European/global roll-out. This can only be achieved when reference data is available. A qualitative inspection was implemented instead.

As for the other layers, filtering improves significantly the look and feel by reducing noise. For the use intensity layer, a filter of 4 pixels in size was applied. All areas within the grassland mask were filtered, so that there is no patch for one of the two intensity classes smaller than 5 pixels in the end. Within small grassland patches, it might happen that e.g. 3 pixels are classified as extensive and 2 are classified as intensive. In such cases, the filter would cause the class values to jump between classes with each filter iteration without getting a patch of 5 unique values. If so, it was filtered in favor of intensive use because most of the areas are used intensively in the demo site. In any case, the number of mowing events layer, which is the previous step to the binary extensive/intensive use decision, is available for consultation. This layer is also useful to check for natural grasslands if it is

assumed that the latter are present when no mowing events are detected at all. For this assumption to be more reliable, a longer time series (e.g., several years) should be considered.

The figure below shows the test in Central of the grasslands use intensity in 2018. It can be observed that grasslands are extensively managed in Alpine regions whereas more intensively in valleys around settlements.

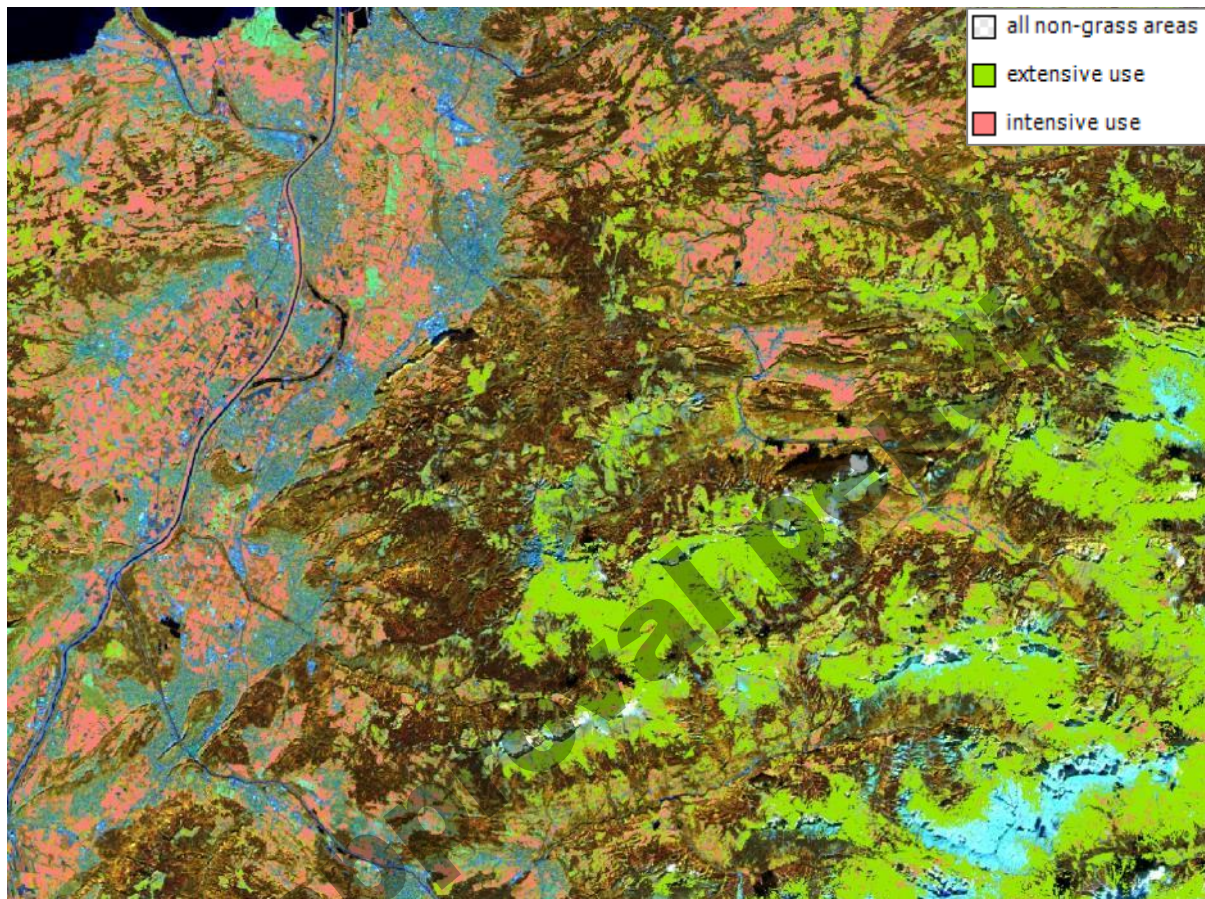


Figure 3-100: Grassland use intensity in Central 2018.

A further approach based on Kalman filtering has been tested. The development and validation of the proposed method is based on the previously mentioned INVEKOS Austria data set, which provides a thorough characterization of the agricultural use of land including the mowing frequencies of grassland. Two tiles (32TNT and 32TPT) of the Central demo site are partially covered, therefore a test site representing the intersection of the tile boundaries with the bounding rectangle of the INVEKOS layer has been defined.

Sentinel-2 images for the respective tiles acquired from the relative orbits R065 and R022 have been downloaded in order to create a time series from March to November 2018. Images with a nominal cloud cover >85% according to the COPHUB metadata have been discarded. Both TOA (level L1C) and BOA (level L2A) versions of the images have been acquired. The L1C data is needed to compute cloud masks using the Fmask 4.0 tool.

TRACKING WITH KALMAN FILTER:

A further approach based on Kalman filtering has been tested. The development and validation of the proposed method is based on the previously mentioned INVEKOS Austria data set, which provides a thorough characterization of the agricultural use of land including the mowing

frequencies of grassland. Two tiles (32TNT and 32TPT) of the Central demo site are partially covered, therefore a test site representing the intersection of the tile boundaries with the bounding rectangle of the INVEKOS layer has been defined.

Sentinel-2 images for the respective tiles acquired from the relative orbits R065 and R022 have been downloaded in order to create a time series from March to November 2018. Images with a nominal cloud cover >85% according to the COPHUB metadata have been discarded in the first place. Both TOA (level L1C) and BOA (level L2A) versions of the images have been acquired. L1C images are required to compute cloud masks using the Fmask 4.0 tool, whereas the masked L2A images represent the input to the Tasseled Cap tracking algorithm previously outlined used to estimate the number of mowing events. Note that the algorithm does not require additional disk space to store intermediate results and the only required output is a single raster file specifying the estimated number of mowing events. The intensity classification depicted in Figure 3-102 has been derived from this result.

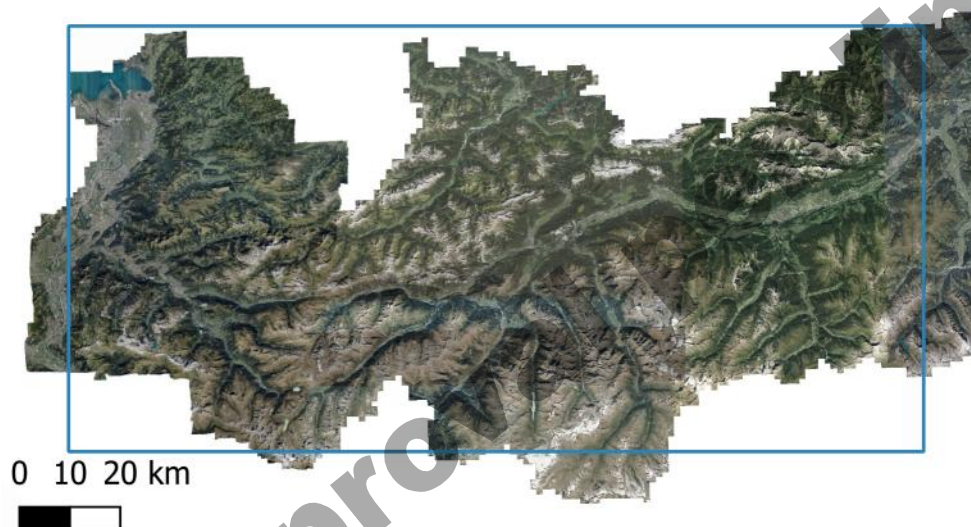


Figure 3-101: Location of the test site in western Austria (Background © [basemap.at](https://www.basemap.at))

The INVEKOS dataset provides polygons delineating agricultural parcels. Each polygon features a class label according to its land use type. The following table lists all classes corresponding to grassland together with the summed area of the associated polygons. Three class labels explicitly state the number of expected mowing events, while for the remaining classes it needs to be assumed. The only class clearly corresponding to high mowing intensity is “Meadow (3 or more mowings)” which also represents more than 38% of the area within the test site. Two other classes are assumed to be mowed intensively, however they represent only a marginal portion of the total area and therefore have little weight.

Table 3-54: INVEKOS grassland classes and associated area within the test site

Class (translated)	Original class label (in German)	Area [ha]	Area %
Mountain meadow	Bergmähder	1818.873	2.27%
Pasture	Dauerweide	2469.965	3.09%
Meadow (1 mowing)	Einmähdige Wiese	5382.739	6.72%
Fodder grass	Futtergräser	55.477	0.07%
Fallow grassland	Grünbrache	19.535	0.02%
Fallow grassland	Grünlandbrache	23.686	0.03%
Pasture	Hutweide	11607.86	14.50%
Clover	Kleegras	643.247	0.80%
Meadow (3 or more mowings)	Mähwiese/-weide drei und mehr Nutzungen	30487.732	38.08%
Meadow (2 mowings)	Mähwiese/-weide zwei Nutzungen	25106.346	31.36%
Litter grass	Streuwiese	2442.283	3.05%
Sum		80057.743	100.00%
Extensive mowing (0 - 2 times)			
Intensive mowing (>2 times)			



Figure 3-102: Mowing intensity map based on Tasseled Cap tracking (within INVEKOS grassland mask)

In order to assess the result of the grassland mowing intensity mapping, the INVEKOS polygons are rasterized to 10m resolution with pixel values corresponding to the mowing intensity indicated by the class label. Figure 3-103 shows the agreement between map and reference layer. Positive agreement outweighs in many parts of the test site, but there are also quite large regions of predominant disagreement. The confusion matrix in

Table 3-55 reports an overall agreement of 76.54% and indicates that intensively managed areas are underestimated. The illustrations of Figure 3-104 to Figure 3-108 document two conditions leading to differences between map and classification.



Figure 3-103: Agreement between the INVEKOS reference layer and the mowing intensity map.

Table 3-55: Confusion matrix for the mowing intensity map

Mowing intensity mapping using Kalman filter		REFERENCE			User Accuracy	Confidence Interval
		Extensive	Intensive	Total		
PRODUCT	Extensive	52.12%	14.54%	66.66%	78.19%	4.16%
	Intensive	8.92%	24.42%	33.34%	73.24%	6.36%
	Total	61.04%	38.96%	100%		
	Producer Accuracy	85.39%	62.67%		76.54%	Overall Accuracy
	Confidence Interval	3.04%	4.48%		3.49%	Confidence Interval
					0.49	Kappa
					0.82	F1-score
					0.67	F0-score

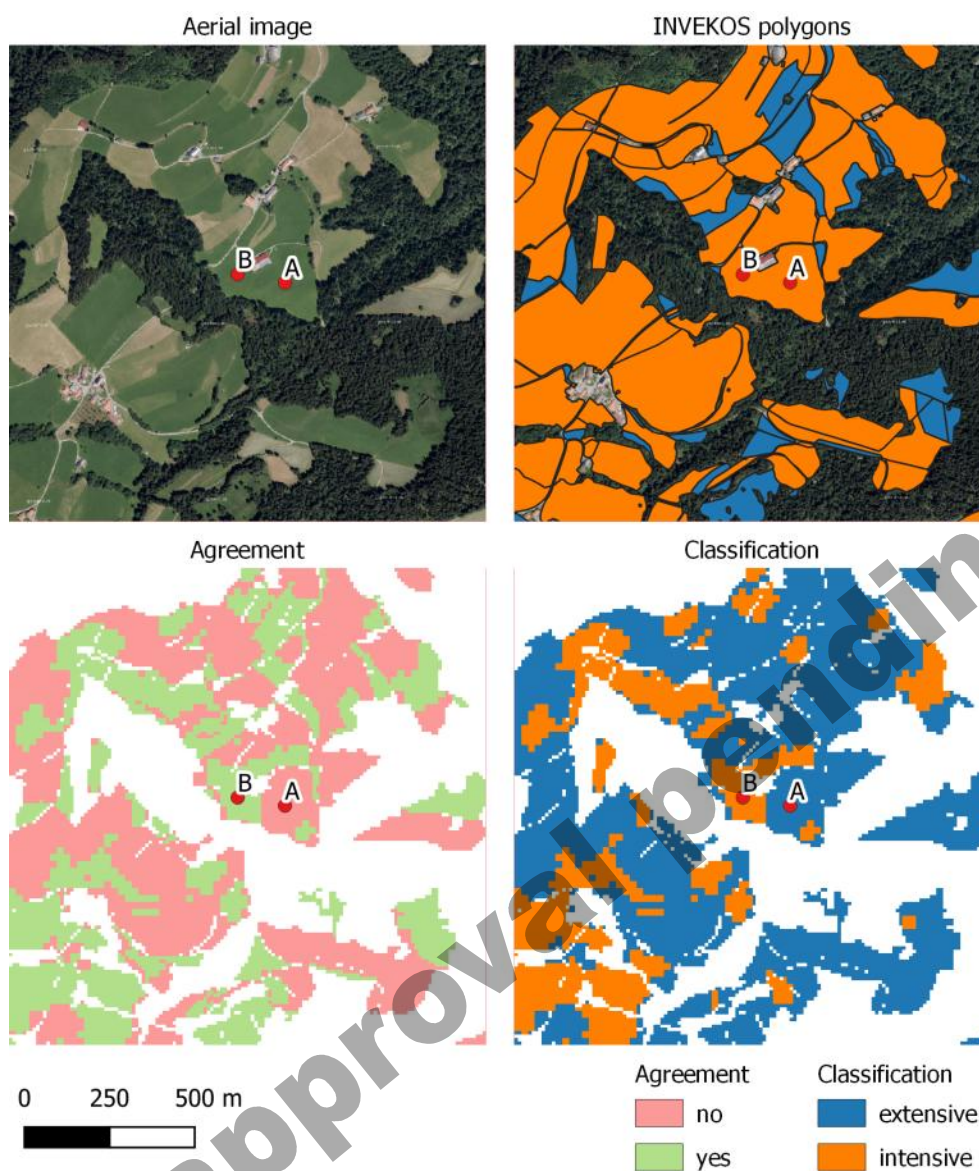


Figure 3-104: Detail analysis of an area with poor agreement between reference and map. Inspection of the available observations for sample pixels A and B (see Figure 3-105) indicates problems caused by data gaps.

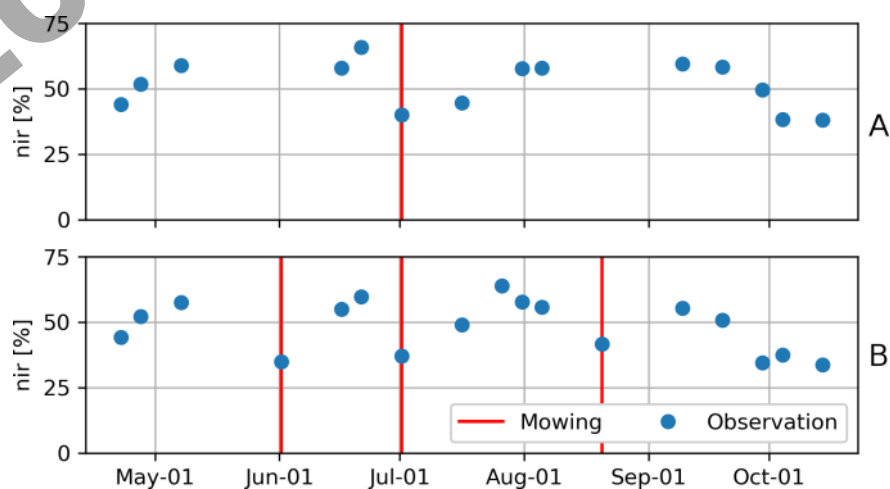


Figure 3-105: NIR time series of sample pixels A and B (see Figure 3-104). Two key observations required to detect mowing events are missing in series A, but not in B.

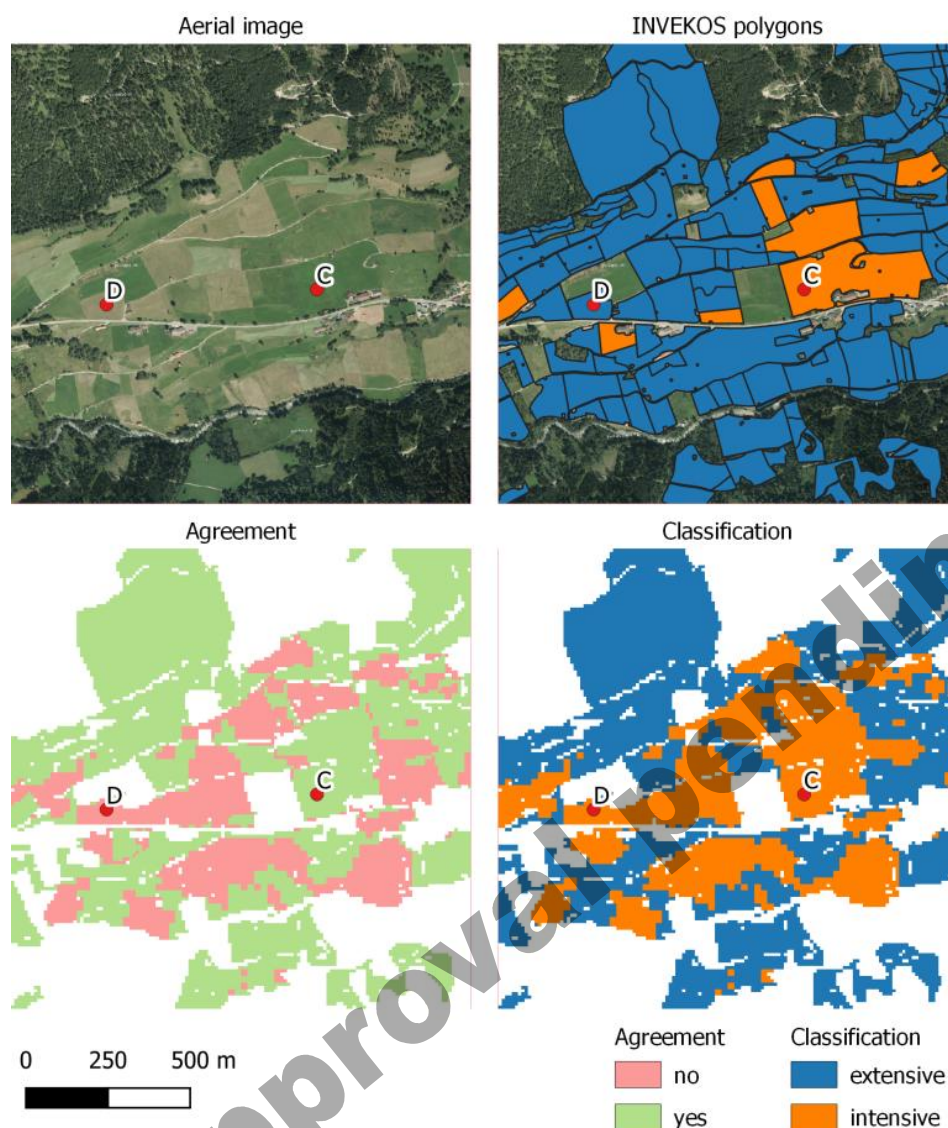


Figure 3-106: Analysis of an area with moderate agreement between reference and map. Inspection of the estimated state variables for sample pixels C and D (see Figure 3-107 and Figure 3-108) suggests that the reference could be wrong in this case.

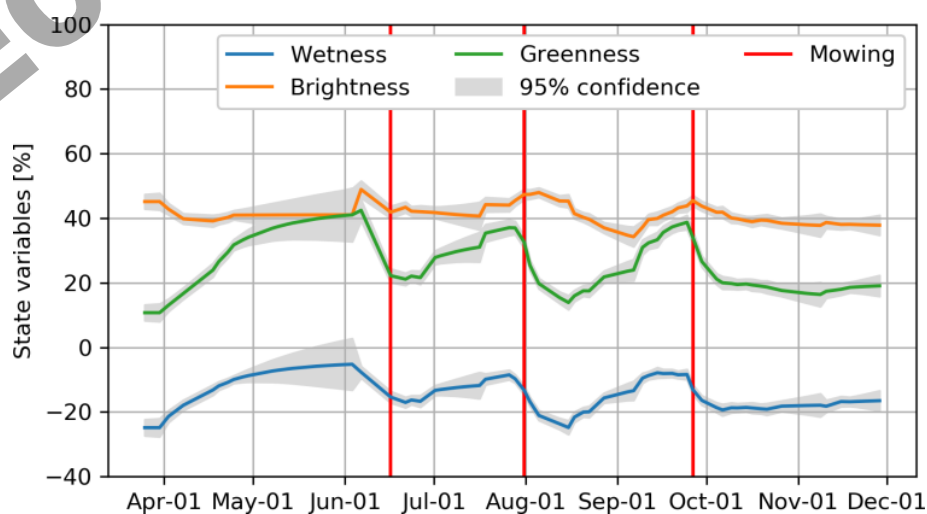


Figure 3-107: Estimated state variables of sample pixel C (see Figure 3-106). Greenness and Wetness patterns indicate three mowing events, which is in agreement with the reference.

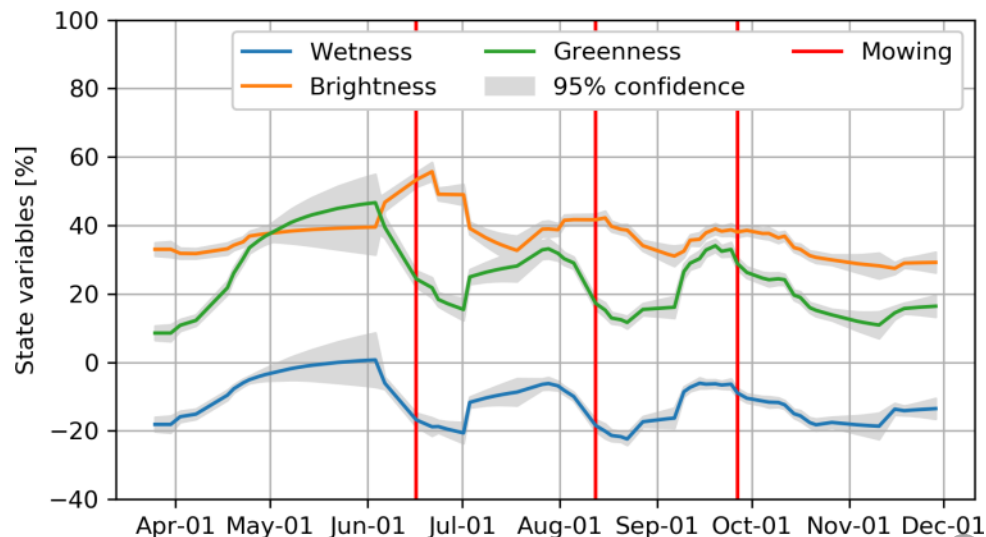


Figure 3-108: Estimated state variables of sample pixel D (see Figure 3-106). Greenness and Wetness patterns indicate three mowing events similar to sample pixel C. However, the reference states extensive usage.

Several concluding remarks can be made based on the presented results:

- Mowing events can be detected from optical remote sensing time series with reasonable reliability if the observation density is given
- If data from overlapping orbits is not available, the number of clear-sky observations can become critically low and a reliable detection of mowing events is unrealistic.
- The statistical significance is influenced not only by the vectors' magnitude, but also by the length of the time gap between consecutive observations.
- Large gaps in the time series will result in a lower sensitivity of the detection method, because the algorithm has not enough information to distinguish between abrupt and gradual signal changes

3.3.3.3 Demonstration site SOUTH-EAST

In the demonstration site SOUTH-EAST benchmarking was only performed during phase 2 of the project. While the classifier, i.e. random forest, was fixed already after the tests made in phase 1, benchmarking was applied in terms of predictor set selection, that is, S1, S2 and S1+S2, and feature selection.

REFERENCE DATA

For the testing the grassland classification in the SOUTH-EAST site, the LUCAS 2018 points were used as trainings samples. In total 3871 LUCAS samples were available in the demonstration site, of which 743 belonged to the grassland classes. LUCAS data were filtered by identical criteria to the demonstration site WEST lined out in Table 3-41. After filtering, 2168 LUCAS samples remained (482 grassland samples), of which 25% were set-aside for internal validation. The LUCAS land-cover classes were then converted into binary form, that is, "grassland" and "non-grassland".

MAPPING ALGORITHM:

The number of trees in the random forest was set to 1000, although performance did already stabilize at around 500 trees in most cases. The number of predictors to evaluate per split was set to the square root of the total number of predictors, which is the recommended default value for discrete responses. Initial attempts at tuning this parameter did not improve the results and were therefore not pursued further.

FEATURE DEFINITION:

From the S2 Level-2A data the following spectral indices were calculated for each acquisition date: NDVI, GNDVI, NDWI, NDRE1, NDRE2, MSAVI2, mean SWIR; IRECI, CI_red_edge, PSRI, REP and MCARI, Tasselled Cap Brightness, Tasselled Cap Wetness and Tasselled Cap Greenness. From the S-1 data gamma naught backscatter values were used.

Based on the single-date spectral bands, spectral indices and backscatter coefficients a suite of multi-temporal metrics were derived for relevant time-periods throughout the vegetation period. The statistics derived were the 10%, 50% (median) and 90% percentiles, the mean, the standard-deviation as well as the coefficient of variation. The aggregation time-periods which were chosen were:

- March – October for capturing the yearly characteristics during the vegetation period;
- Bi-monthly intervals: March – April, May – June, July – August and September – October;
- Tri-monthly intervals March – May, June – August and September – October.

FEATURE SELECTION:

Feature selection was accomplished by backwards feature selection. Overall, 1200 Sentinel-2 and 66 Sentinel-1-based time-series predictors were tested, resulting in a total number of 1266 potential predictors. Based on this pool of potential predictors the random forest Gini-based feature importance was used in a backward, element-wise feature selection procedure. For S1 the optimal set comprised 24 predictors, for S2 optimal performance was achieved with 39 predictors. The performance of the combined S1/S2 model was superior throughout and peaked at 36 predictors.

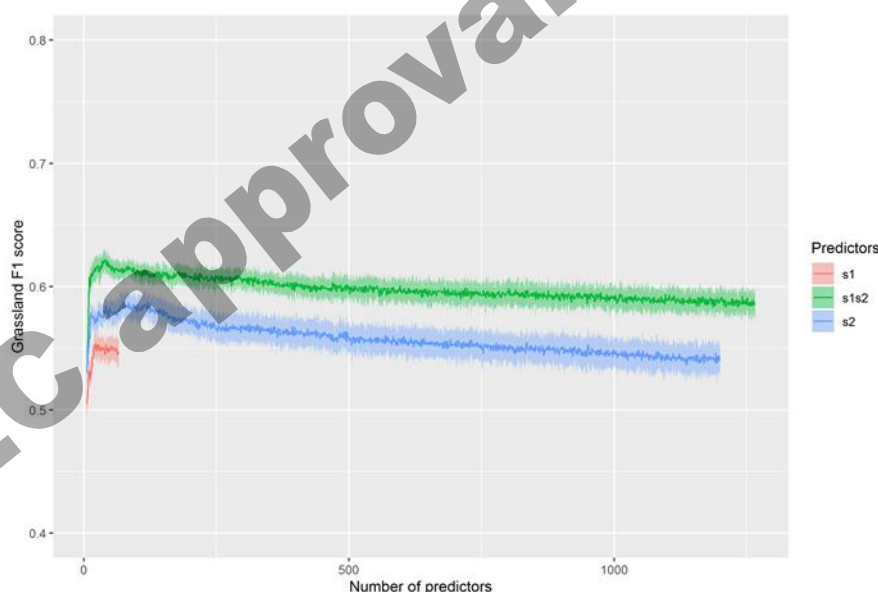


Figure 3-109: Development of out-of-bag grassland F1 score during Gini-based backward feature selection for the predictor sets: Sentinel-1 (S1), Sentinel-2 (S2) and combined features (S1/S2).

The observation that model performance increased with a decreasing number of predictors (Figure 3-109) is an indication of a sample size limitation. Even though the random forest algorithm is to a large degree robust to very high-dimensional feature spaces, this behaviour shows that it is likely that an improvement could have been possible had more training samples been available.

While there is a high variability in the ranking of the feature importance due to their extreme co-linearity, some insights can be gained from identifying features which were selected both in 2017

and 2018. Table 3-56 shows this comparison. From the optical data, primarily S-2 Band 3 (red) and S-2 Band 11 (SWIR) played an important role, as well as the red-edge position quantified by the REP spectral index, which locates the red-edge inflection point and is for the first time available operationally and in high-spatial resolution due to the S-2 mission. In both years 2017 and 2018 further vegetation indices played a role, but due to their high correlation they were not stable in their importance and could be handled interchangeably between years. The S-1 SAR features were much more stable in their relative importance in both years, which is likely due to their more consistent time-series, which are barely influenced by cloudcover and illumination effects. Notably primarily mean and high-quantile backscatter of both VV and VH over selected time-periods turned out to be important in both years, while standard deviation as a measure of dispersion was not frequently found to be important.

Table 3-56: Important variables identified for Grassland status layer production in demonstration site South-East in both years, 2017 and 2018 for the combination of S1+S2 features. Temporal statistics: q10 = 10% percentile, q50 = 50% percentile (median), q90 = 90% percentile, sd = standard deviation, cv = coefficient of variation. TC green = tasselled cap greenness component, TC wet = tasselled cap wetness component.

Type	Variable	Time period (months)	Statistic
S2, spectral band	Band 3	03-10	q10
S2, spectral band	Band 5	03-10	q10
S2, spectral band	Band 11	09-10	q10
S2, spectral band	Band 11	09-10	q50
S2, spectral index	REP	03-10	q90
S2, spectral index	REP	05-06	q90
S2, spectral index	REP	05-06	sd
S1, gamma naught	VH	03-04	q90
S1, gamma naught	VH	03-10	q90
S1, gamma naught	VH	06-08	q90
S1, gamma naught	VV	03-05	q90
S1, gamma naught	VV	03-05	mean
S1, gamma naught	VV	03-10	q90
S1, gamma naught	VV	03-10	mean
S1, gamma naught	VV	07-08	mean

PERFORMANCE BENCHMARKING S1, S2, S1+S2

Benchmarking the predictor input sets was accomplished based on the test-set set aside before feature selection. The test-set benchmarks for the years 2017 and 2018 are shown in Table 3-57 and Table 3-58, respectively. In all cases, model performance of S-2 was superior to using S-1 only, however model performance was highest when using both S-1 and S-2 input features. This means that on the one hand there is a notable amount of overlapping information contained in S-1 and S-2 data, however, there is also complementary information, which is essential to achieving the highest possible performance. While processing both S-1 and S-2 data for a potential pan-European roll-out, certainly entails a non-negligible processing overhead, these results demonstrate the added value of doing so.

Table 3-57: Test-set count based accuracy metrics (in %) for random forest based classification for 2017 using S1, S2, and S1+S2 features.

2017	S1	S2	S1+S2
Producer Accuracy	41.79	49.23	54.72
User Accuracy	60.85	68.06	68.24
Overall Accuracy	81.46	82.80	83.87

Table 3-58: Test-set count based accuracy metrics (in %) for random forest based classification for 2018 using S1, S2, and S1+S2 features.

2018	S1	S2	S1+S2
Producer Accuracy	40.30	45.13	66.51
User Accuracy	62.79	70.81	73.26
Overall Accuracy	81.79	83.85	88.50

3.3.3.4 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 in 2016, 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).

As the results of Task 3 have shown, the synergetic use of temporal features derived from optical and SAR data streams enhances the accuracy of the classifications in comparison to either single optical or SAR only approaches. High-frequency optical and SAR acquisitions over the growing seasons were therefore used to derive temporal features over the grassland growing season as input for the classifications. The generation of suitable time features, especially considering upscaling to pan-European or global levels, is challenging and requires large computational capacities. Automated feature selection at the bio-geographic level is therefore applied, to reduce the computational effort for potential future operational large area roll-out. While attributing an absolute importance to each feature is not straight forward due to their high co-linearity, the ability to reduce the number of features which need to be computed beforehand without compromising mapping performance is the most important outcome. From the perspective of an operational implementation, a valuable observation is that the number of input features may remain moderate around at a maximum of 50 predictors.

Some commission errors have been detected in orchards and frequently occurring mixtures of grassland and shrubland. Cropland can be excluded from the layer with a higher reliability if the ploughing event is captured in the time series. If that is not the case, the crop areas share the similar spectral signatures with grass areas and cannot be excluded. The ploughing event is the main character to discriminate the grassland areas from crop areas. The grassland over detection due to missing ploughing events in the time series is the primary issue regarding the classification for 2017, where the time series is a lot sparser. Changing or abnormal environmental conditions between seasons also have an effect on the classification result. The drought period in summer 2018 resulted in a under detection of grassland areas as their spectral signature has been similar to some crop areas. Further, some misclassifications with fruit orchards remain as the grass cover between the orchards influences the spectral signature and the algorithm cannot separate them. Nevertheless, most of the orchards can be excluded using S1 and S2 depending on the tree size and management practices. It must be taken into account that no historical time series were included in the grassland classification process to detect ploughing events and exclude grassland younger than 5 years, therefore agricultural grassland is included in the status layer. Considering data availability and the not homogeneous data situation across Europe, another consideration towards larger area and constant production, is that the time period should not be too short, as otherwise results might not be meaningful due to the likeliness of limited number of scenes in some regions. Changing or abnormal environmental conditions between seasons also have an effect on the classification result.

3.3.4 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 in Europe, plus the experiences in the African test sites.

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). Some of the requirements related to agriculture layers have been compiled in WP21 [AD05].

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. The interest in this topic was tackled in the CLMS H2020 sister projects second meeting in Spain in April 2019, and the IACS Workshop event. ECoLaSS experiences with the agriculture testing and prototypes were presented, and fruitful interactions with the other projects (e.g., SENSAGRI, SEN4CAP) benefited the developments in phase 2. WP41, WP44 and WP33 implementations embed this trend and requirements compiled in WP21 for the agriculture products in ECoLaSS.

To summarize, ECoLaSS aims at deriving methods for a potential future HRL Agriculture that additionally could provide information on a yearly basis to be used as additional input to CAP, if desired.

Due to the differences between the biogeographic regions, and the need of adopting case-wise approaches and WP33 targeting methods compendia, the subsection on Agriculture, is structured differently. Within, section 3.3.4.1 describes the tests carried out in the Central test site in Germany/Austria, section 3.3.4.2 describes testing carried out in the Belgium site and section 3.3.4.3 compiles the experiences in the tests developed in the African sites.

3.3.4.1 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.

3.3.4.1.1 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 S Sentinel-1 or Sentinel-2 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 (Sentinel-1, Sentinel-2, and Sentinel-1 & Sentinel-2), different time windows 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. Particularly, the following research questions were analyzed for the Central Site:

- accuracies to 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 (phase 1 + 2),
- improvements and performances of the crop types classification by selecting a suitable crop type structure (phase 1 + 2), aggregating the results on field basis (phase 1 only),
- optimization of a number of features with respect to the full feature set without a significant accuracy decrease (phase 1 + 2),
- effect of data and features from the late season of the previous years over crop types accuracies (phase 1 only),
- how well the crops can be classified during the growing period (phase 1 + 2), and
- whether it is possible to provide comprehensive information that enables users to assess the reliability of a prediction (focus of phase 1).

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 Sentinel-1 and Sentinel-2. 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 could be saved. This was assessed in phase 1. It is important to stress, that no segmentation of the image data was performed. However, the crop type reference data was at least partially 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. 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.

3.3.4.1.2 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.

3.3.4.1.3 Implementation and results of benchmarking

CLASSIFICATION INPUT DATA

In phase 1, 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 November 2017 were downloaded and pre-processed for the two Sentinel-2 tiles T32UNV and T32UNU (Table 3-59). The number of available scenes for each tile is shown in Table 3-59.

In phase 2, the test sites are located in the tiles 32UNU and 32TNT, for which the Sentinel-1 and Sentinel-2 time series of the vegetation period starting from Mid-March 2018 to Mid-October 2018 were downloaded and preprocessed. The autumn and winter season of the previous year is not included anymore because testing in phase 1 revealed only little benefit. Instead, the approach of phase 2 focused on the main vegetation period during spring and autumn. Taking into account the highly heterogeneous growing periods of the different crops, the classification approach tried to cover the whole growing period starting in Mid-March and ending up in Mid-October where EO imagery offers most information on spectral characteristics and texture, basing on varying phenology and biophysical aspects. The respective number of available scenes for each tile is shown in Table 3-60.

Table 3-59: Phase 1 - Number of Sentinel-2 (< 50% Cloud cover) and Sentinel-1 scenes for the period October 2016 - December 2017.

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

Table 3-60: Phase 2 - Number of Sentinel-2 (<90% Cloud cover) and Sentinel_1 scenes for the growing period Mid-March 2018 to Mid-October 2019

	32UNU	32TNT
Sentinel-1	151	87
Sentinel-2	139	98

In both phases, the Sentinel-2 imagery was atmospherically corrected and topographically normalized using the ESA Sen2Cor software [AD07]. Since in 2018 the cloud cover was generally high, imagery with cloud cover up to 90% was used for classification and analysis in order to get time series as dense as possible. 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 Layer produced by Sen2Cor. As an additional consequence of the high cloud cover throughout 2018 and caused by the product tiling of the Sentinel-2 data, the number of available scenes per tile varies and ends up in differing density of imagery for classification especially in the northern tile of the test site compared to the Southern one. This fact is represented by differing Data Score Layer of the neighboring tiles. This issue has been observed in both testing phases of the project(Phase 1: Figure 3-110, Phase 2: Figure 3-111).

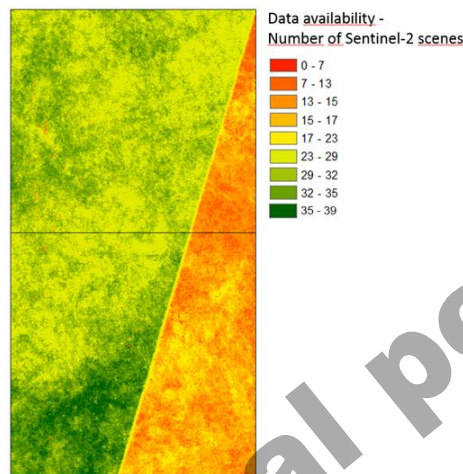


Figure 3-110: Phase 1 - Sentinel-2 data score (inverted cloud value count) for ECoLaSS central test site (T32UNU+32UNV tiles) for the time period March-Nov 2017.

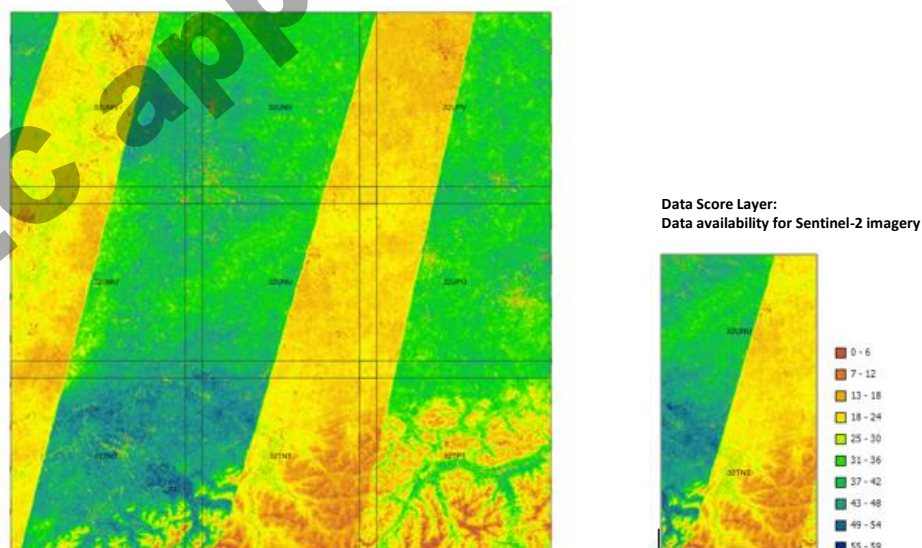


Figure 3-111: Phase 2 - Sentinel-2 data score (inverted cloud value count) for ECoLaSS central test site (left: whole demo site, right: test site T32UNU/TNT tiles) for the time period Mid-March – Mid-Oct 2018.

The Sentinel-1 Ground Range Detected (GRD) data (VV and VH polarization) were pre-processed to Gamma0 values and a multi-temporal filter was applied on the time series (further details are found

in [AD07]. The pre-processing was done using the ESA SNAP toolbox. In contrast to phase 1, in phase 2, both ascending orbit 15 and descending orbit 168 with a minimum tile coverage of 20% (referring to Sentinel-2-tile) have been used to get a denser time series.

Figure 3-112 describes the monthly data availability of Sentinel-2 and Sentinel-1 data for the two test tiles in phase 1. 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% including an intentional data gap between December 2016 and March. 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 for the Sentinel-2 data, since Sentinel-2B data is only available starting from July 2017, meaning that there is a limited availability of Sentinel-2 scenes from October 2016 to June 2017.

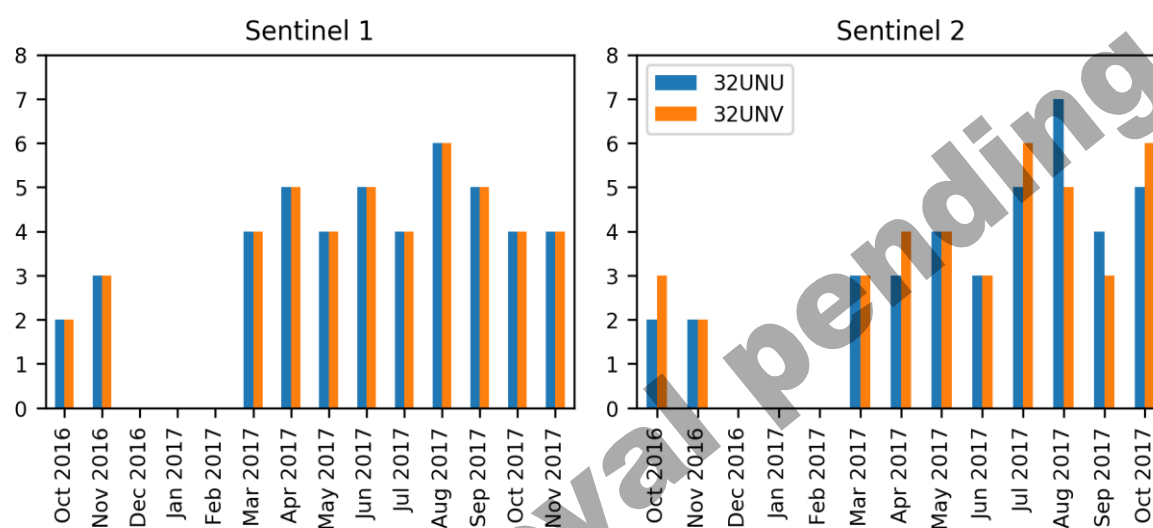


Figure 3-112: Phase 1 - Imagery used in phase 1: monthly data availability for the two test tiles of Sentinel-1 (left) and Sentinel-2 (right) with cloud cover <50%.

The winter season has been left out due to the fact that in most middle European regions snow and ice cover as well as low temperatures induce a vegetation rest, providing only little information suitable for crop area and crop type identification. However, the autumn season of the previous year has been included in order both compensate the data scarcity of the 2017 year and to capture the extended vegetation period of some crops and also the resting period.

As an outcome of phase 1, the late autumn's period isn't included anymore: testing revealed only little benefit in doing so. Instead, the approach of phase 2 focused on the main vegetation period during spring and autumn where plants show highest vitality. Taking into account the highly heterogeneous growing periods of the different crops – plus the shifts in sprouting, growing, vegetation peaks and withering of winter vs, summer cereals –, the classification tried to cover the whole growing period starting in Mid-March and ending up in Mid-October. During this time window, EO imagery offers most information on spectral characteristics and texture, basing on varying phenology and biophysical aspects.

Both Sentinel-1 A + B and Sentinel-2 A + B being fully operational, the available number of scenes has strongly increased in phase 2. However, the limiting factor in 2018 being the high cloud cover, the total number of Sentinel-1 imagery by far exceeds the total number of Sentinel-2 imagery (Figure 3-115).

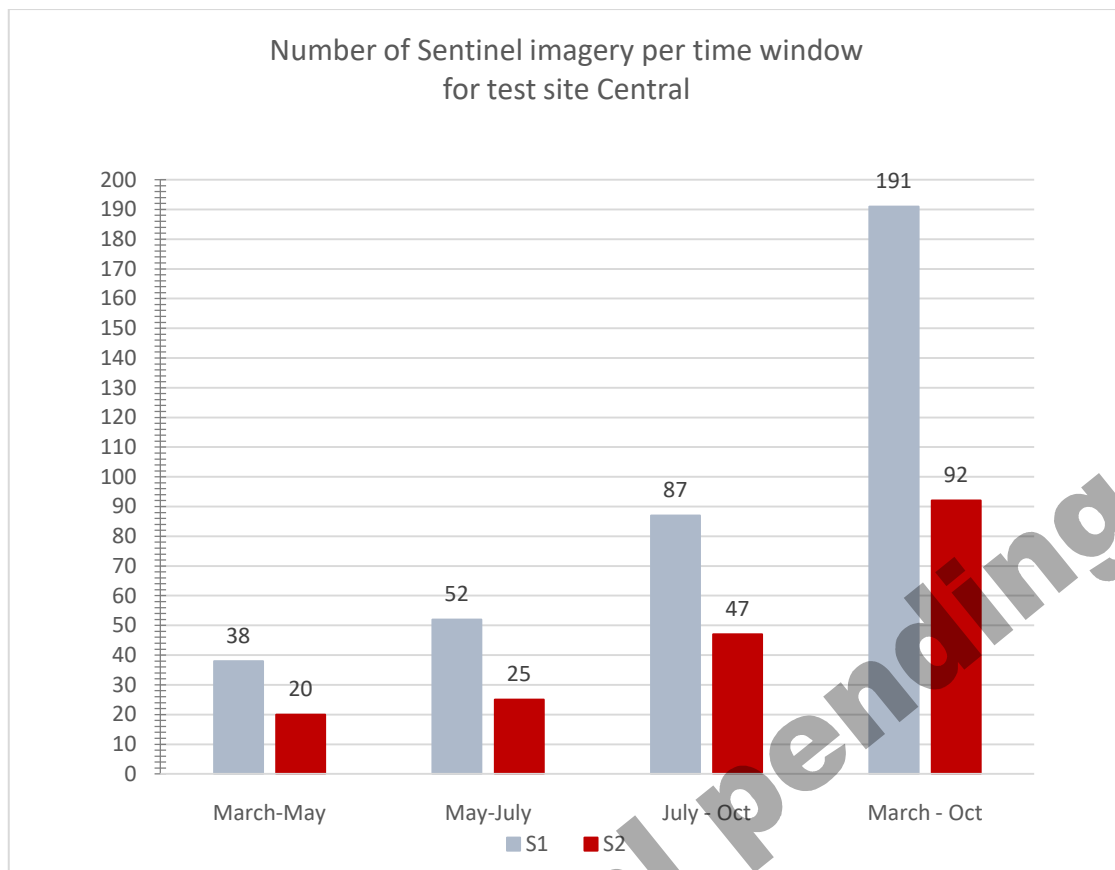


Figure 3-113: Phase 2 - Available imagery for S-1 and S-2: number of scenes per time window for the test site (32UNU + 32TNT) with cloud cover up to 90% (S-2 data) and minimum tile coverage of 20% (S-1)

As the vegetation period strongly depends on temperature, precipitation and altitude within a certain region, the approach needs to be adapted. Differing conditions have been observed still in the small area of the test site and this applies all the more for other regions in the Pan-European perspective. Mediterranean regions for instance, which are characterized by predominantly temperate climates with rainy a mild winter months and drought events in summer, could benefit from using imagery from December to June whereas Northern regions, such as the Scandinavian areas might require a reduced period of April to September to get suitable information on crop vegetation.

TIME FEATURES

For the crop mask and crop type classification, all time features described in 3.1.4 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 in phase 1. Table 3-61 gives an overview of the number of features for the different sensors and periods.

Table 3-61: Phase 1 - Overview of the number of features for the different sensor and period combinations in phase 1

Sensor(s) Name	Period	Sentinel-1	Sentinel-2	Sentinel-1&Sentinel-2
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

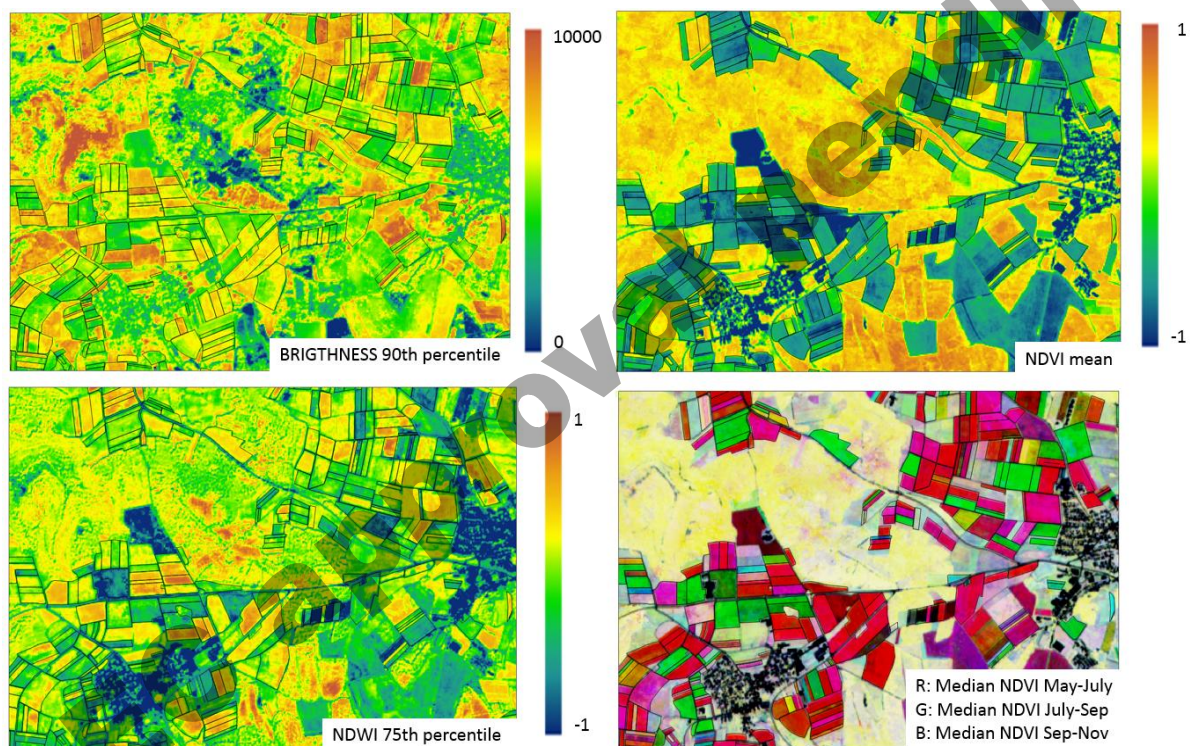


Figure 3-114: Phase 1 - Exemplary selected time features from the Mar-Nov 2017 period (brightness 90th percentile, NDVI mean, NDWI 75th percentile) and an RGB composite of different two-month periods.

In phase 1 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 3-62 shows the number of features used in these two settings.

Table 3-62: Phase 1 - Comparison of the number of features when excluding and including the October and November 2016 data.

Sensor(s) Dataset	Sentinel-1	Sentinel-2	Sentinel-1&Sentinel-2
2016 excluded	92	95	187
2016 included	160	166	326

In order to analyse the growing season, 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 geospatial 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. For the two shorter periods the same features as in case of 201703m9 were 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 3-63: Phase 1 - 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

Considering the experiences in the other tests sites and demos in phase 1, and also the outcomes in WP41, in phase 2 the time windows focused now on the growing season, since it turned out that within this time slot all relevant information on phenology, texture, crop vitality and development could be gathered. This is – as already mentioned – the case for the middle European conditions at the test site and has to be adapted in regions showing different climate conditions. Several tests were carried out for changing time windows: the early winter crop mapping and the sprouting of summer crops (15 March-14 May), the period when all crops are in place respectively growing and peak of vitality for most crops (15 May-14 July), the time for decrease in vitality, withering of crops and harvesting (15 July -14 Oct) and the extensions to cover the whole main growing season (15 March-14 October) including the harvest period. The time windows definitely vary locally. The slightly changed time windows (mid-month instead of beginning of month and 3-months period

instead of 2-months period) derives from experiences from phase 1. Limited time windows focusing on main vegetation phases are an ideal trade-off between maximum of information with minimum data input this should help to save processing time and costs but at the same time get high accuracy.

A list of all calculated time features is given in Table 3-64 and Table 3-65, and an overview over the number of features created per period and sensor is given in Table 3-66 (notation as following: Mid-March to Mid-May 2018 is referred to as 201803m2 which stands for the start year and month and the total number of months comprised by the time period).

Table 3-64: Phase 2 - Overview over calculated features per band and index for Sentinel-1

sensor	S-1	bands + indices	VV VH NDVVH RATIOVVH	features	covariance (cov) standard deviation (std) mean min max median p010 p025 p050 p075 p090 pdiff075025 pdiff090010

Table 3-65: Phase 2 - Overview over calculated features per band and index for Sentinel-2 in phase 2

sensor	S-2	bands + indices	band 03 band 04 band 08 band 11 band 12 NDVI NDWI IRECI BRIGHTNESS	features	covariance (cov) standard deviation (std) mean min max median p010 p025 p050 p075 p090 pdiff075025 pdiff090010

Table 3-66: Phase 2 - Total number of time features per sensor and per time period

Sensor Name	Period	Sentinel-1	Sentinel-2	Sentinel-1&Sentinel-2
201803m3	Mid-March to Mid-May 2018	52	117	169
201805m3	Mid-May to Mid-July 2018	52	117	169
201807m4	Mid-July to Mid-Oct 2018	52	117	169
201803m8	Mid-March to Mid-Oct 2018	52	117	169
SUM		208	468	676

Within the Central region, some differences appear due to altitude and micro climate characteristics. In this sense, stratification is recommended. The following images of some exemplarily selected time features point out the capability of the time feature approach to capture the changing vegetation cover in the selected growing phases. The benchmarking results of the research questions are based on these features.

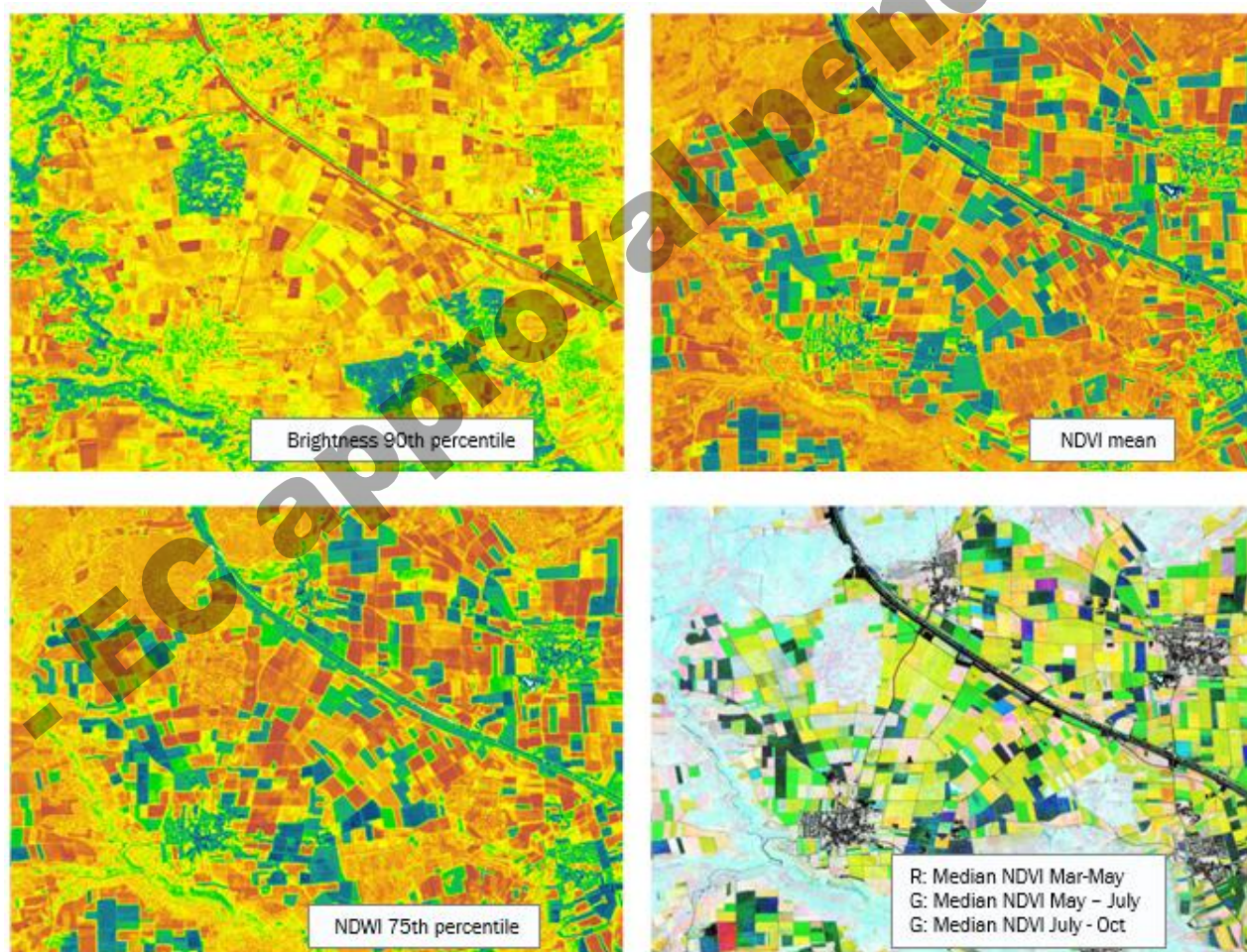


Figure 3-115: Phase 2- Exemplary selected time features from the Mid-March - Mid-Oct 2018 period (brightness 90th percentile, NDVI mean, NDWI 75th percentile) and an RGB composite of 3 different periods.

Feature Selection

The bands and indices given in Table 3-64 and Table 3-65 could be seen as pool of information which is in general a very good and proven basis for vegetation classification and highly suitable for crop differentiation. However, the number of useful features and the ranking differs from region to region.

Since the processing cost increases with the number of features it is desirable to reduce the number of features without sacrificing accuracy. A feature selection algorithm can be applied in the classification workflow as described above to serve this purpose. According to this workflow, a large set of features is generated for the training samples polygons. Basing on this information, a subset of most informative features can be selected for the final classification model. Only the selected features then need to be computed on the complete raster data. As for the method of selecting the most suitable features, two different methods have been tested in phase 1 respectively phase 2. A recursive feature selection (Guyon et al. 2002) was tested for selecting a subset of features for benchmarking in phase 1. 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.

The outcomes of phase 1 and the testing experiences in Task 3 for agriculture in Central contributed to the definition of the classification final parameters and confirmed suitable temporal windows and features that were performing best. In that regard, the Random Forest classification algorithm which is in use in the ECoLaSS project provides information about the importance of certain features for the respective classification. For agriculture and grasslands, and some of the forest products, the Grouped Forward Feature Selection method has been applied in phase 2. This feature selection method - adapted and embedded in the Random Forest classification process - is based on the sequential feature selector which is integrated in the machine learning package (python module scikit-learn in the machine learning extension MLxtend). It selects the most suitable time features by using the information input of the training samples, builds a classification model and provides a subset of features which then can be used for the roll-out classification process on the whole raster.

The method involves reducing an initial d-dimensional feature space to a k-dimensional feature subspace where $k < d$. Generally, feature selection aims at two aspects: improving the computational efficiency and reducing the generalization error of the model by removing irrelevant features or noise. The sequential feature selector removes or adds one feature at the time based on the classifier performance until a feature subset of the desired size k is reached. The Recursive Feature Elimination, method might be less complex in a computational effort perspective, using the feature weight coefficients (e.g., linear models) or feature importance (tree-based algorithms) to eliminate features recursively. However, the advantage of the Forward Feature Selection is that it eliminates or adds features based on a user-defined classifier/regression performance metric. The algorithm finally results in a specific combination of the features guaranteeing the potentially highest accuracy and seems to be the ideal feature selection method for crop classification. The selected number of features still depends on the complexity of the classification aim and on the number of classes. As crop classification needs a lot of detailed input for an accurate classification, even the reduced number of best-off features is still high.

REFERENCE SAMPLES

In phase 1, LPIS data have been used as reference basis for both, Crop Mask and Crop Type Mask. 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 within the test site covers grassland which is not examined in the present analysis. For the study, only major crop types were identified and

grouped into several categories. Table 3-67 lists these relevant crops by category and number of available reference parcels.

Table 3-67: Phase 1 - Overview of the type and number of reference parcels used for crop type classification.
Crop code and crop name derived from LPIS

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 3-116 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 3-116 right).

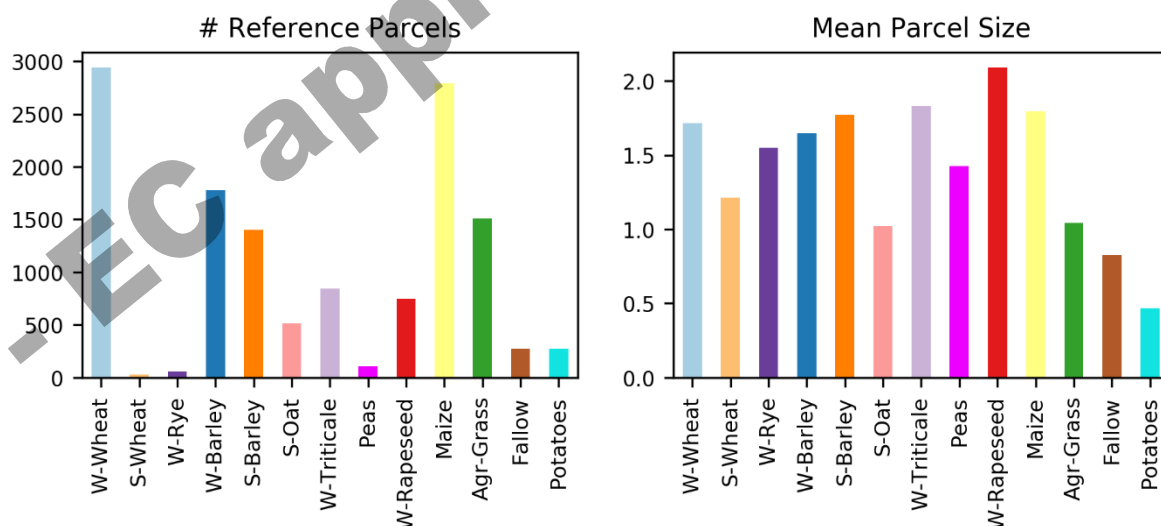


Figure 3-116: Phase 1- Frequency (left) and mean parcel size (right) of the reference samples used for crop type classification.

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 of the LPIS reference data (proportion = 7:3) to ensure unambiguous reference data.

For the crop mask classification, additional reference samples that cover the basic land cover 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 3-68 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 3-68: Phase 1 - 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

- EC approval pending -

In phase 2, two different input data sets have been used: The reference data for the crop mask consist of LUCAS 2018 data for the crop classes and some non-crop classes, complemented by samples from the HR Layer 2015 (which includes partially data from 2017) covering the non-crop classes. It must be noted that LPIS dataset was not available in Switzerland nor in the Northern part of the Central region. As the number of cropland samples derived from LUCAS was still very low, additional manual samples have been added for certain classes such as orchards or vineyards. Table 3-69 gives the LUCAS classes which have been used as basis for the sampling layer for the Crop Mask.

Table 3-69: Phase 2 – Overview of the LUCAS classes that were used as sample base for classes grassland, forest, imperviousness and water bodies.

Classes used for cropland samples	GRA	FOR	IMD	WaW
B11 Common wheat B12 Durum wheat B13 Barley B14 Rye B15 Oats B16 Maize B17 Rice B18 Triticale B19 Other cereals B21 Potatoes B22 Sugar beet B23 Other root crops B31 Sunflower B32 Rape and turnip rape B33 Soya B35 Other fibre and oleaginous crops B36 Tobacco B37 Other non-permanent industrial crops B41 Dry pulses B42 Tomatoes B43 Other fresh vegetables B44 Floriculture and ornamental plants B51 clover B52 lucerne B53 other leguminous+mixed fodder B70 Fruit trees B71 B75 berries	E10 Grassland with sparse tree/shrub cover E20 E10 Grassland without tree/shrub cover	C10 broadleaved C20 Coniferous C21 spruce dominated coniferous woodland C31 spruce dominated mixed woodland C33 other mixed woodland CXX3 Alpine Forests	A10 Roofed built-up areas A22 non built-up linear features	G10 inland water bodies G20 inland running water

Table 3-70 shows the number of samples for each class for the crop mask classification. Splitting the dataset into training- and test set was performed for crop mask and crop type classification, meaning that approx. 50% were used for training and 50% - after several test it turned out that this percentage showed the highest accuracy in phase 2 (different from phase 1 where the 70:30 percentage showed best results).

Table 3-70: Phase 2 - Overview of the reference samples used for the crop mask classification

Class code	Classname	# of samples	Source
0	broadleaved trees	1800	HRL 2015 LUCAS2018
0	coniferous trees	2583	HRL 2015 LUCAS 2018
0	imperviousess	1800	HRL 2015 LUCAS 2018
0	waterbodies	1206	HRL 2015 LUCAS 2018
0	grassland	1800	HRL 2015 LUCAS 2018
1	cropland	2664	LUCAS 2018 Manual sampling

The cropland area of the crop mask has to be as accurate and comprehensive as possible because it will be the basis for the crop type mask. In this sense, the quality of the sampling is essential: issues in the crop mask are propagated in the crop type map within the crops. One first conclusion is that it is very important that the crop mask accuracy is very high to achieve a satisfactory crop type map. The main difference between the Crop Mask and the Crop Type Mask in this respect is, that in general agriculturally used areas remain quite stable whereas at the same time the crop management on these areas is highly variable due to crop rotation in all farming managements systems. The sample base for the identification of cropland vs. Non-cropland therefore can be more general and tolerates impreciseness concerning crop type differentiation and lacking timeliness. In the end the information of cropland or non-cropland is needed and LUCAS data are very suitable for that purpose. In contrast, Crop Type mapping strongly depends on highly accurate reference data to provide accurate results.

Even though the LUCAS data is not used for crop type differentiation due to the lack of detail, validity and timeliness in the attributes regarding the selected crop types for classification, the class structure of the crop type mask finally is oriented towards the LUCAS class structure (aiming at the potential of LUCAS data being a source of information available in most EEA countries in the future). The selected crop type structure comprises the most common crop types for the crop groups of winter and summer cereals, vegetables, dry pulses and legumes, industrial crops, root/tuber crops, fodder crops and permanent crops.

From the test experiences, it became clear that it was necessary to add manual sampling for the crop mask because LUCAS points are insufficient partially for location, partially for not covering specific crop types or for not being representative for the actual crop type in the respective region [AD15]. Sample representatives for orchards and vineyards that are also mixed with grasslands and forest in not pure crop samples result in mixed pixels. Spectral unmixing could potentially be applied, although this was not explored further in order to not compromise cost-efficiency for larger areas developments. Instead, manual samples for those classes have been added which improved the classification but unfortunately hasn't solved the issue. It was also proved that shape matters (e.g., trees in line that are not similar to other orchards with different patterns). In this case, the classifier is often not able to make reliable decisions and opts for grassland instead of orchards (omission error), which should not be part of the Crop Mask.

A known issue is that grassland, fodder crops/temporary grassland and crop types which by nature show also high percentage of grassland are difficult to differentiate and cause misclassifications. The confusion matrix of the Crop Type Mask gives more information about how grassland and crops mix with many other classes. One option could be to use a higher sampling weight for those classes which are highly affected by commissions caused by other crop types. Using the probabilities could be another potential solution for these issues as the underrepresented crop areas tend to show probabilities between 40-50%, whereas the classifier decides for a specific crop type starting with probabilities of 51%. Further tests will be needed to find out if this would be a legitimate option even for a larger region.

Crop Type nomenclature

The crop type nomenclature agreed upon for testing in phase 2, is partially based on the outcome from phase 1 tests and the demos implementation but also from an anew analysis of spectral characteristics, specific growing phases, phenology of plants and taking into account their distribution respectively significance within the region. The crop type nomenclature is a prototype that considers a potential Pan-European roll-out. As experienced in other projects dealing with crop type mapping, and as it is always the case in land cover classifications, hierarchy, definitions and nomenclature are not trivial matters. That is why the data model and standard definitions are so relevant for global mapping and for enabling product updates and comparisons. In particular, occurrence, phenology, vegetation period and farming system of crop types show high variability. Thus, a hierarchical legend is a good compromise to be able to define a common legend at larger scales (e.g., Pan-European level), while matching with local conditions in subsequent levels by class aggregation or disaggregation. The class structure in ECoLaSS takes into account the degree of prevalence and extent of the cultivated area for the specific crop type (e.g., considering the crops more representative in terms of surfaces in Europe, cross-checking with available LPIS and LUCAS datasets and other sources available), the potential of spectral separability (crop-specific, and directly linked to the workflows using EO time series data inputs), phenology and growing cycles. The class structure is oriented to some extent to the LUCAS land cover structure of vegetation with crop groups at the highest level, crop classes at the 2nd level and crop types at 3rd level. In the case of the Central demo site, rice and olive groves, though relevant at the Pan-European scale, are left out. This is an example on how the legend must be adapted in the end to the local conditions.

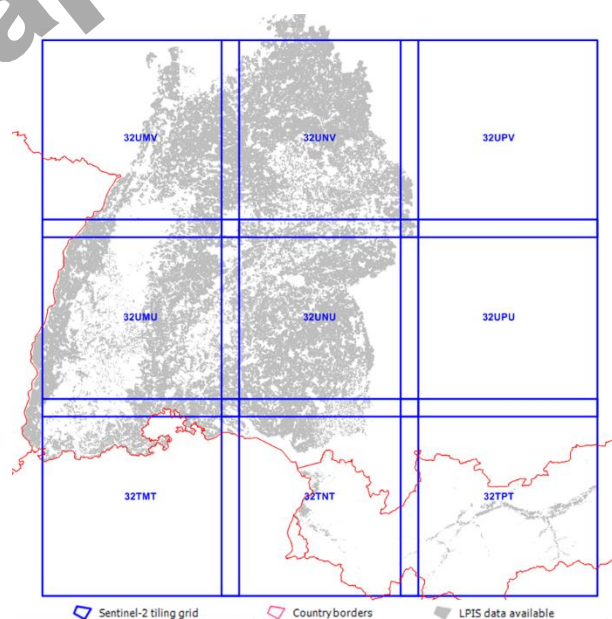


Figure 3-117: Phase 2 - available LPIS data for the 2 test tiles 32UNU and 32TNT for 2018

In contrast to phase 1, in phase 2 the class structure showed a higher number of crop types – oriented to the LUCAS class system and hierarchically structured to be customized to regional conditions (see Table 3-71). The higher number of classes is part of the lessons learned in phase 1 where it became clear that a classification of single crops and a later aggregation to crop groups would provide better classification results.

The LUCAS survey identified certain crop groups and crop classes which are supposed to be relevant in all European countries. As these crop groups and crop classes are part of the Pan-European survey and proved to work under different conditions in many European regions, it makes highly sense to keep these crop groups and - as far as possible - also the crop classes for the crop type classification. The idea is to then summarize characteristic regional crop types under these crop classes referring to as regional adaptation.

Samples for the crop types consist of sample extraction from LPIS data from Baden Württemberg and InVeKoS data from Austria (referring to the test site tiles 32TNT and 32UNU). They have been customized to the crop class/crop type structure implemented in phase 2 which means that only distinct samples which are representative for a crop type, have been chosen. Aiming at a sufficient sample basis for the high number of crop type classes, 250 polygons (if available) have been randomly chosen as input for the Feature selection, the model building and the subsequent classification. Some crop types have been declared as substantial for the class structure but are rarely cultivated within the test site, therefore less than 250 polygons have been available. This is the case for class 4-winter oats, class 7-summer rye and class 13-legumes.

The hierarchical structure of fixed crop groups on the first level, largely fixed crop classes on the second level and regionally varying crop types which has been developed in phase 2 for the crop type mapping could be seen in the following table:

- EC approval pending

Table 3-71: Phase 2 - ECoLaSS crop type nomenclature in phase 2 revealing the hierarchical structure to be customized to regional or local characteristics.

Land cover mask	Crop Group	ECOLaSS Paneuropean Crop class	Reference Data for test site Central				ECOLaSS Crop type classes for CTM	8 sample excluding small areas <0.1ha				
			LPIS BaWü		Austria InVekos							
Crop mask	1	Cereals	11	Winter cereals	111	winter wheat	115 - Winterweizen 112 - Durum/Winterhartweizen	WINTERWEIZEN	1	31369		
					112	winter barley	111 - Wintergerste	WINTERGERSTE	2	17369		
					113	winter rye	121 - Winterroggen	WINTERROGGEN	3	870		
					114	winter oats	112 - Winterhafer	WINTERHAFER	4	53		
			12	Spring/summer cereals	121	summer whe	115 - Sommerweizen 113 - Durum/Sommerhartweizen	SOMMERWEIZEN SOMMERHARTWEIZEN (DURUM)	5	665		
					122	summer barley	112 - Sommergerste	SOMMERGERSTE	6	10141		
					123	summer rye	122 - Sommerroggen	SOMMERROGGEN	7	4		
					124	summer oats	113 - Sommerhafer	SOMMERHAFER	8	5313		
			13	Maize	131	Maize	171 - Silomais 171 - Körnermais/CCM 172 - Mais (Biotas) 313 - Saatmais 174 - Zuckermais	KORNERMAIS KORNERMAIS GRÜNMALIS	9	38785		
			14	Rice					999			
				2	Vegetables, dry pulses, berries	21	Vegetables	211	vegetables	211 - Gemüse FELDGEMÜSE ERNÄHRL. TYP FELDGEMÜSE MEHRFAK. TYP FELDGEMÜSE VERARBEITUNG MEHRFAK. TYP FELDGEMÜSE VERARBEITUNG ERNÄHRL. TYP SPESSETURBS OLIVEN		10
			22							dry pulses/legumes	221	peas and beans
						222	lentils	222 - Linzen (Speise-Linse)				
	223	legumes				223 - Lupinen 223 - Espasette		13	52			
	224	legumes				224 - Sojabohnen		14	422			
	41	Soybeans	411			soya beans	411 - Sojabohnen		15	76		
				42	sunflowers		420 - Sonnenblumen	SOMMELNEN	16	6160		
	43	Rapeseed	431	rape seed	431 - Winteraps 432 - Sommeraps							
					44	Other oleaginous, fibre, biofuel, and beverage crops	441 - Ölsamige/Faserpflanzen 442 - Mais 443 - Getreide	ELEFANTENGRA	17	1803		
	5	Root/tuber crops	51	Potatoes	511	potatoes	602 - Speisekartoffeln 601 - Stärkekartoffeln 603 - Pflanzkartoffeln	SPESIELARTOFFEL FRÜHARTOFFEL	18	1516		
							52	beet crops	603 - Zuckerrüben 413 - Futterrüben (Runkelrüben)	FUTTERRÜBEN/RUNKELRÜBEN	19	418
	6	Fodder crops	61	Temporary grasslands (<5 y.) - fodder/agrarian grass	611	temorary grassland	441 - Wiesen (Grünlandneueinsaat weniger als 5 Jahre zurücklegend) 442 - Mahwiesen (Grünlandneueinsaat weniger als 5 Jahre zurücklegend) 443 - Weiden (Grünlandneueinsaat weniger als 5 Jahre zurücklegend)	MAHWIESEN-WEIDE (DREI UND MEHR NUTZUNGEN)	20	38042		
							612	agrarian grass	425 - Klee-Luzerne-Gemisch	LUCERNE		
									423 - Luzerne, Hofenklapp, Gelbklee, Bastardluzerne/Sandluzerne	KLEE		
									421 - Flor-/Voll-/Alexandrier-/Kleiner-/Erd-/Schweden-/Fischer Klee	KLEEGRAS		
									422 - Klee, Luzerne-Gras-Gemenge			
							613	agrarian grass	424 - Ackergras 432 - Klee Mischung aus 421, 422, 431			
7	Permanent crops	71	Grape vines	711	vine growing	643 - Bestockte Rebfläche 643 - Tafeltrauben	WEIN TAFELWEIN	21	1074			
						72	Olives groves		999			
						73	Other permanent crops	731	fruit trees/orchards	602 - Kern- und Steinobst 603 - Kirschen	ZWETSCHKEN TAFELAPFELN TAFELAPFEL KIRSCHEN	22

A class structure with a number of classes of this range seems reasonable for Central. It is quite generic in view of a homogeneous pan-European crop type map production, while at the same time reflects the regional situation on the test site. Tricky crop groups identified from the tests are Crop Group 2-*vegetables, dry pulses, legumes* as well as partially Crop Group 4-*industrial crops* (especially the crop class 44 – *Other oleaginous fibre, biofuel, and beverage crops*). These groups are characterized by small parcels, locally heterogeneous phenology and local occurrence) which is verified in the demo site as well. Our findings suggest that it is very important that the reference data structure is aligned to phenology and spectral characteristics in order to make it possible to distinguish between crop types in the region. Season peaks and indicators must be homogeneous within one crop group. Taking this into account, dubious samples must be left out which reduces the number of suitable samples. Another issue when dealing with several databases is that consistency and definitions and nomenclature homogeneity are not guaranteed. Both issues could lead to either a highly reduced sample base and/or unwanted heterogeneity of samples within one crop class/crop type both of which is strongly affecting the accuracy of the subsequent classification. Without a Pan European standard concerning survey system and naming convention to name some, these issues cannot be solved.

A specific issue is how to deal with minority classes (Table 3-71). Class 4-*winter oats, 7-summer barley* and 13-*legumes* show a very low number of parcels and at the same time very small parcels (Figure 3-118), subsequently only a small number of samples could randomly be extracted. Since this caused misclassifications, they have been left out.

One option to deal with minority classes could be to use the SMOTE approach in order to get a well-balanced sampling set for classification, respectively to get a higher sampling base for underrepresented classes and thus to improve the accuracy for those crop types. The Synthetic Minority Over-sampling Technique, uses synthetically generated samples along the line joining the minority samples and its 'k' minority class farthest neighbors in order to obtain a relatively balanced dataset for higher accuracy of underrepresented classes. Another approach has been used here: instead of artificially increasing the sample set for very small classes, as it would be the case with SMOTE - which in fact are hardly represented in specific areas - those classes have been left out. In the test site these left-out classes of 4-*winter oats, 7-summer barley* and 13-*legumes* with very low number of and at the same time very small parcel sizes. Polygons under 1 ha have therefore been left out to limit unclear information caused by spectrally mixed pixels.

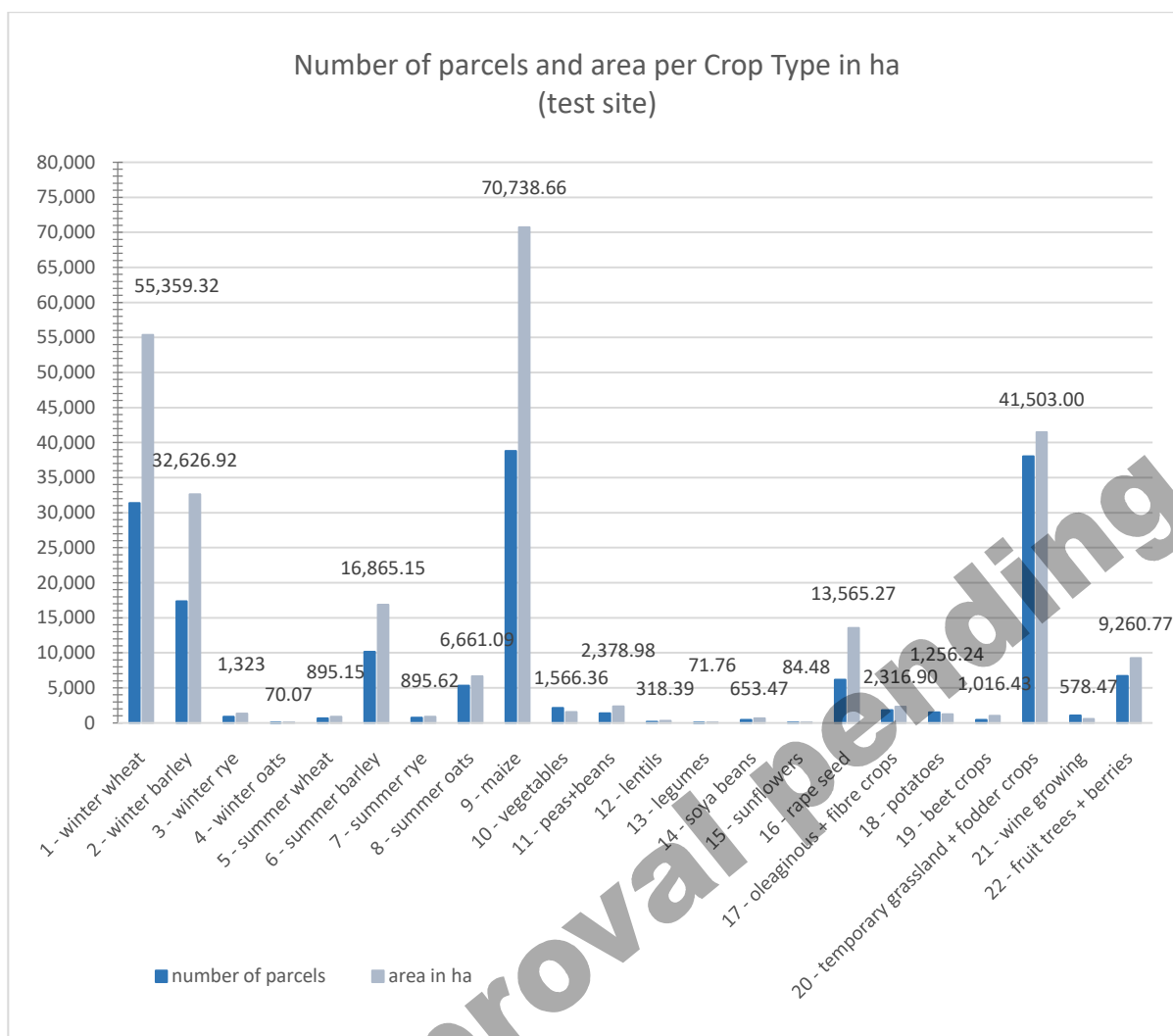


Figure 3-118: Number of parcels per crop Type and total area per Crop Type in ha, indicating the average parcel size as well as the degree of representation within the test site

This approach is consistent, since the classes *rice* and *olive groves* are left out as well where they do not occur. This, in fact, is a very fundamental issue, since there will be hardly any crop type that occurs in every European country.

However, it must be said that the low number of samples is not the only determining aspect for keeping or leaving out crop types. Spectral characteristics, phenology and size of parcels should also be taken into account. Sunflowers for example show also a minor number of samples but are very distinct in spectral information and are cultivated in both, smaller and larger parcels (Figure 3-119). Thus, they can be well differentiated and detected.

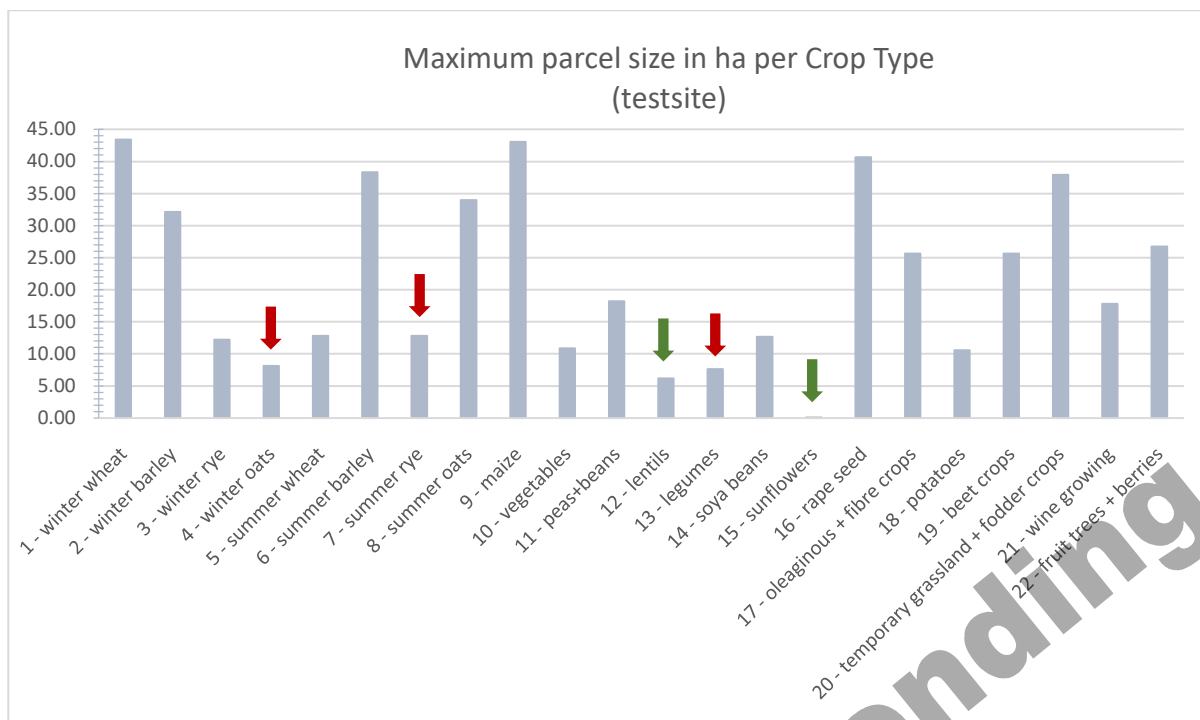


Figure 3-119: Phase 2 - Maximum parcel size per Crop Type in ha; classes marked in red have been left out due to small parcel sizes, the green ones have been kept – they offer distinct spectral information and can be well classified.

CLASSIFICATION AND RELIABILITY LAYERS

A Random Forest Classifier was used for all classifications both for phase 1 and phase 2 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, although the largest probability is the one provided together with the prototypes delivered:

- largest probability (maximum probability)
- largest probability - second largest probability (breaking ties) (Luo et al. 2015)
- entropy - $\sum_{c=1, \dots, \text{#Classes}} (p_c \log p_c)$ 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.

3.3.4.1.4 Classification Results and Validation

In the first place, for the crop mask, the multi-sensor approach and the pixel/object level were assessed. In phase 1, the results were generated by using data from reference year 2017 plus data from autumn months of 2016, whereas in phase 2 an adequate time window of Mid-March to Mid-Oct 2018 has been chosen in order to cover all crucial growing stages of both winter and summer crops. In the Central site the classification accuracies based on Sentinel-1 data are significantly lower than the respective accuracies based on Sentinel-2 data (OA of 89.5% for Sentinel-1 vs. OA of 92.3% for Sentinel-2 on pixel level). This was the case in phase 1 and has been approved also in phase 2, as it is commonly the case in various agriculture studies elsewhere. The combination of the two sensors does not significantly improve the classification accuracy when compared with the accuracies using simply the Sentinel-2 data (Table 3-72 and Figure 3-120). However, in order to get a sufficient time series density and due to the fact that optical data availability might be limited, the integration of Sentinel-1 imagery is strongly recommended. The contribution of SAR features will indeed play a crucial role in regions/time windows with very high cloud cover.

Table 3-72: Phase 1 - Kappa Coefficient (K) and Overall Accuracy (OA) for the different crop mask experiment setups (Sentinel-1, Sentinel-2, and Sentinel-1 & Sentinel-2 on pixel and field level).

	K * 100	K * 100	OA	OA
	pixel	field	pixel	field
Sentinel-1	68.1	78.8	89.5	89.3
Sentinel-2	79.2	84.9	93.0	92.3
Sentinel-1 & Sentinel-2	80.4	85.2	93.9	92.6

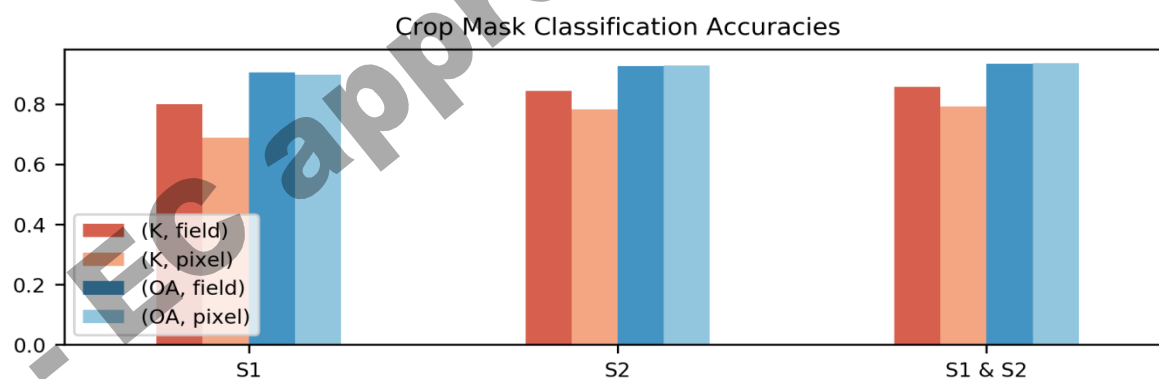


Figure 3-120: Phase 1 - Barplot of Kappa (K) and Overall Accuracy (OA) for the different experiment setups (Sentinel-1, Sentinel-2, Sentinel-1 & Sentinel-2 on field and pixel level).

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, the overall accuracies increase with the field size. This is an expected pattern due to, e.g., the presence of speckle, as for the SAR data, but happens also for optical data due to the naturally uneven growth of the crops on the field. This important finding indicates that - even if the real field polygons are not available, segmentation could be considered, all the more if optical data availability is not sufficient (due to high cloud cover) and SAR data must be used as primary data source.

Considering incremental updates, intra-seasonal monitoring (e.g., see WP41 and user requirements compiled in WP21), and an intended large scale production the pixel based approach is a good compromise. Further examples of field-based approaches are provided in other sites.

Phase 2:

Assessing the multi-sensor approach was still a focal point, and it has been confirmed, that for both, Crop Mask as well as Crop Type Mask, optical data are indispensable for high accuracies (Table 3-73 and Figure 3-121), however, the combined approach offers high potential of complementing the data base where high cloud cover limits the benefit of optical data.

Table 3-73: Phase 2 – Crop Mask: Kappa Coefficient (K) and Overall Accuracy (OA) for the different experiment setups (Sentinel-1, Sentinel-2, and Sentinel-1 & Sentinel-2 (pixel level)).

	K*100	OA
Sentinel-1	64,9	88,3
Sentinel-2	80,8	94,4
Sentinel-1 + & Sentinel-2	81,8	94,4

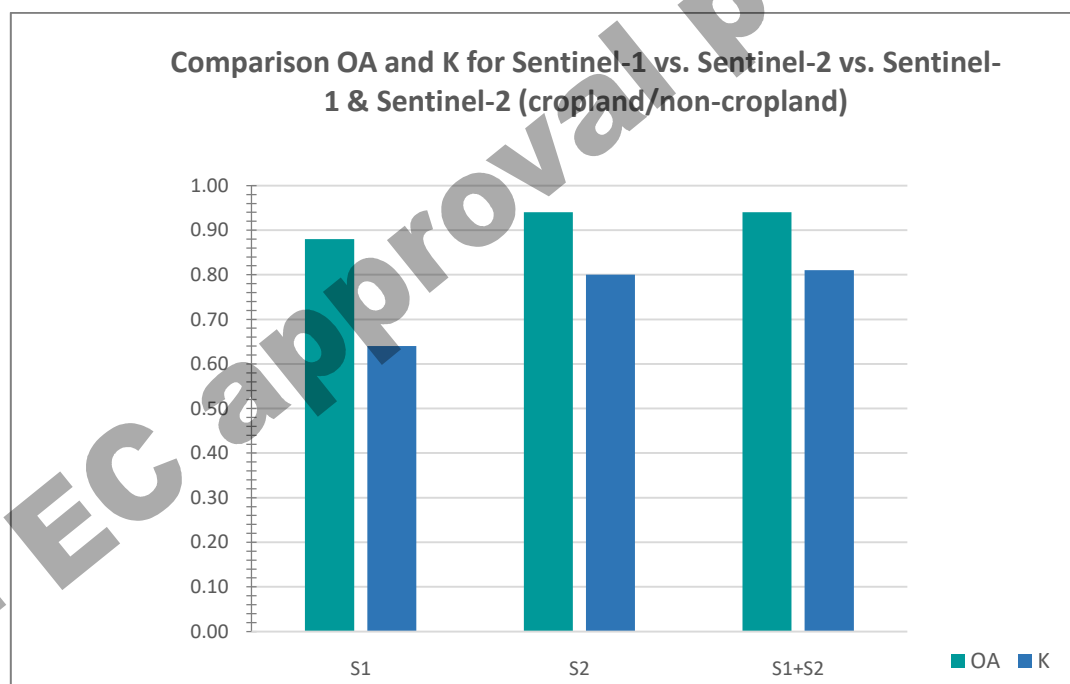


Figure 3-121: Phase 2 – Crop Mask: Overall Accuracy (OA) and Kappa Coefficient (K) for the different experiment setups for Sentinel-1, Sentinel-2, Sentinel-1 & Sentinel-2 (pixel level).

In phase 2, the pixel based approach is still kept for both products for the Central demo site, due to the satisfactory accuracies achieved on the one hand, and the optimal cost-efficiency in production on the other hand. Since the pixel level accuracies versus field level accuracies have already been analyzed in phase 1, this has been left out in phase 2.

As already mentioned, crop mask and crop type classification have been produced with different input data in phase 2: LUCAS points for the crop mask, LPIS data for the crop types (however limited to a

constraint area of Baden-Wurttemberg and Austria for the test site). The main reason to use different input data in phase 2 was the generally limited availability of up-to-date LPIS data in a Pan-European context. Using LUCAS data could be a meaningful alternative for the production of a crop/non-crop layer, as LUCAS data provide large-scale coverage (except only few countries) and, contrary to the crop types, the basic crop areas remain quite stable and in general only show minor changes over short time periods. Potentially outdated sample points can – to a certain extent - be handled by the classifier due to its robustness. So, LUCAS data are a time and cost effective basis for a classification of crop/non-crop areas in larger or even Pan-European context.

However, as for the crop type differentiation, the classifier strongly relies on detailed and up-to-date information on the type of crop as well as on its location. Therefore, this product still depends on correct and up to date information as well as on short-term availability. This indeed is the main limitation of a large-scale field approach for crop type mapping. The following statistics give an impression of the potential of both approaches if full availability would be possible.

Table 3-74 (Phase1) and

Table 3-75 (Phase2) give a summary of the results with regard to the processing cost, accuracy and other benchmarking criteria. The listed sensor and data specific problems could 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 [AD07], which leads also to misclassification in the final raster. These issues have been faced on both testing phases.

Table 3-74: Phase 1 - Benchmarking criteria and specific problems of the different crop mask experiment setups.

	Accuracy (K*100)	Processing Cost	Specific Problems
Sentinel-1 pixel level	68,1	+	Foreshortening, layover in strong relief, speckle
Sentinel-2 pixel level	79,2	+	Clouds/cloud shadows
Sentinel-1 & Sentinel-2 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.
Sentinel-1 field level	78,8	++	Foreshortening, layover in strong relief, segmentation
Sentinel-2 field level	84,9	++	Clouds/cloud shadows, segmentation
Sentinel-1& Sentinel-2 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

Table 3-75: Phase 2 - Benchmarking criteria and specific problems of the different experiment setups.

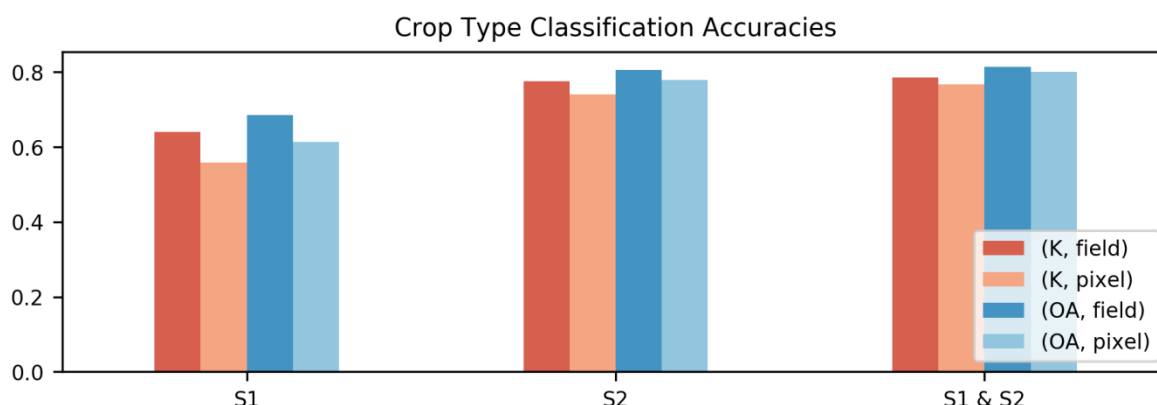
	Accuracy cropmask (K*100)	Processing Cost	Specific issues
Sentinel-1 pixel level	64,9	+	Foreshortening, layover in strong relief, speckle
Sentinel-2 pixel level	80,8	++	Clouds/cloud shadows
Sentinel-1& Sentinel-2 pixel level	81,8	++	As in S1/S2; the strength of one sensor type can compensate for the weaknesses of the other and vice versa.

Crop Type Mask in Phase 1

The patterns of the crop type classification accuracies are similar to those of the crop mask (Table 3-76 and Figure 3-122). In case of the pixel classifications the Sentinel-2 based classification 2017 yields better results than the Sentinel-1 based classification (Kappa*100: + 18.2) while the combination of the two sensors does not improve the Sentinel-2 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 Sentinel-2 based classification, Kappa*100 increases from 74 to 77.5 and for the Sentinel-1/Sentinel-2 based classification there is an increase of 1.9. In case of the Sentinel-1-based classification there is an improvement of Kappa*100 from 55.8 to 64.

Table 3-76: Phase 1 - Kappa Coefficient (K) and Overall Accuracy (OA) for the different experiment setups (Sentinel-1, Sentinel-2, and Sentinel-1 & Sentinel-2 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 3-122: Phase 1 - Barplot of Kappa (K) and Overall Accuracy (OA) for the different experiment setups.

The class-wise F1-Scores (mean of User's and Producer's Accuracy) depicted in Figure 3-123 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 Sentinel-1 as with Sentinel-2 and the field level aggregation does not improve the classification accuracy importantly. In case of maize, the accuracy of the Sentinel-1 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 Sentinel-1, for maize significantly higher accuracies can be achieved with Sentinel-2. 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 3-67 (section 3.3.4.1.3, 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.

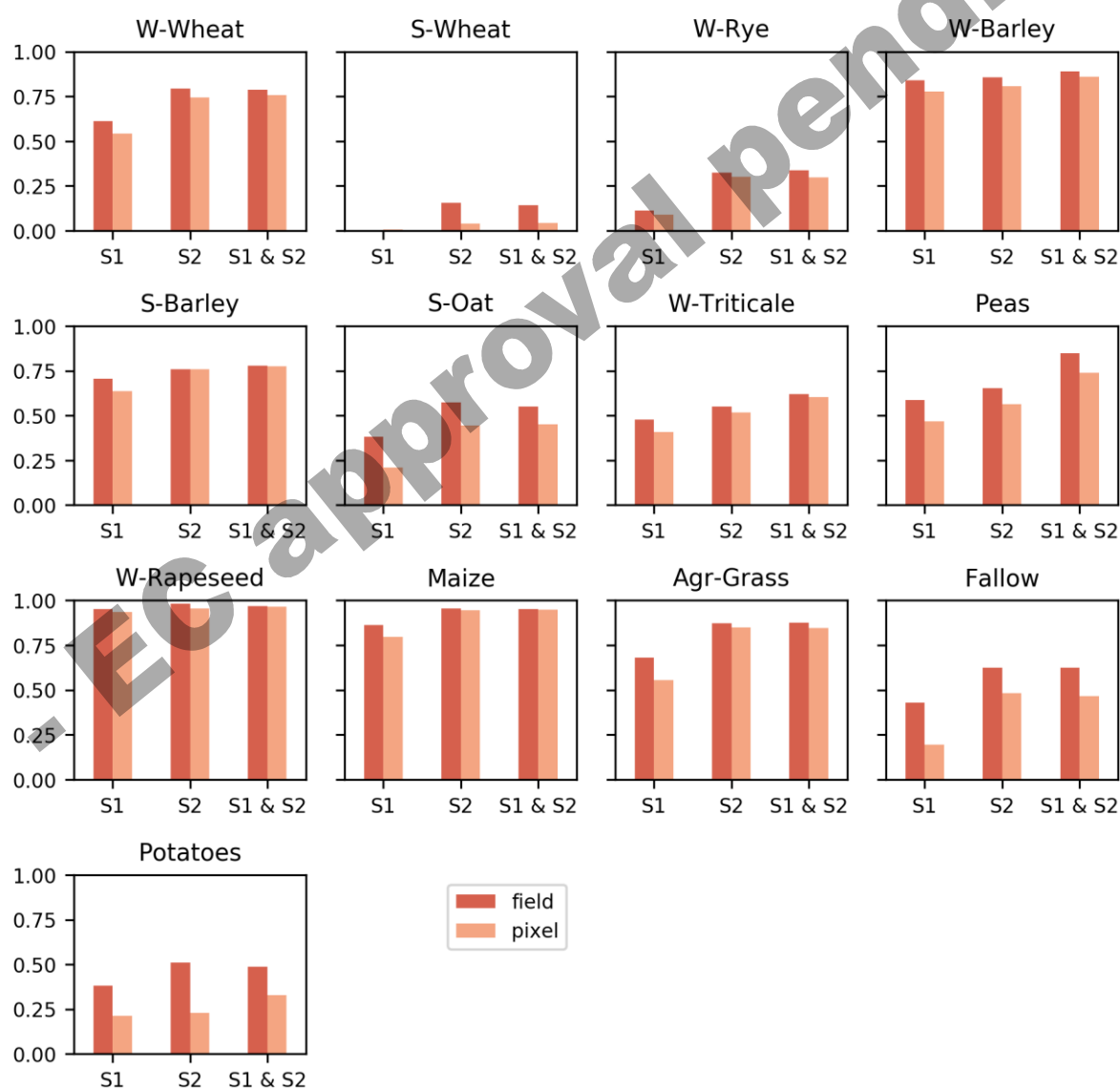


Figure 3-123. Phase 1 - Class-wise F1-Score (mean of User's and Producer's Accuracy) for field vs. pixel-based classifications, reference year 2017.

Figure 3-124 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.

Predicted Label	W-Wheat	S-Wheat	W-Rye	W-Barley	S-Barley	S-Oat	W-Triticale	Peas	W-Rapeseed	Maize	Agr-Grass	Fallow	Potatoes	User's Acc.
W-Wheat	630	0	1	29	4	3	42	0	0	1	5	0	1	87
S-Wheat	2	1	0	0	1	1	0	0	0	0	0	0	0	20
W-Rye	8	0	11	10	0	0	3	0	0	9	5	0	1	23
W-Barley	12	0	1	450	1	1	10	0	0	1	3	0	1	93
S-Barley	40	4	0	2	305	9	2	0	0	0	0	0	1	84
S-Oat	46	2	0	4	100	126	1	0	0	0	7	3	13	41
W-Triticale	121	0	1	28	0	1	183	0	1	0	5	0	0	53
Peas	1	0	0	0	3	0	0	31	0	2	0	1	3	75
W-Rapeseed	7	0	0	0	0	0	2	0	223	0	2	0	2	94
Maize	2	0	0	0	0	2	1	0	0	783	7	5	11	96
Agr-Grass	10	1	3	7	3	6	7	0	1	7	393	2	4	88
Fallow	2	0	0	1	4	6	1	0	0	21	19	66	9	51
Potatoes	1	1	1	2	0	0	0	1	0	12	7	5	36	54
User's Acc.	71	11	61	84	72	81	72	96	99	93	86	80	43	81
	W-Wheat	S-Wheat	W-Rye	W-Barley	S-Barley	S-Oat	W-Triticale	Peas	W-Rapeseed	Maize	Agr-Grass	Fallow	Potatoes	Producer's Acc.
Reference Label														

Figure 3-124: Confusion Matrix of the crop type classification on field level based on the combination of Sentinel-1 and Sentinel-2 time features.

Figure 3-125 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, the conclusion is that it might not be worth to include 2016 data. Therefore, the following experiments were conducted with 2017 data only.

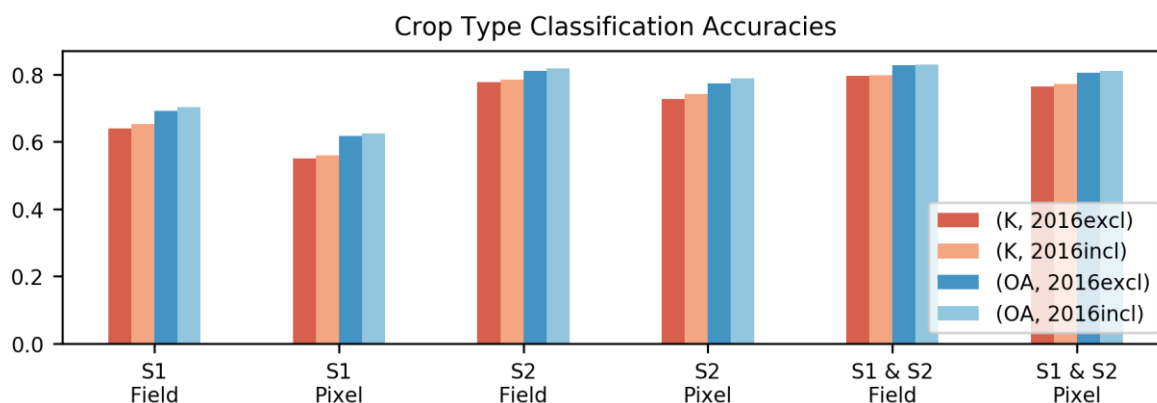


Figure 3-125: Barplot of Kappa (K) and Overall Accuracy (OA) for the different experiment setups.

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 in phase 1. 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.

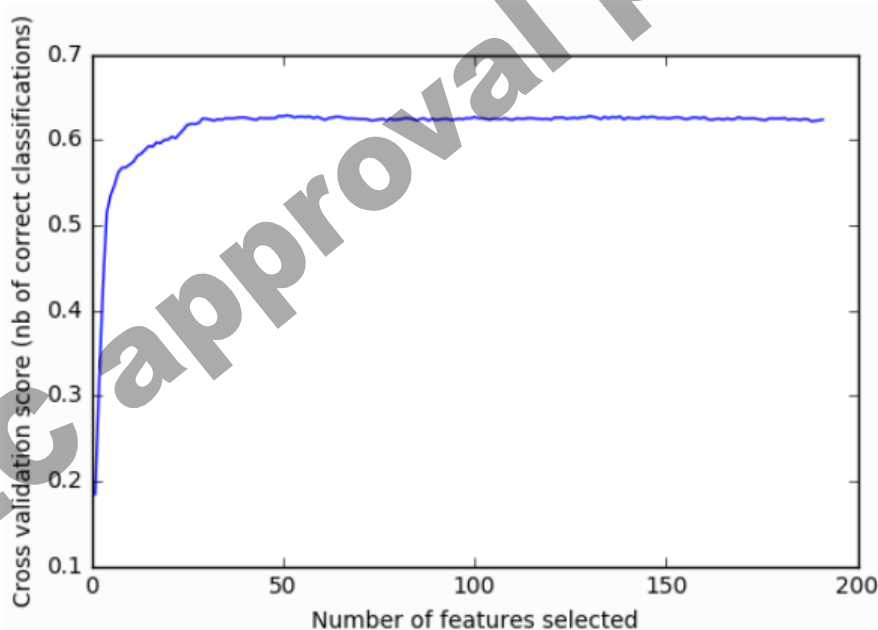


Figure 3-126: Overall accuracy (OA) based on the cross-validated training samples dependent on the number of selected features.

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 3-127). 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. This was confirmed in phase 2.

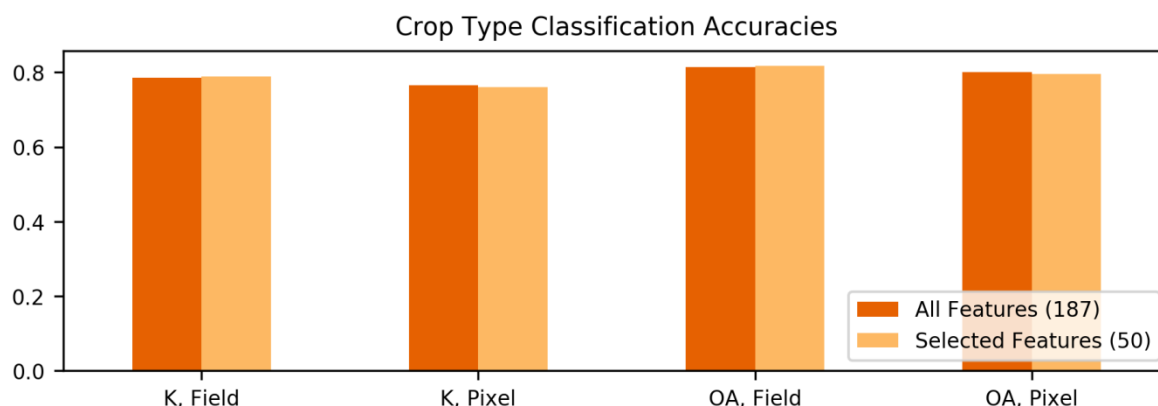


Figure 3-127: Barplot of Kappa (K) and Overall Accuracy (OA) for the classification based on all features, and the 50 selected features.

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: Sentinel-1, Sentinel-2 and Sentinel-1 & Sentinel-2 (Figure 3-128). Instead, the improvement of the classification accuracies between the mid-July and the full period scenario is only marginal.

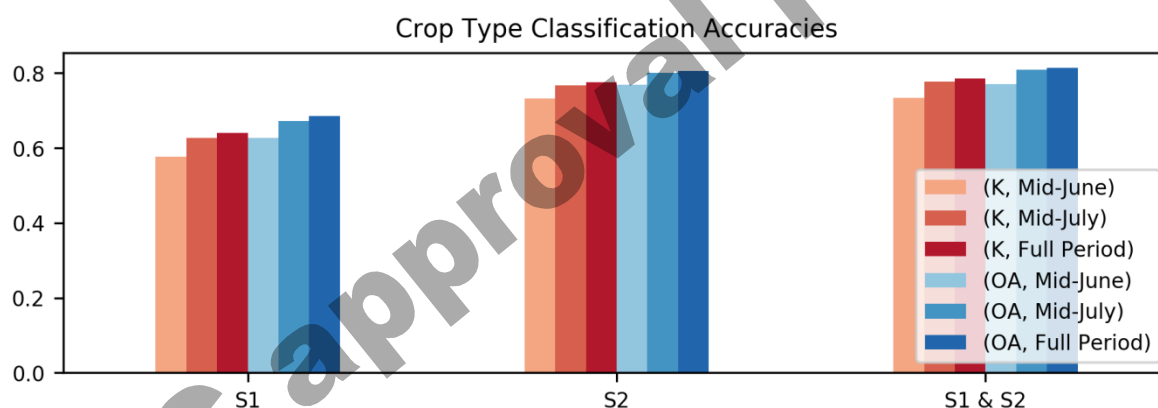


Figure 3-128: Kappa and Overall Accuracy on field level of the for the different experiment setups, particularly the three considered periods.

Looking at the class-wise classification accuracies, Figure 3-129 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 Sentinel-2 data is used. Instead, with Sentinel-1 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.

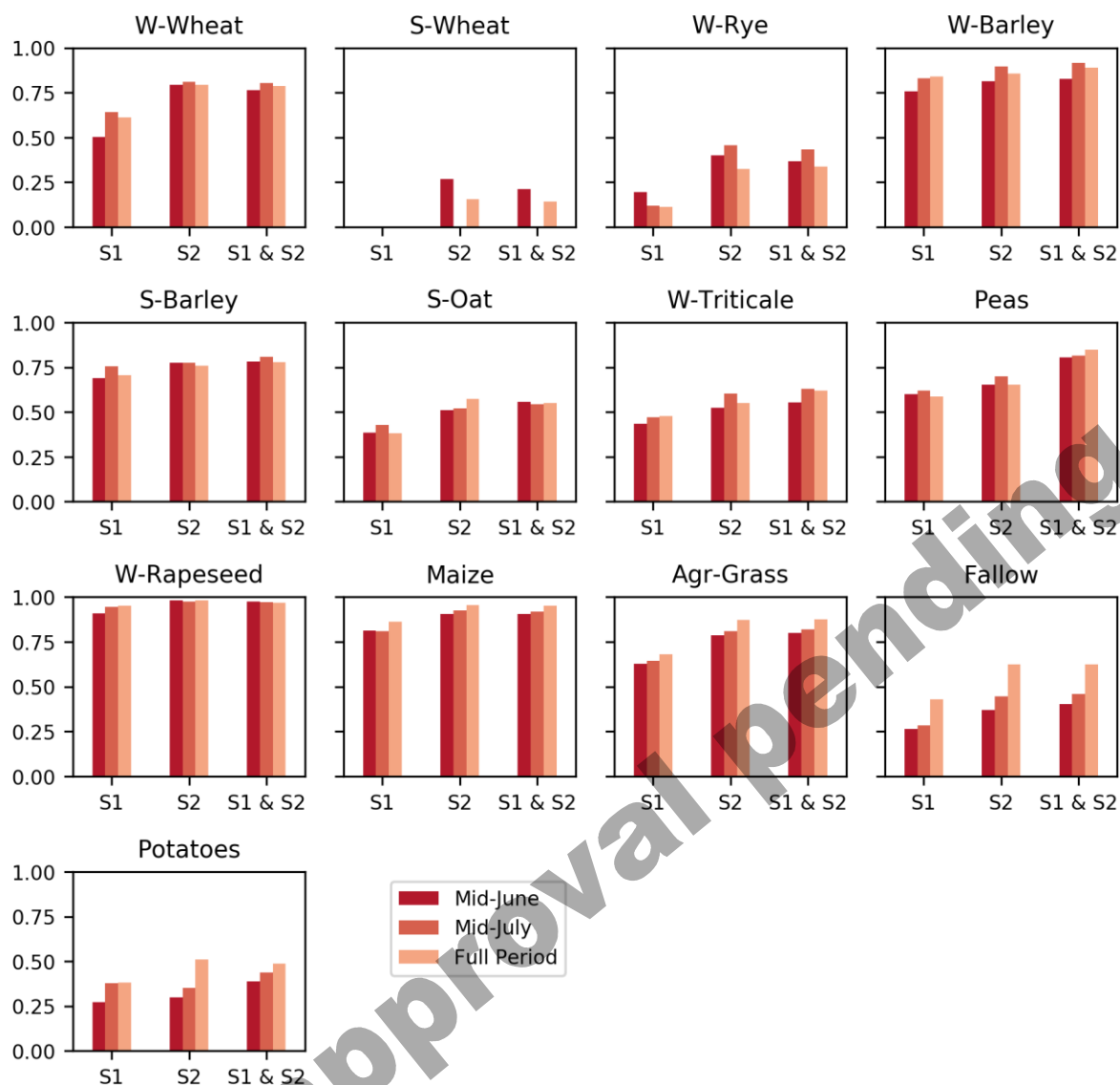


Figure 3-129: Phase 1 - Class-wise field-level F1-Scores (mean of User's and Producer's Accuracy) for the different experiment setups, particularly the three considered periods.

With respect to the corresponding probability layers, high classification reliabilities in the respective layers 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. In fact, such decision support systems have been already implemented locally (Serra et al., 2007) and are in line with the current trends of the EO Monitoring in support of the CAP policies and payment systems. 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.

This can be observed in Figure 3-130 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.

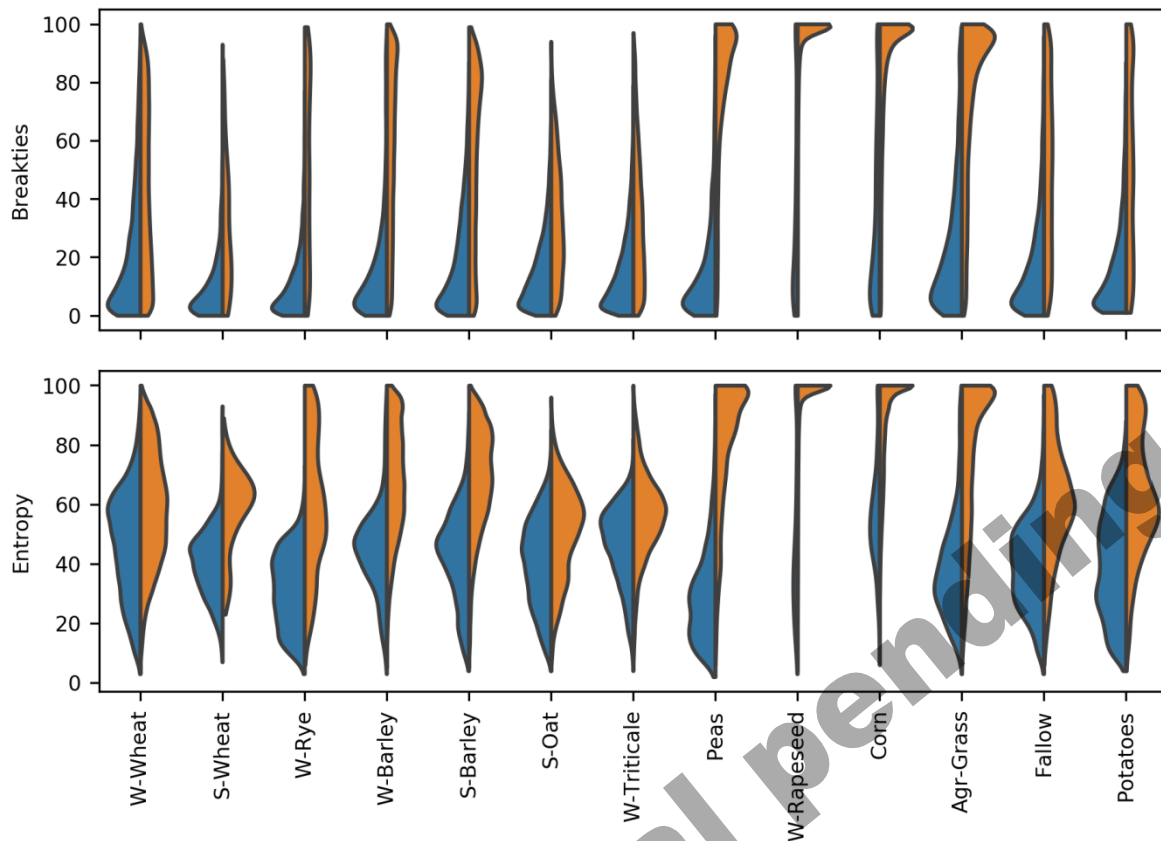


Figure 3-130: Phase 1 - Distributions of the breaking ties and entropy reliabilities of wrong (blue, left) and correct (orange, right) predictions grouped by the predicted crop type.

The following Figure 3-131 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.

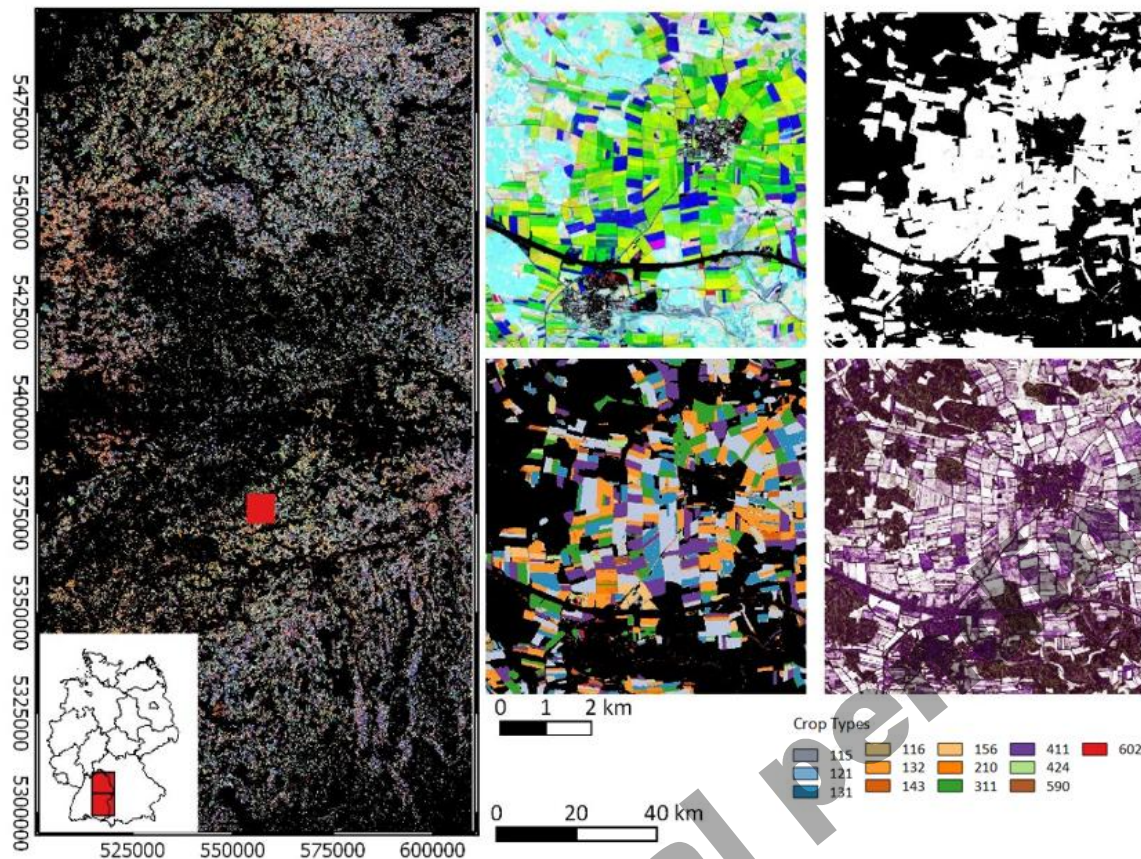


Figure 3-131: Final crop types map with the crop mask overlaid over the two processed Sentinel-2 tiles (left). The insets show an RGB composite of the median NDVI layers of the three two-month periods (upper left), the crop mask (upper right), the crop types with the Crop Mask overlaid (lower left) and and RGB composite of th three reliability layers maximum, breaking ties and entropy.

Figure 3-132 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.

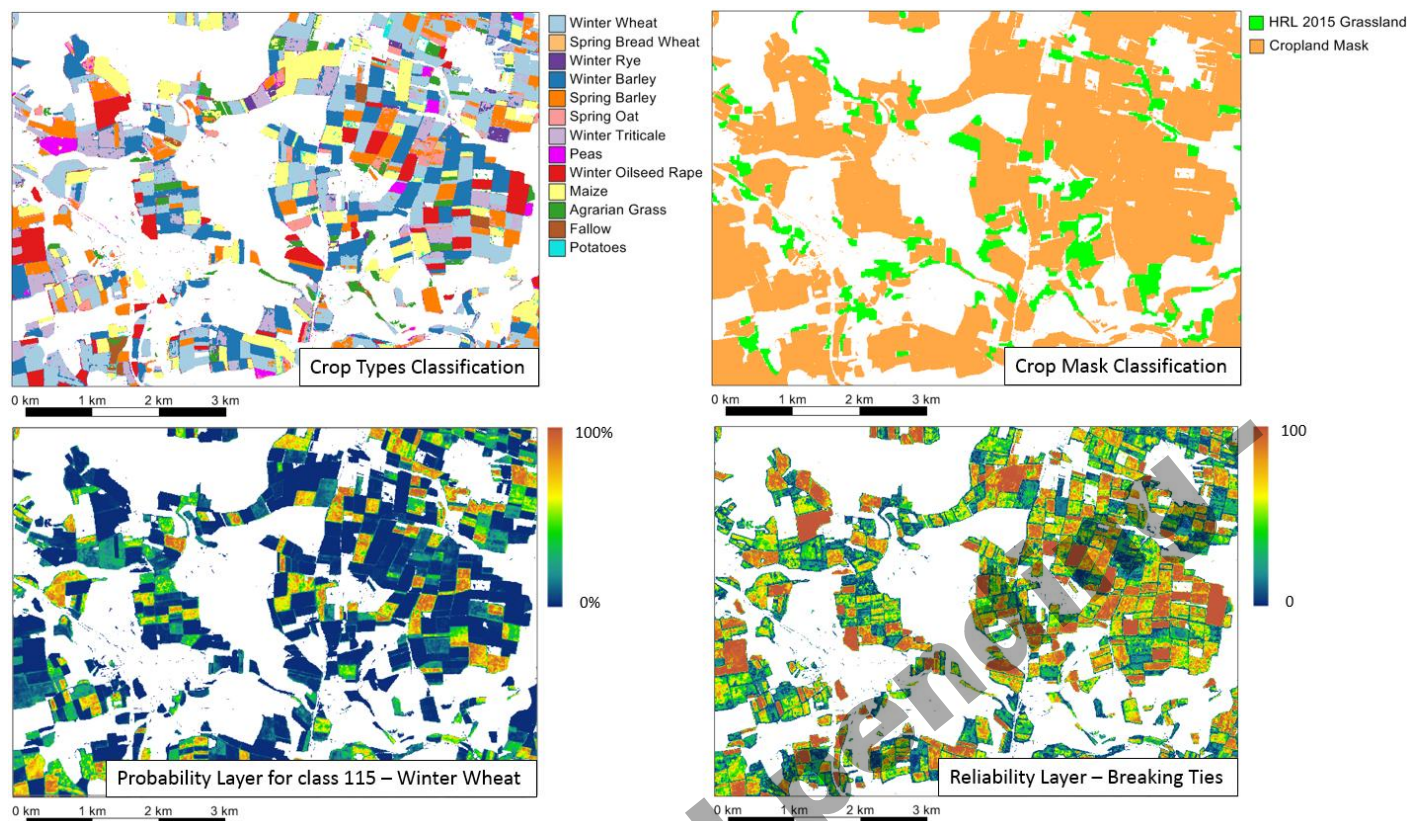


Figure 3-132: Phase 1 Details - Top left: detailed view of the crop type classification for the 13 crops. Top right: crop mask classification together with the HRL 2015 Grassland layer. A good distinction between crops and grassland was achieved. Bottom right: Example of the reliability layer 'breaking ties' as described in chapter 3.2.4.2. Bottom left: probability layer for the class winter wheat.

Results of Crop Mask and Crop Type Mask in Phase 2

The classification results in phase 2 largely confirm the findings of phase 1 in terms of the significance of the sensor: Sentinel-2 only leads to higher accuracies than Sentinel-1 only. This is confirmed for both, Crop Mask and Crop Type Mask (Table 3-77). Altogether the overall accuracy achieved for crop mask and crop type with 94 % and 86% respectively is very satisfying, even more when taking into account that both reference bases, the LUCAS data as well as the LPIS data have their limitations (see previous chapters).

Table 3-77: Phase 2 - accuracies for Crop Mask compared to Crop Types at Pixel level and with different sensor experiments; strengths and weaknesses of all three experimental set ups

	OA Crop Mask	K Crop Mask	OA Crop Type	K Crop Type	Processing Cost	Specific Problems
S1 pixel level	88,00%	0,64	73,00%		+	Foreshortening, layover in strong relief, speckle
S2 pixel level	94,00%	0,80	82,00%		+	Clouds/cloud shadows
S1&S2 pixel level	94,00%	0,81	86,00%		++	As in S1/S2 at pixel level; the strength of one sensor type can compensate for the weaknesses of the other and vice versa.

However, different to phase 1, the combination of the two sensors does in most cases improve the S2 based classification. This is especially the case for the cereal classes, for *peas+beans*, and for *lentils*, where the combination of both sensors seem to develop their full potential () and show a remarkable increase of the F1Score.

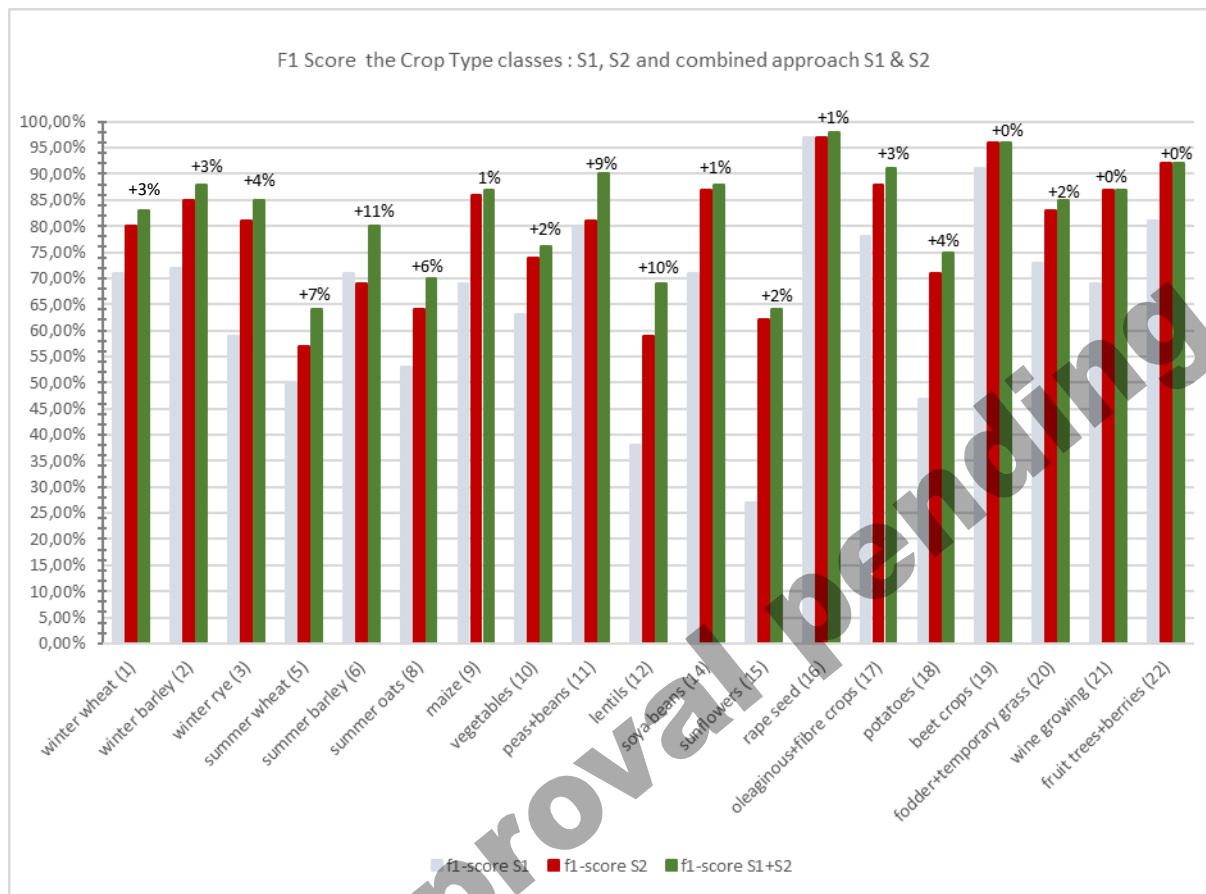


Figure 3-133: Phase 2 - F1 Score for all Crop Type classes: S1, S2 and combined approach S1 & S2 and the improvement of accuracies by using both sensors

The class-wise F1-Scores (mean of User's and Producer's Accuracy) depicted in the figure above (Figure 3-133) show that *maize* and *winter wheat*, *summer barley* and *rape seed* which account for 27.32%, 21.38%, 6.51% respectively 5.24% of the whole crop area, can be classified with very high accuracies, but also other crop types with less percentage show similar accuracies. That indicates that the level of representation within the crop area might not be the only aspect for being well detected and differentiated.

Instead, Parcel size seems to have a very high impact on the classification result, as shown in Figure 3-134. It shows the maximum size of parcels per Crop Type which corresponds largely with an average size of parcels. LPIS data suggest a tendency to grow for example cereals, *maize* or *rape seed* on larger parcels whereas *vegetables* or *lentils* usually grow on smaller parcels.

It was one of the aims in phase 2 testing to find out if and how various cereal types could be identified and differentiated from each other. The tests in phase 2 focused on the main cereal types in Central Europe, which are the groups of summer and winter cereals split into various cereal types. Wheat, barley, oats and rye can be found growing all over Europe and can be well detected with means of remote sensing. Further cereals such as spelt, emmer, kamut tend to be more grown than in the past but still play only regional roles.

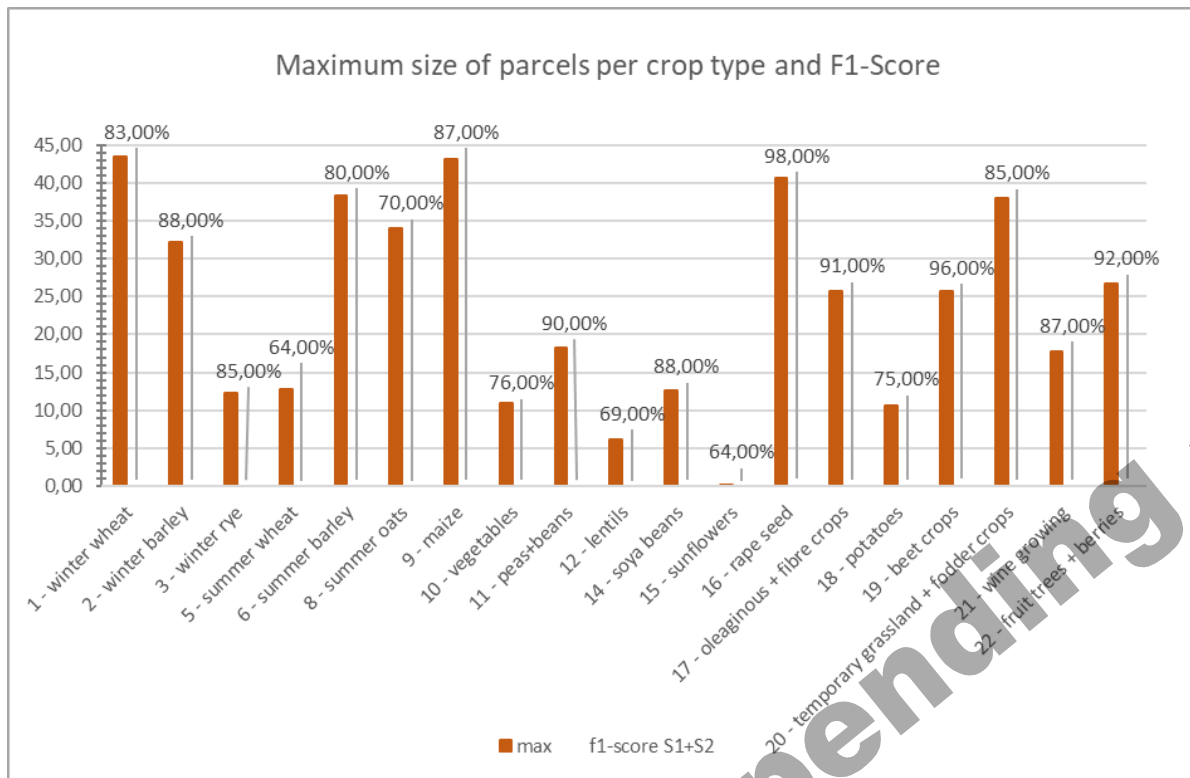


Figure 3-134: Phase 2 - Maximum size of parcels and F1-Score per Crop Type

The fact that the chosen cereals (classes 1-8) show slightly lower accuracies in the classification compared to most other crop types such as *oleaginous + fibre crops*, *peas + beans*, *soya*, *beet crops* or *fruit trees + berries* is caused by the high similarity in terms of farming management (similar sowing and harvesting times), similarity in vegetation development and in their high similarity in phenology which leads also to difficulties in spectral differentiation.

These two aspects – percentage of area covered by a crop type plus maximum parcel size are the main reasons for high accuracy in crop type classification. Limitations considering these aspects led to leaving out the classes *winter oats* and *summer rye* (low occurrence in the test site + small parcel size). However, the classes *lentils* and *sunflowers* indicate that a specific phenology and very distinct spectral characteristics could compensate those limitations and still lead to reasonable accuracies (Figure 3-134 above and Figure 3-135 below).

A detailed Crop Type classification aiming at high accuracy in differentiation therefore has to consider

- a sufficient representation within the region marked by number of parcels and even more by size of parcels in order to get an adequate sample base
- high homogeneity **within** the plants representing one crop type crop indicated by similarity in vegetation period, phenology and spectral characteristic
- high differentiability **between** the Crop types indicated by differences in vegetation period, phenology and spectral characteristic.

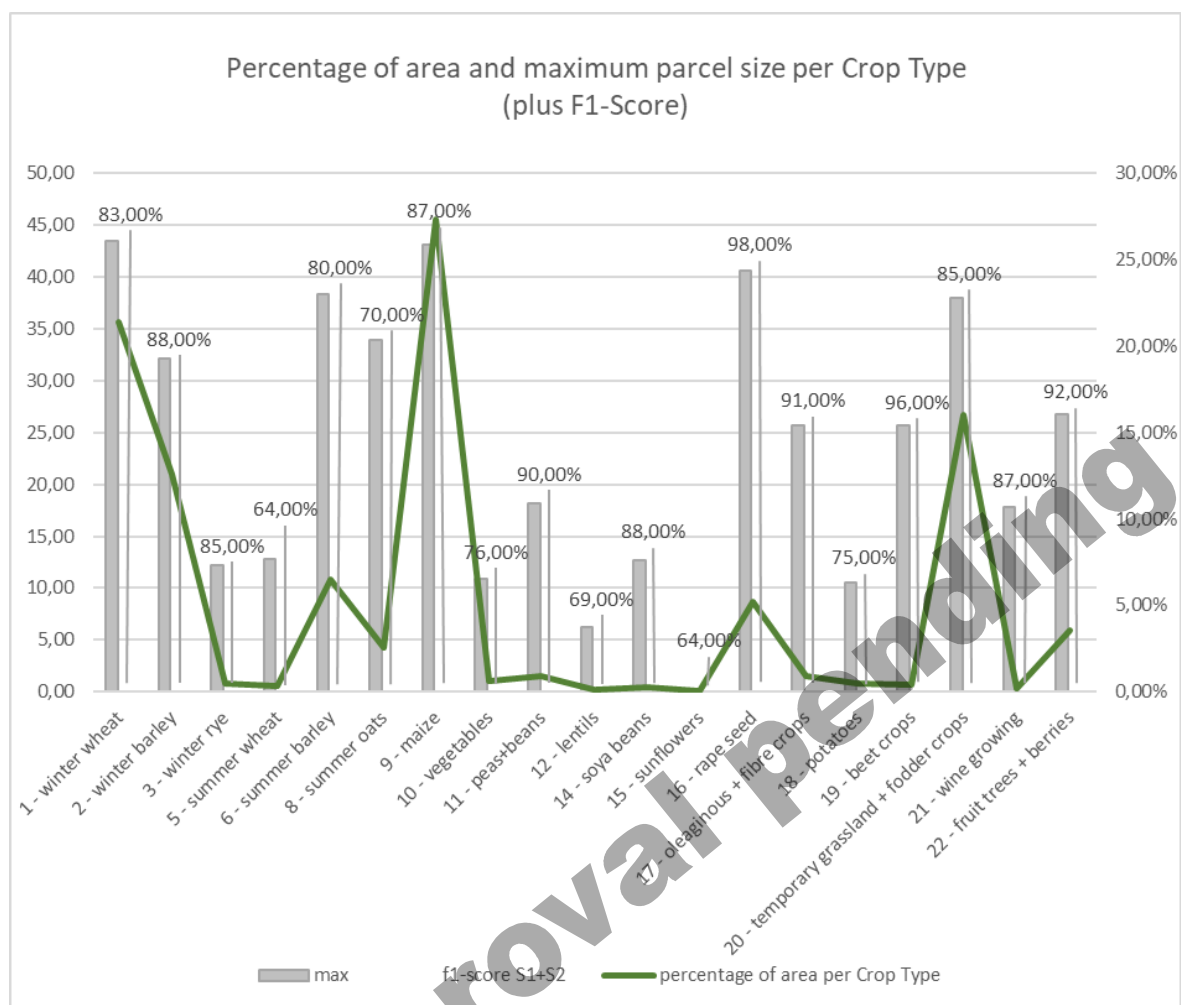


Figure 3-135: Phase 2 - Connection between percentage of area covered by each Crop Type and accuracy within the classification

Despite high accuracies (see Figure 3-136), it can be seen that for some crops the confusion is high. Especially cereals are mixing up with each other. Further, *vegetables* and *lentils*, are mixed with several other classes. Also the classes of *temporary grassland + fodder crops* and *wine growing* show strong overlaps.

In case of cereals, the reason is first and foremost the high similarity, already mentioned above. The reason for the mixing of *vegetables* and *lentils* with other crop classes might be caused by the small parcel size leading to limited number and quality of samples, but might also be caused by similarities in vegetation periods and phenology. As for the *temporary grassland + fodder crops* and *wine growing*, the heterogeneous land cover caused by grassland patches between the vines, lead to high commission errors for *temporary grassland + fodder crops*. That is also the case for class *fruit trees + berries*, depending on the growing status of the trees and the level of covering the (grassland) ground.

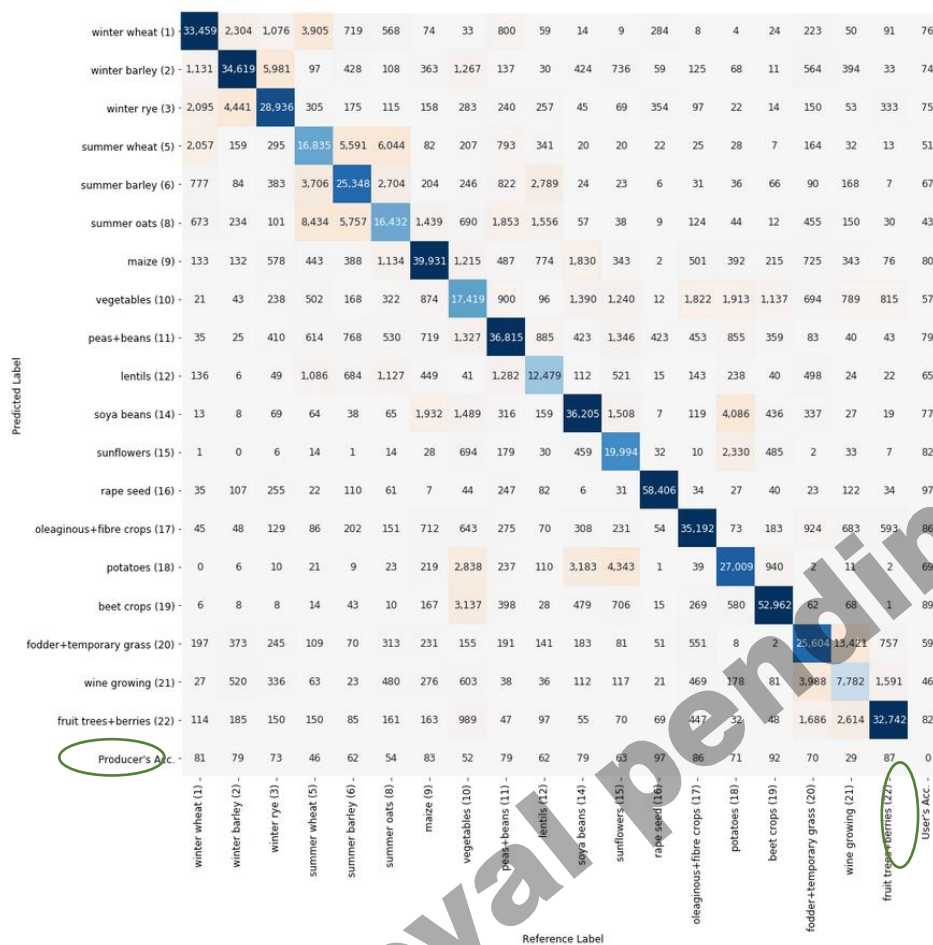


Figure 3-136: Phase 2 - Confusion matrix for Crop Type classification at Pixel level (PA indicated down left, UA down right)

In phase 2, the approach of grouped Forward Feature Selection (FFS) led to a reduction of features from 676 to 221 for the Crop Type Mask. The number is still very high but could be explained by the high number of classes and the necessity of collecting a broad range of information for a large variety of crops in an extended time window. In order to detect every necessary information for every time step for every crop type, only a broad ensemble of time features is able to provide suitable classification accuracies for 19 crop type classes. This fact is displayed the figure where the curve is not as steep as usual or as the curve for the crop/non-crop classification above (Figure 3-138) but achieves the saturation value at a later stage with a higher number of best-of time features (Figure 3-137).

The ranking among all features depends upon the time window. Usually indices like NDVI and NDVVH play an important role both by number and by ranking and have therefore been included in the feature settings. Also features like min, max and the percentiles have high impact on the classification and are therefore selected. However, the feature selection for the sensors differs. As for Sentinel-1, nearly all NDVVH features accumulate with high numbers in the time window of Mid-May to Mid-July, whereas features of VV and VH predominate in all other time windows. For Sentinel-2 is different: NDVI features predominate during Mid-March to Mid-May, NDVI, NDWI and IRECI are highly ranked from Mid-May to Mid-July (also indicated by the high number of these features in the respective time windows), band features seem to be of importance in all time windows. As for the type of features: min, max, mean and percentages have proven to be always selected with high ranking for vegetation detection. Only the Ratios have been completely neglected by the FFS process and thus could be left out for further testing.

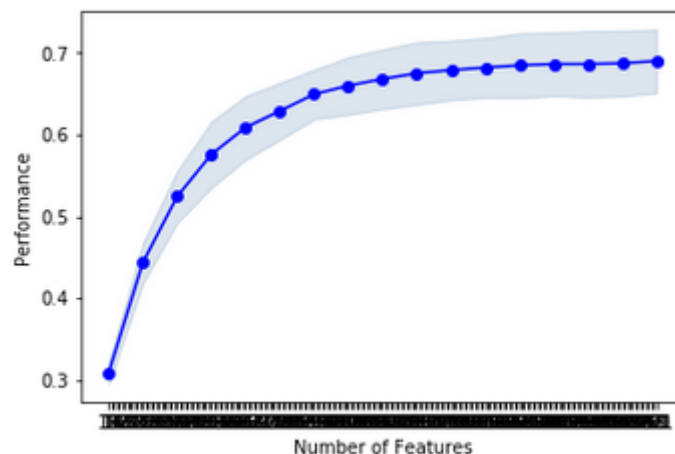


Figure 3-137: Phase 2 - Overall accuracy OA based on the cross-validated training samples dependent on the number of selected features for the Crop Type classification. The curve describes the saturation process where the slope is less steep than usual indicating that a higher number of time features is necessary for getting sufficient accuracies.

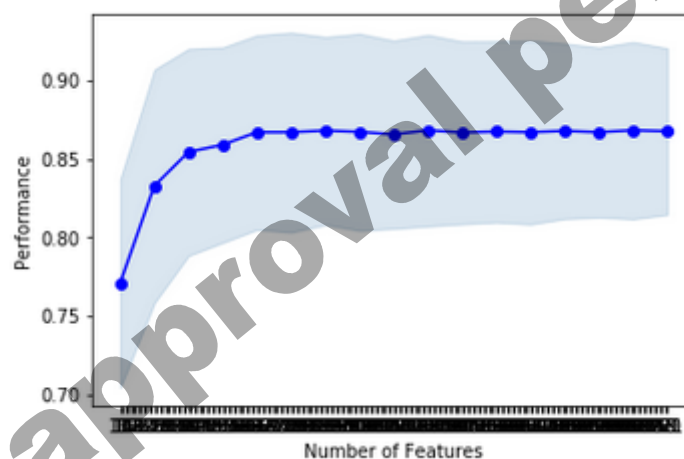


Figure 3-138: Phase 2 - Overall accuracy OA based on the cross-validated training samples dependent on the number of selected features for the Crop Mask classification

As the comparison of accuracies already implies (Figure 3-110Figure 3-133), the importance of Sentinel-2 time features is significantly higher than that of Sentinel-1 time features (Table 3-78). It should be mentioned that the highest number of 39 Sentinel-1 features accumulate in the time window Mid-May to Mid-July followed by that of Mid-March to Mid-May, whereas the Sentinel-2 features of Mid-May to Mid-July show the highest number of 65 features followed by a still noteworthy high number for Mid-March to Mid-July and same for Mid-July to Mid-October. This indicates on the one hand the high complementary potential the combined approach of sensors as well as the meaningfulness of the chosen time windows. The number of features for the overall period Mid-March to Mid-October is smaller but high enough to make it reasonable to add these information as additional value to the classification process. The F1 Score values for the combined approach for all crop types shows that for some crops like winter rye and lentils even a low F1 Score for Sentinel-1 improves the accuracy in a combined approach.

Table 3-78: Number of Time Features selected by the grouped Forward Feature Selection per sensor and per time window

	S1	S2
Mid-March to Mid-July	26	26
Mid-May to Mid-July	39	65
Mid-July to Mid-October	13	26
Mid-March to Mid-October	13	13
total	91	130

In contrary to phase 1, no field level analysis has been conducted for the test site due to limited coverage with reference data. The available LPIS data cover only parts of Baden-Württemberg and Austria and thus only a constraint part of the two test tiles 32 TNT and 32UNU, why this analysis was done in other test sites.

From the user's perspective, the outcome on real field level and also the visualisation in field level might be more interesting than the pure pixel result. Thus, using geometries could not only improve the classification by providing majority results but will also correspond more to the real. This investigation however, has been part of the prototype analysis.

Probability measures:

The analysis of the probability layers of the crop type classification of phase 2 confirmed the findings of phase 1: high classification reliabilities in the respective layers usually correspond to correct predictions. However, it is not recommended to use this information for direct actions referring to subsidy policies, because the validity of the reliability layers is crop type dependent. Concerning well separable crop types, the reliability layers are informative with respect to the likelihood of a correct classification, such as *maize* or *rape seed*. For less separable crops (see confusion matrix) an application cannot be recommended.

Results for the Crop Mask:

The crop mask for test site 32UNU and 32TNT for 2018 (Figure 3-139) comes with an OA accuracy and Kappa indices of 88% and 0.4 for Sentinel-1 approach, 94% and 0.80 for the Sentinel-2 approach, and 94% and 0.81 for the combined approach (Sentinel-1 & Sentinel-2).

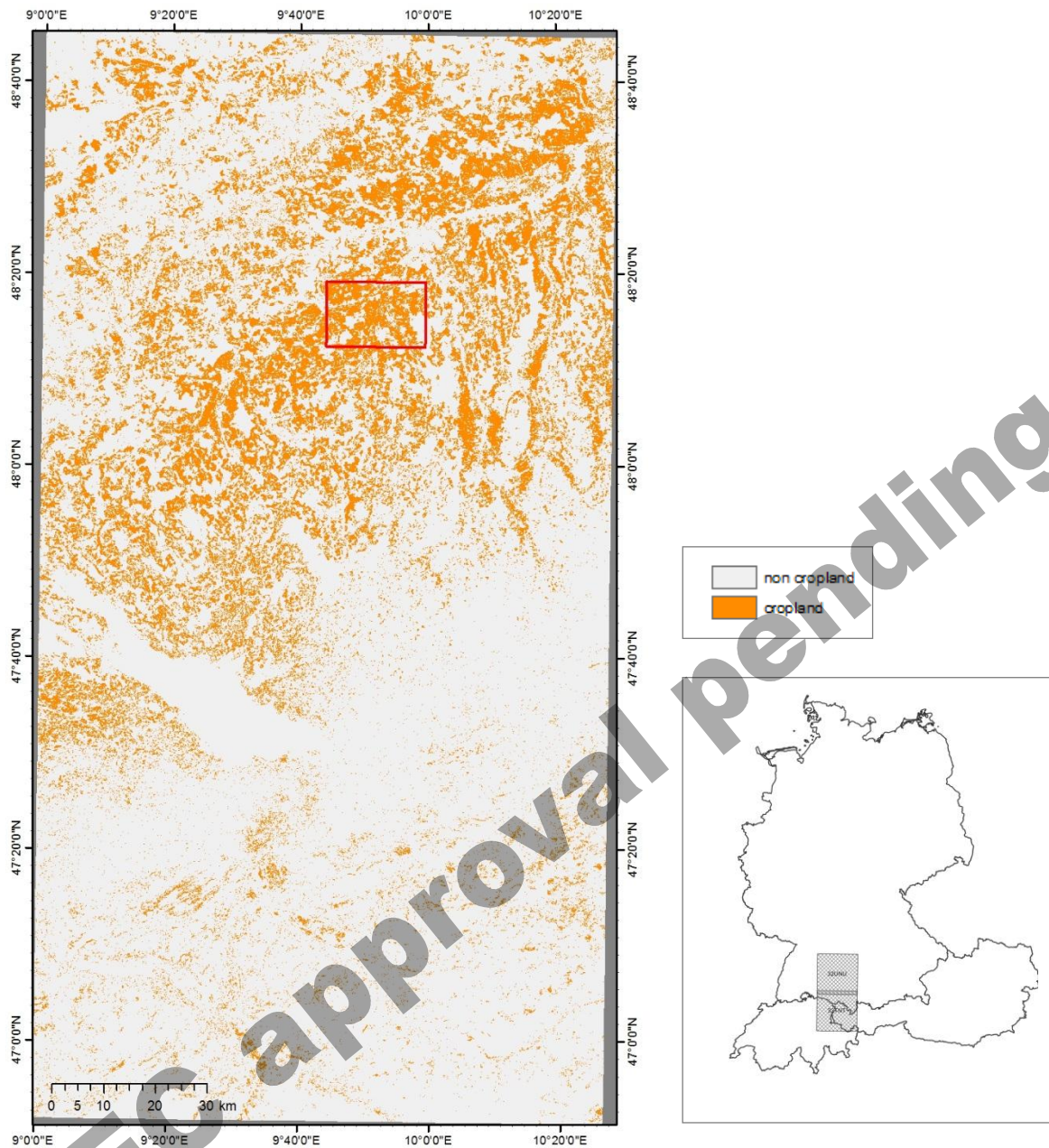


Figure 3-139: Phase 2 - Crop Mask for test site tiles 32TNT and 32UNU (left) and location of the test site within the border region of Germany, Switzerland and Austria (right)

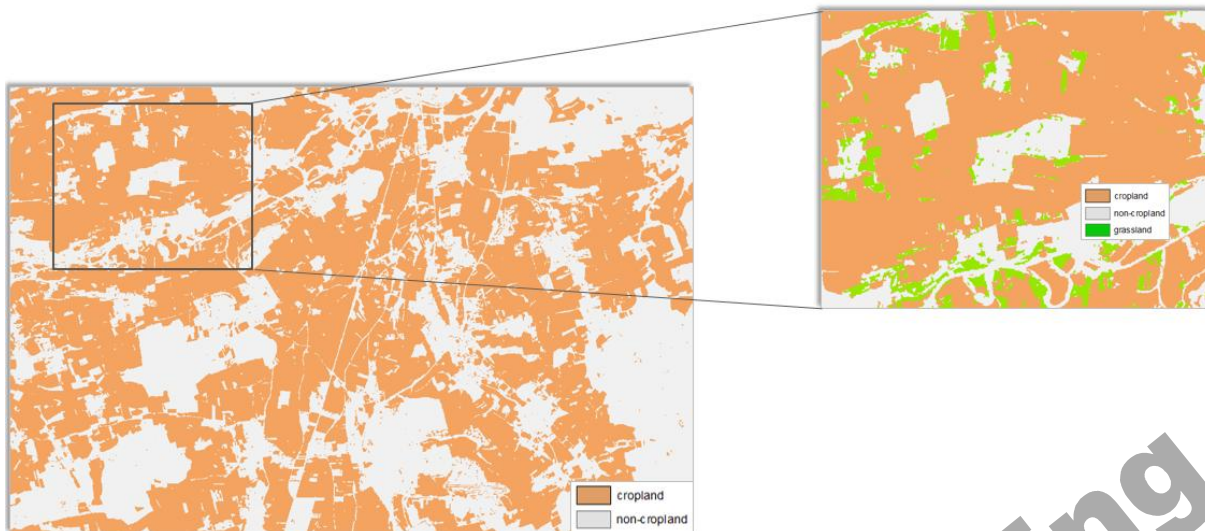


Figure 3-140: Phase 2 - Detail of crop mask 2018 complemented by the grassland mask 2018.

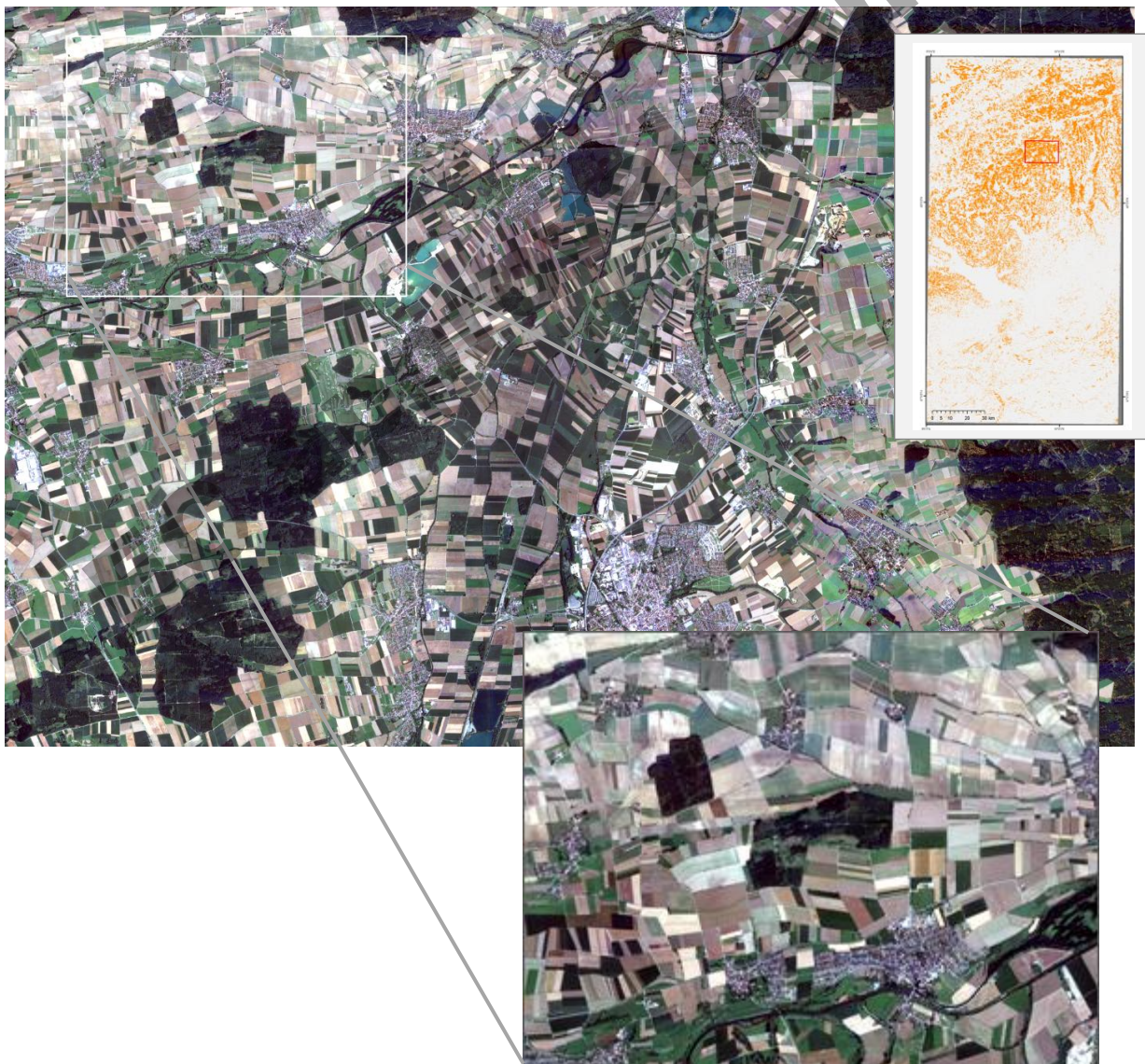
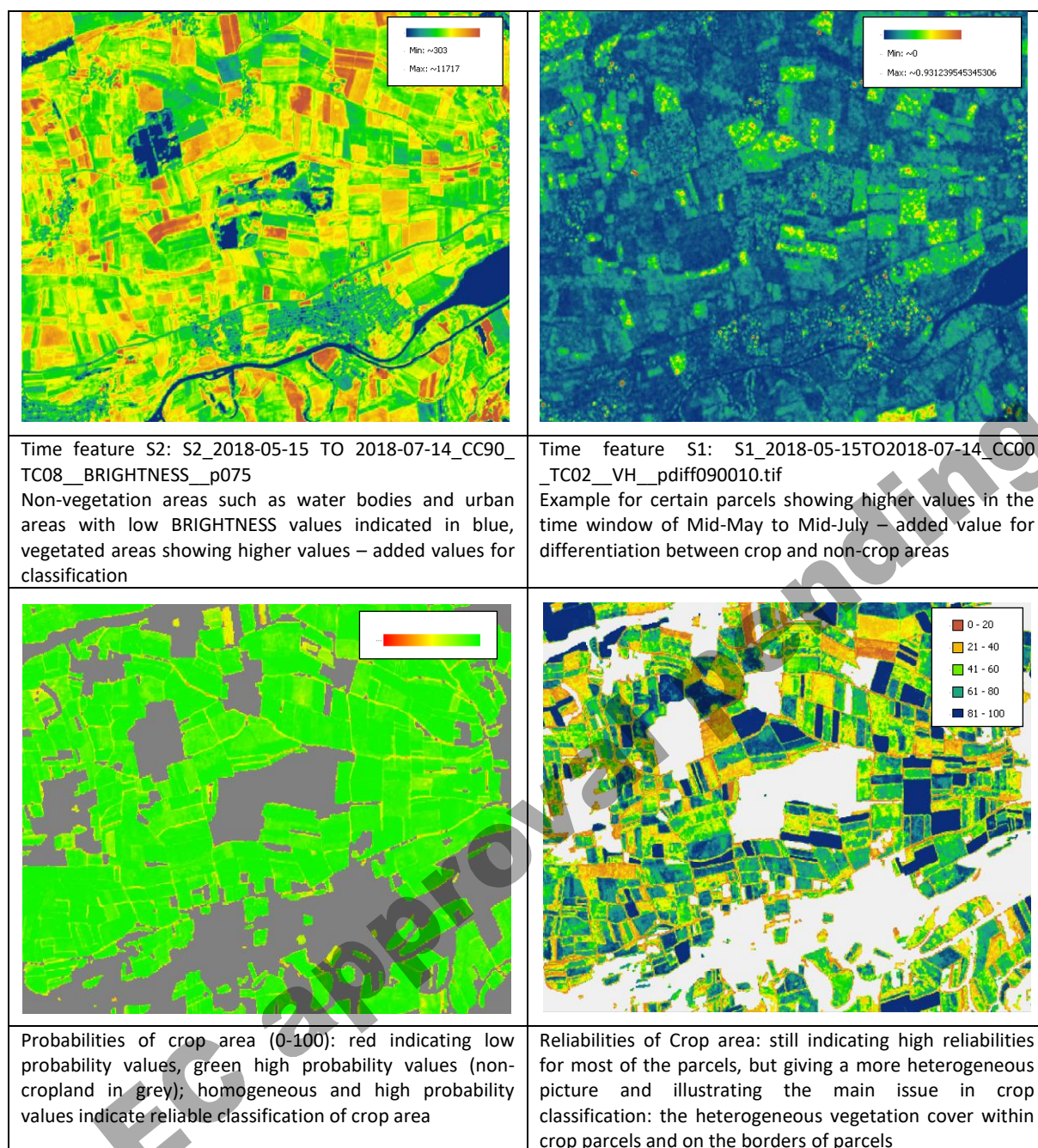


Figure 3-141: Phase 2 - surroundings by google Earth imagery in 2018: Laupheim, SW of Stuttgart, Baden-Wurttemberg



Figure 3-142: Phase 2 - Part of the Crop Mask and Details for test site Central

Detail of Crop Mask, displaying cropland and non-cropland	Sentinel-2 imagery from 2018_05_07 (NIR-SWIR-RED)



The result for the Crop Mask for test site Central is shown above (Figure 3-139, Figure 3-140, Figure 3-141, Figure 3-142). Details of the classification for an area South West of Stuttgart near Laupheim (Figure 3-139, indicated in red) are given in the following insets. The Sentinel-2 image from 2018_05_07 (NIR-SWIR-RED) indicates parcels with vital vegetation in red but gives only the partial picture since other land cover such as water bodies, urban areas or trees are difficult to identify. Agricultural management systems, especially in Middle Europe leads to very heterogeneous land cover patterns from beginning of spring until End of October. In order to cover the agricultural area in comprehensive manner, it is necessary to collect information on tilling, growing, vegetation peaks and harvesting over the whole time period. The dense time series of Sentinel-1 and Sentinel-2 joins these information of the time features per time window, such as the Sentinel-2 feature displaying the BRIGHTNESS from Mid-May to Mid-July (S2_2018-05-15 TO 2018-07-14_CC90_TC08_BRIGHTNESS_p075) or Sentinel-1 feature from the same time window (S1_2018-05-15 TO 2018-07-14_CC00_TC02_VH_pdiff090010.tif) and at the same time supports differentiation from non-crop areas. The high probabilities for the crop area as well as the high reliability values verify the effectivity of the approach.

Results for the Crop Type Mask:

The result for the Crop Type Mask for the test site Central 2018 comes with an OA of 73% and Kappa index 0.81 for the Sentinel-1 only approach, 82% and Kappa index 0.81 for the Sentinel-2 only approach and 86% with Kappa index of 0.85 for the combined approach of Sentinel-1 & Sentinel2.

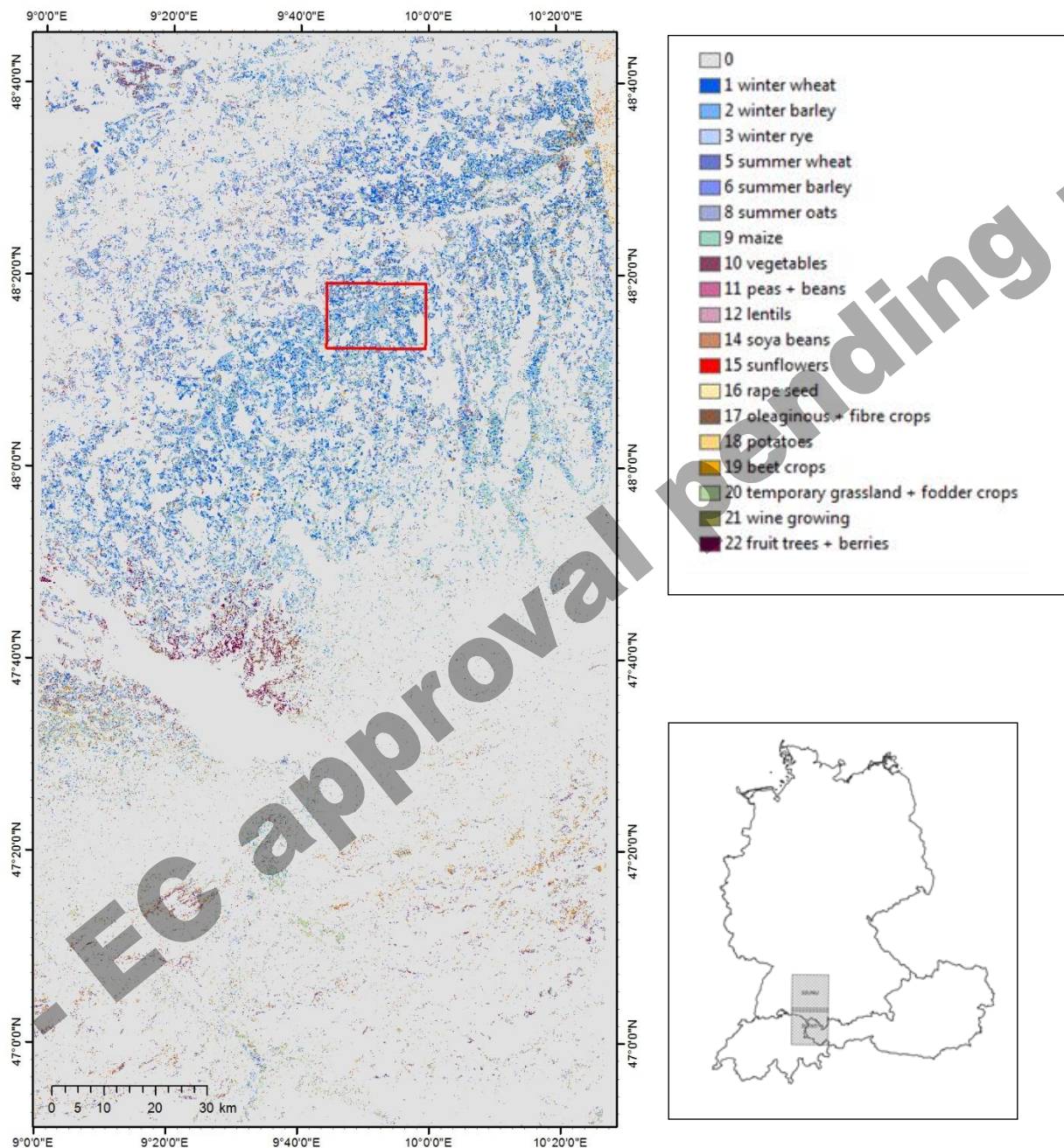


Figure 3-143: Phase 2 - Crop Type Mask for test site in tiles 32TNT and 32UNU (left) and location of the test site within the border region of Germany, Switzerland and Austria (right)

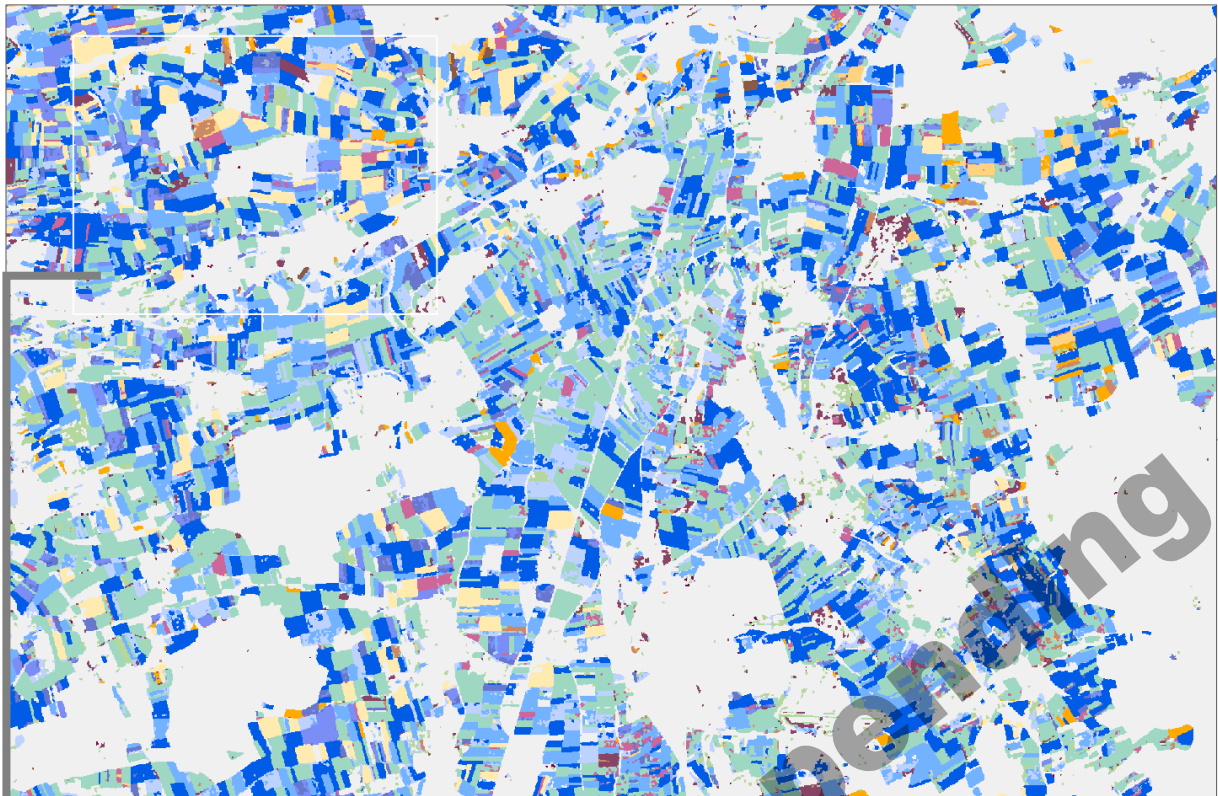
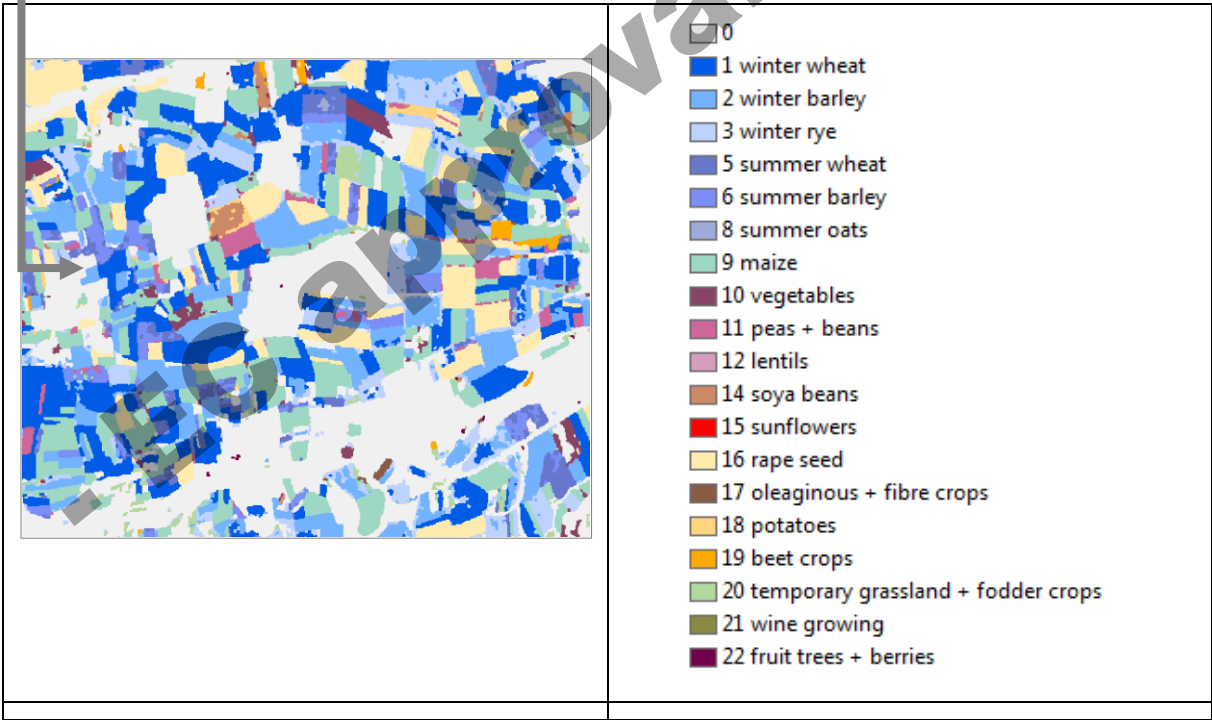


Figure 3-144: Phase 2: Part of the Crop Type Mask 2018 and Details for test site Central 32UNU and 32TNT



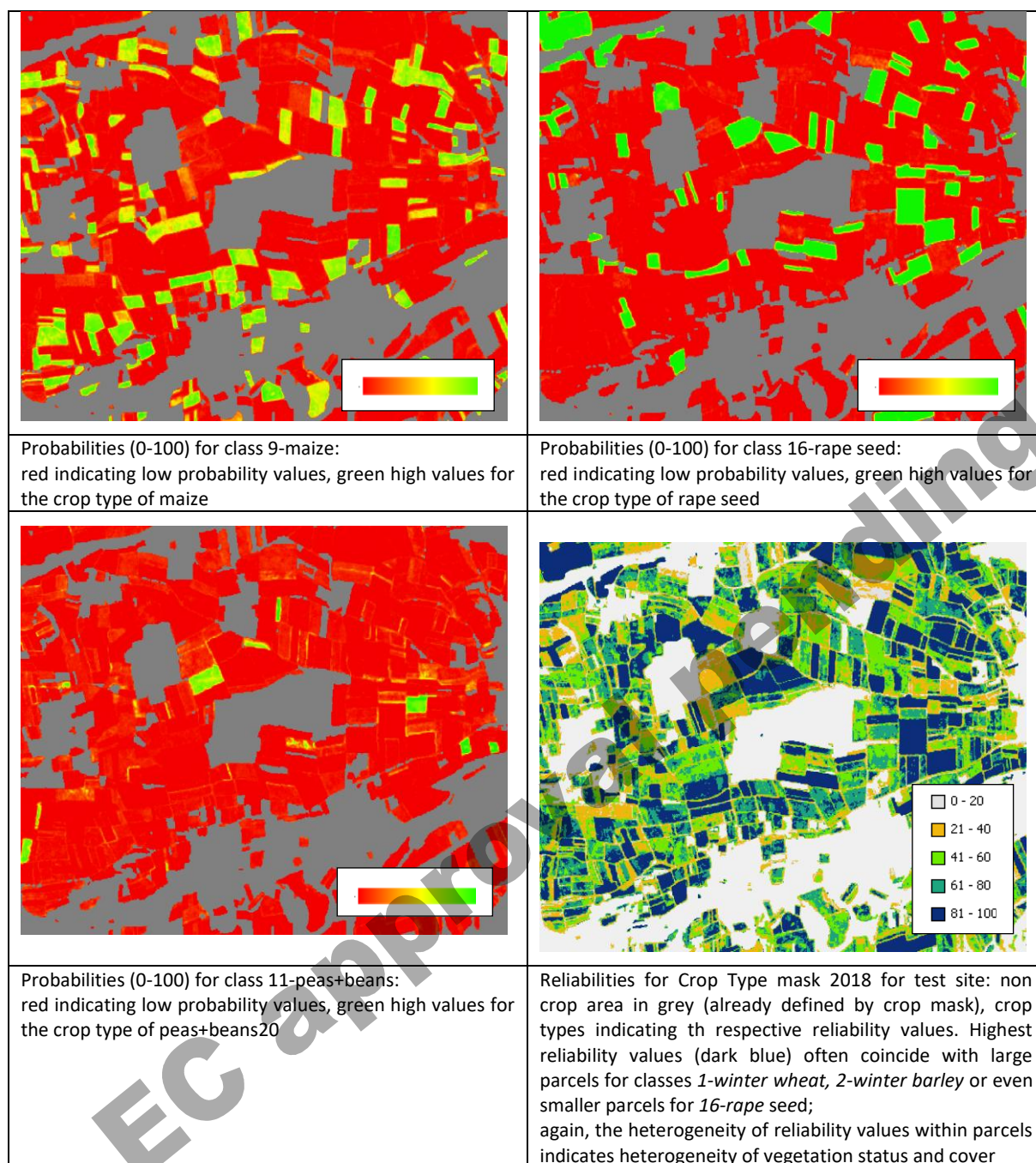


Figure 3-145: Phase 2 - Detail of Crop Type Mask with RGB of NDVI median TF for 3 time windows (March-May, may-July, July, Oct)

A first run combining Sentinel-1 and Sentinel-2 data (using the same technical workflow as for the Crop Mask) provided an OA of 86% and Kappa index of 0.81. The good detection and differentiation of the different classes is promising. When considering the whole demo site, a regional stratification should be taken into account in order to cover the shifted vegetation period of the alpine region in comparison to the area in the North of the demo site. This approach could reduce overlaps between the crop mask/crop type mask (concerning classes agrarian/fodder grass) and the grassland mask as well as reducing the issue of mixed crop types.

This method is only applicable for the larger area of the demo site as it is related to a better fit to differentiated local conditions that are not significant in the test area. This is one example where the across-scale approaches might differ, and that required being tackled in parallel in phase 2 between Task 3 and Task 4 sub activities. With the provision of the LPIS data of Bavaria and the larger region, the

number of samples should be enough to work with this approach. One issue is the overlap of the grassland layer and the class agrarian/fodder grass of the Crop Layer, which is inevitable for a layer based on a one-year period, lacking the historic information of tilling in the previous years. Without a dense time series covering historical data, natural and permanent grassland cannot be distinguished from agrarian grassland and temporary grassland being ploughed regularly.

The mono- and multi-sensory approach results for the crop mask and the crop type's tests are shown in the graphics below (Figure 3-146 and Table 3-79).

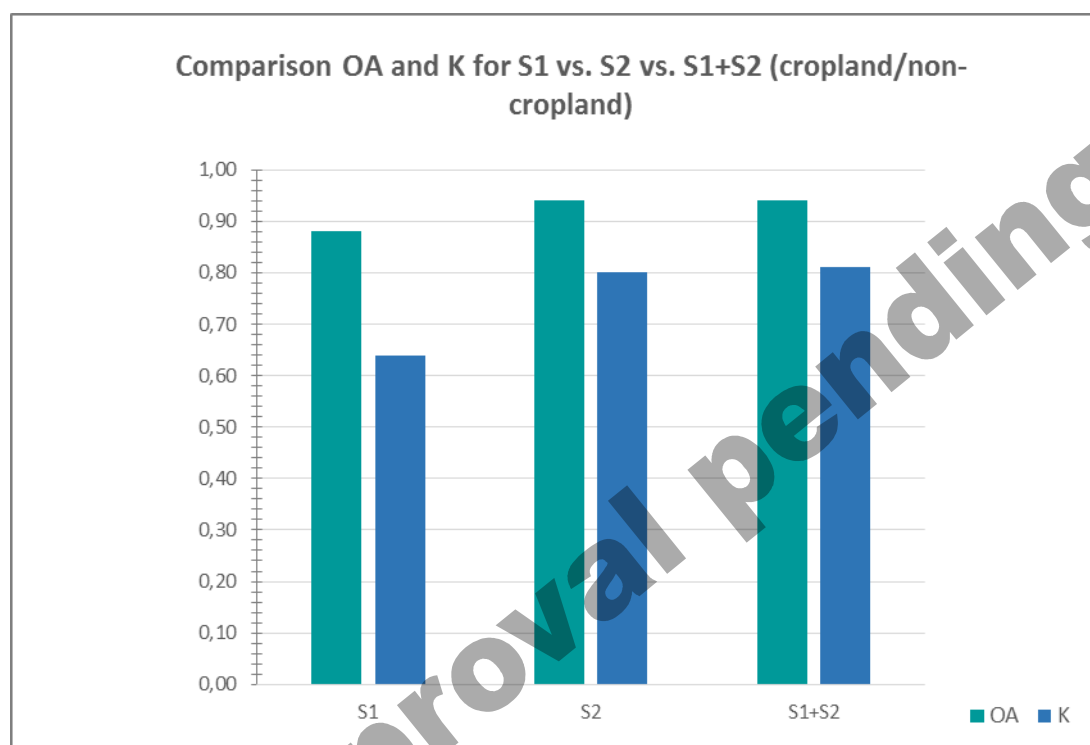


Figure 3-146: : Phase 2 - accuracies for the Crop Mask 2018 for the test site per experimental setup, S1 only, S2 only and the combination of S1 & S2

Table 3-79: Phase 2 - accuracies for the Crop Mask 2018 for the test site per experimental setup, S1 only, S2 only and the combination of S1 & S2

	K*100	OA
Sentinel-1	64,90	88,30
Sentinel-2	80,80	94,40
Sentinel-1 & Sentinel-2	81,80	94,30

The crop mask tests show that the classifications based on Sentinel-2 only data tend to show better results than those based on SAR features. In this case, the combination of Sentinel-1 & Sentinel-2 does not significantly improve the classification accuracy, when compared to the Sentinel-2-only approach.

For the crop types map, the F1 scores for each crop type are shown in Figure 3-147 (grey bars for Sentinel-1 only classification, orange for Sentinel-2 only classification, and green for the combined classification). The corresponding overall accuracies and kappa indices are respectively 73% and 0.71 for

Sentinel-1 approach, 82% and 0.81 for the Sentinel-2 approach, and 0.86% and 0.85 for the combined approach (Sentinel-1 & Sentinel-2). F1 Scores for Sentinel-2 are higher, although for most classes, the combined approach provides the best accuracies. All in all, the overall accuracy for the combined approach is significantly higher than for each of the sensors on its own (see percentages in Figure 3-133). Regarding the contribution of SAR features, when considering that in general Sentinel-2 features are usually preferred in agriculture land cover workflows, it must be noted that even if the cloud situation is not the best in 2018, there is a sufficient number of Sentinel-2 imagery for the tests carried out in the Central region. This means that the benefit of using Sentinel-1 imagery in the combined approach, might prove even higher in areas with very high/nearly permanent cloud cover.

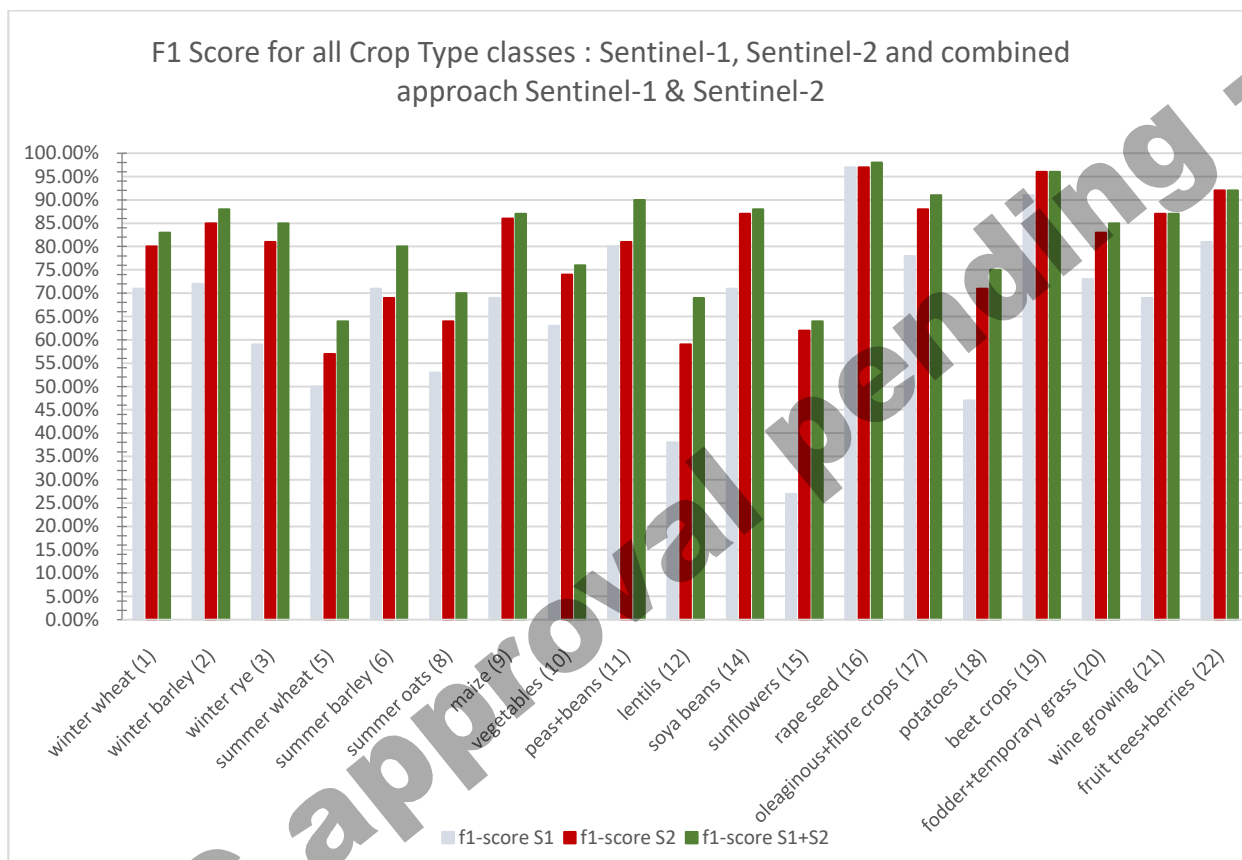


Figure 3-147: accuracies for the Crop Type Mask 2018 for the test site per experimental setup, Sentinel-1 only, Sentinel-2 only, and the combination of Sentinel-1 & Sentinel-2

3.3.4.2 Belgium site

In the following sections, the classification methods applied in the Belgium site are described. Details are given towards the methods (section 3.3.4.2.1) as well as the benchmarking criteria (section 3.3.4.2.2) and the implementation and results of the benchmarking procedure (section 3.3.4.2.3).

3.3.4.2.1 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). These crop features are not considered in the Belgian test site from this benchmark. This method is fully automated but requires *in situ* data for the training.

3.3.4.2.2 Benchmarking criteria

Overall accuracy and kappa are reported for all benchmark scenarios to assess respective classification performances. A F-score for crop types is also provided in particular where low occurrence classes were

evaluated. These metrics refer to the ECoLaSS accuracy assessment guidelines (section 2.4). The approach with the optimum cost-benefit ratio was not considered here as the cost factors did not vary much from one scenario to another.

3.3.4.2.3 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.

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 scenarios for data calibration preparation were compared. 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 pixels located on field borders which is particularly critical for agriculture due to the limited field size. 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 or mixels.

Every classification scenario achieved an overall accuracy higher than 80% (Figure 3-148). 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.

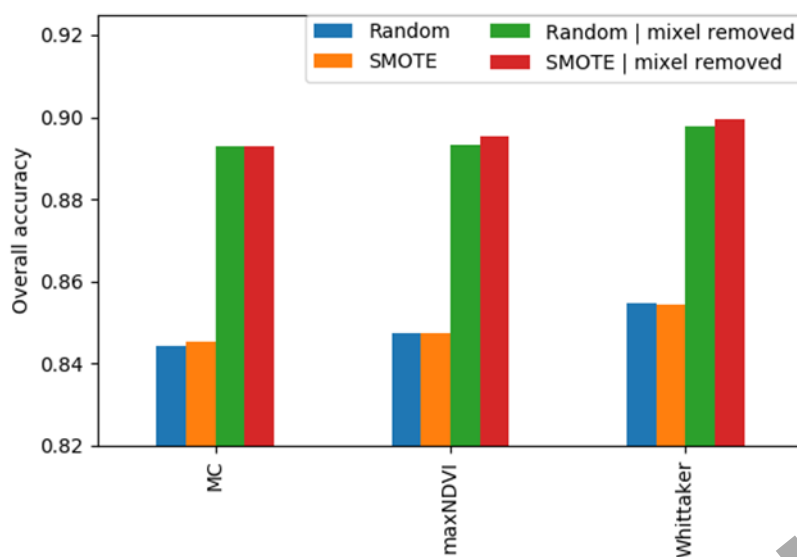


Figure 3-148: Overall accuracy for every classification scenario evaluated.

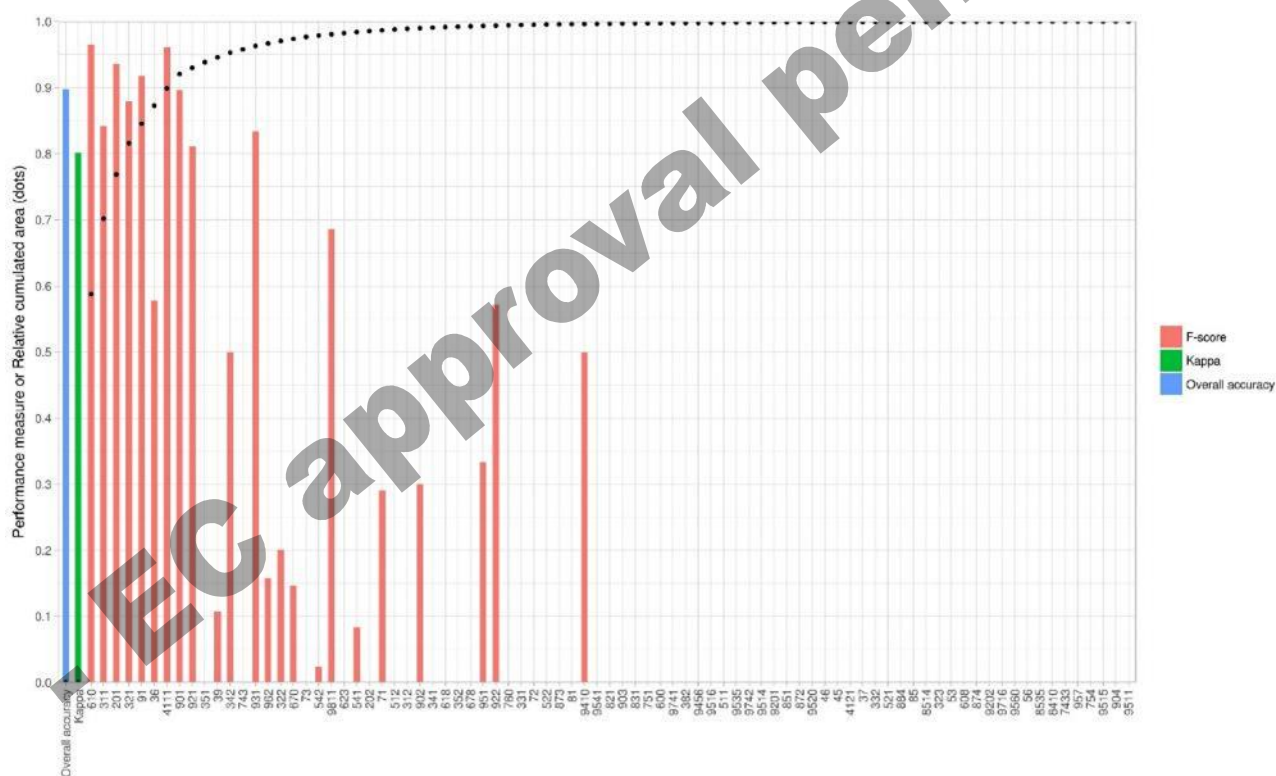


Figure 3-149: Classification F-score for each crop type ID for Whittaker inputs with random sampling and mixel removal (red). Overall accuracy (blue) for classification and Kappa (green). Relative cumulated area of crop types (black).

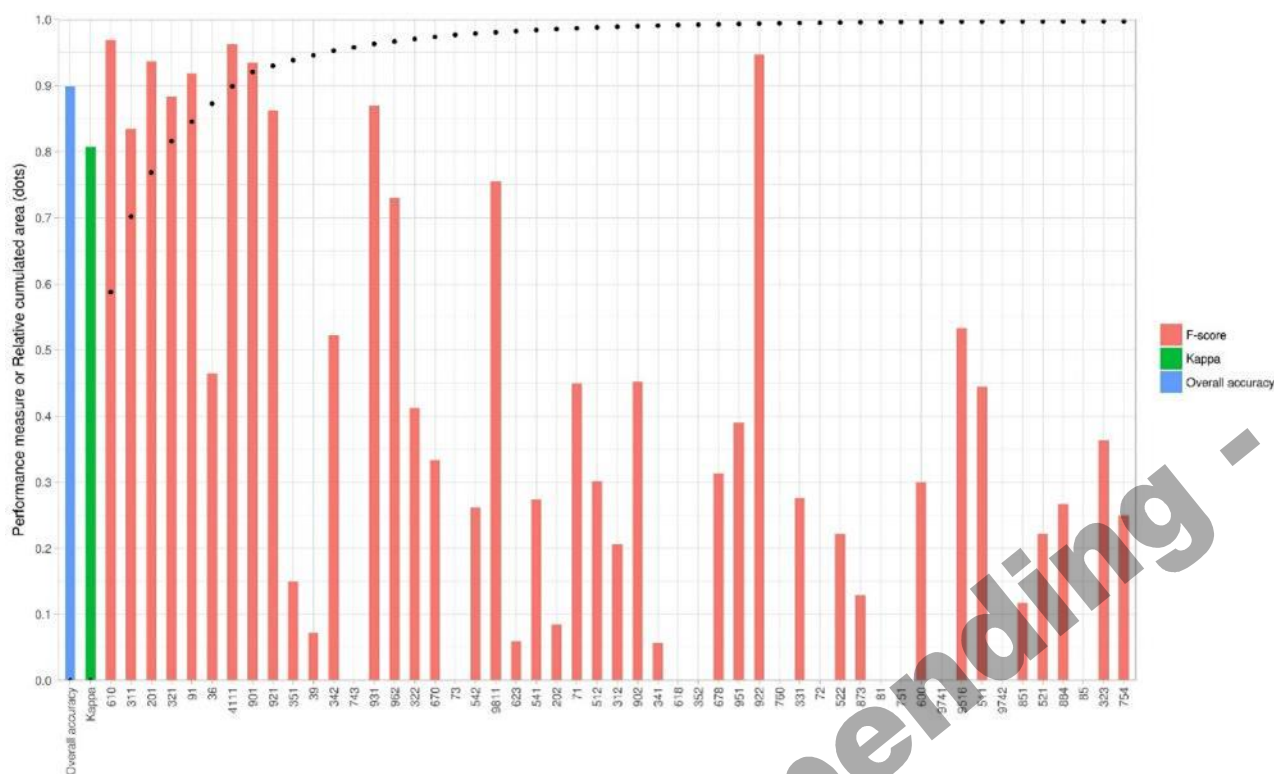


Figure 3-150: Classification F-score for each crop type ID for Whittaker inputs with SMOTE and mixel removal (red). Overall accuracy (blue) for classification and Kappa (green). Relative cumulated area of crop types (black).

To assess the performance improvement provided by the SMOTE method, classification errors on low occurrences classes must be evaluated. As seen in Figure 3-149, random sampling is not able to classify properly classes with small sample size in the *in situ* dataset. Figure 3-150 on the other hand shows that SMOTE method is useful to improve results for small occurrences classes.

3.3.4.3 African site

A classification approach similar to a scenario implemented for the Belgian site has been applied to three tiles of the Western Cape province in South-Africa for 2017 (Figure 3-151). The method used linearly temporally gap-filled Sentinel-2 images as inputs for the RF classifier. Only Sentinel-2 L1C images with a cloud cover below 90% and recorded from the 1st April to 30th November 2017 were calibrated and cloud screened by MAJA code. The pixels located at the field borders were discarded from the calibration dataset in order to reduce the mixels contribution. In addition, the parcels with a size lower than 0.5 ha were also excluded from the analysis.

The in-situ dataset was made available by the Western Cape province and included 200 different crop types. These crop types have been grouped into 16 crop types including weeds and unknown crops. As reported in the Figure 3-152, the OA reaches 70 % and the wheat, the oilseeds and barley which are most frequent crops are mapped with a promising accuracy (F1-score higher than 0,75). On the other hand, the grasses, fallows and fodder crops are poorly discriminated while they cover large extent of the agricultural lands in these areas.

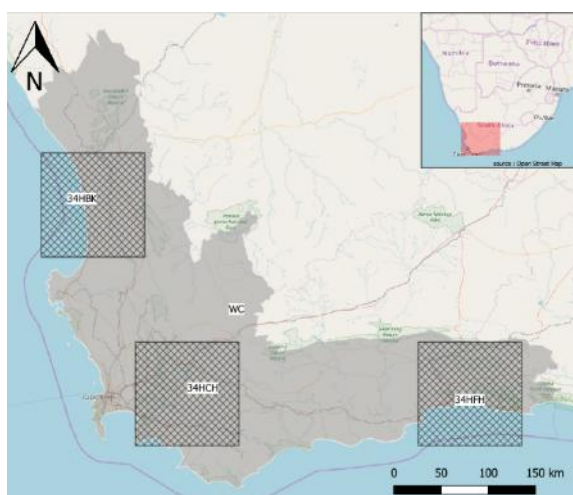


Figure 3-151: Location of the three benchmarking tiles for the South-African sites in the Western Cape province.

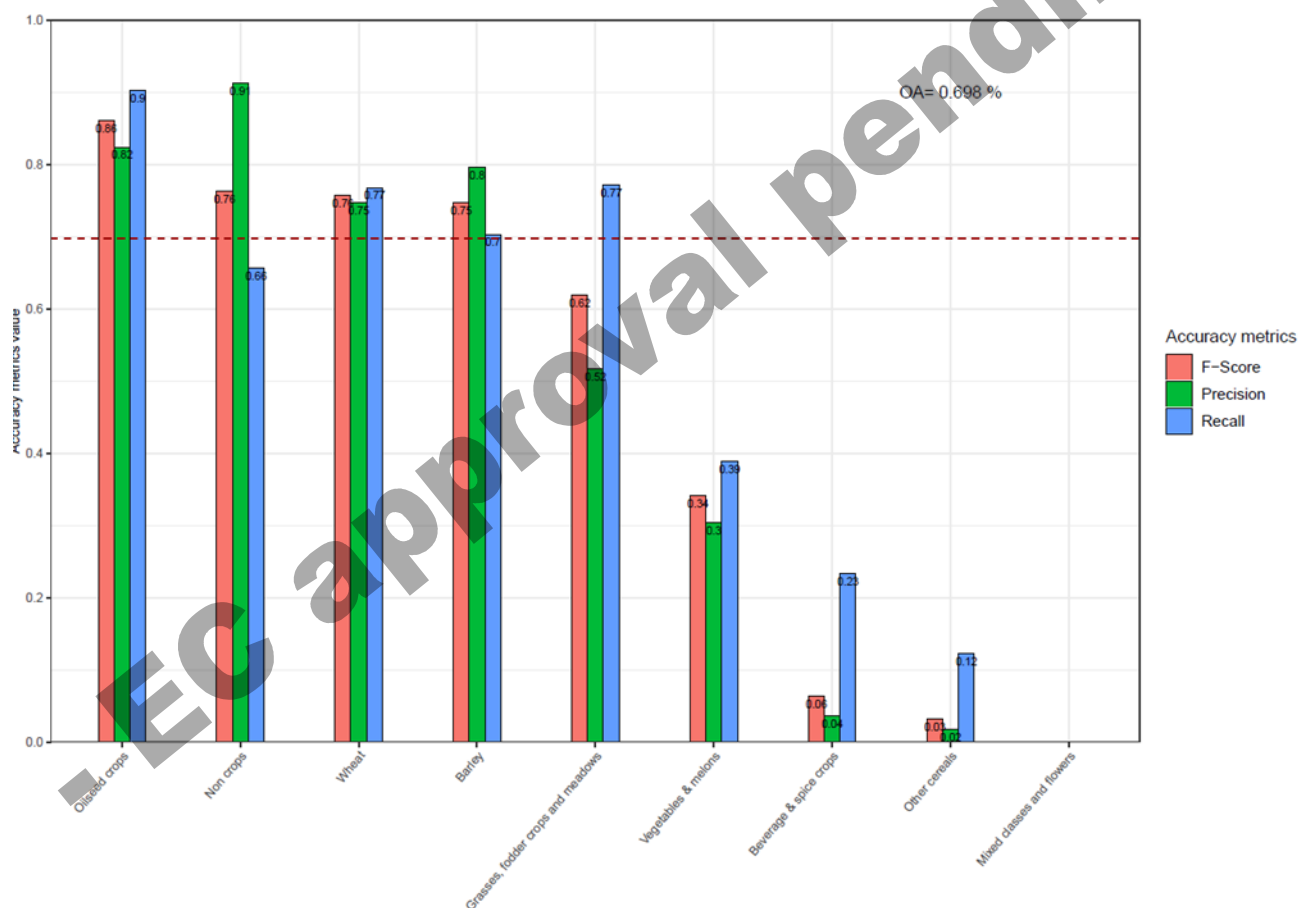


Figure 3-152: Accuracy assessment of the crop type mapping in the South African site (3 tiles).

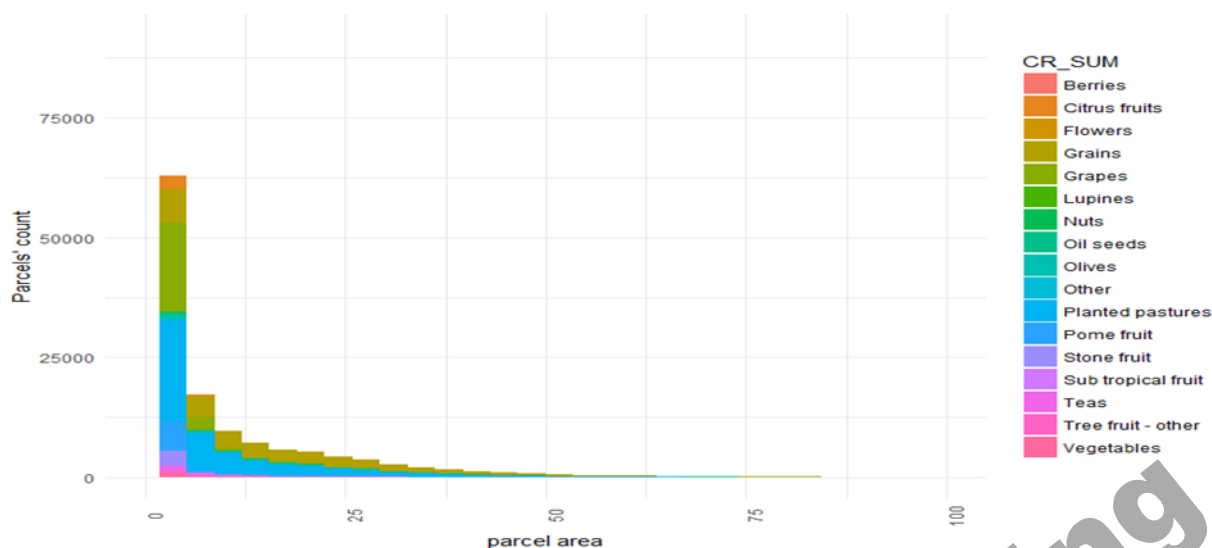


Figure 3-153: Number of parcels per crop Type and parcel area per Crop Type

3.3.4.4 Summary and conclusions

The crop mask and crop type mapping using Sentinel-1 and -2 time series have been addressed in three different biogeographic regions (Central, West and African test sites). This was performed by using the Random Forest algorithm which was applied on several time features extracted from time series from Sentinel-1, -2 and combined approaches. Based on the specifications for a potential future Copernicus HRL on agriculture, methods have developed and proposed.

The tests presented in this study have also been applied on the larger prototyping sites of 6-9 Sentinel-2 tiles in the framework of WP 44. In both phases, the aim was to cover the whole growing season, however the definition of growing phase differed. In phase 1 it was tested to use an extended time window including also autumn months of the previous year. In phase 2, the time windows selection was more constraint, focusing on a shortened vegetation period within one year, but covering the main events/changes per season: starting with sowing, then covering the heterogeneous phase of vegetation peaks for most of the crop types and extending to the harvest seasons. The test focus on the multisensory benchmarking and, considering the time features selection, by means of a pixel based approaches, is aiming at the best tradeoff between cost-efficiency and highly accurate results.

The time interval algorithms are strongly affected by undetected clouds/cloud shadows as well as confusion with bright surfaces in the cloud mask. These algorithms present many artefacts and data gaps due to the short compositing period and the interval between images available in the time series. In phase 1, it was demonstrated that feature-based algorithms are more appropriate as they achieve more spatial consistency and very few data gaps thanks to the use of the entire time series as input. To avoid potential artefacts derived from the presence of unmasked cloud cover pixels, time periods were selected to guarantee a sufficient number of imagery to minimize the distortion that extreme values would pose to the statistics. In the West test site, it was proved that 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.

For the **crop mask**, the accuracies of the classifications based on S2 were significantly higher than those based on S1. The combined approach of S1 and S2 increases the accuracies of crop/non-crop classification only marginally. As a consequence, and in order to reduce the computational effort, the input data for the crop mask classification of 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 is viable and promising.

Concerning **pixel level versus field level approaches**, the accuracy gain of the evaluation based on field level in phase 1, turned out to be much higher for S1 compared to the modest gains for S2. The strong increase in accuracy for the S1 based classifications on field level is expectable due to the reduction of the speckle effect. Despite the multi-temporal filtering carried out for Sentinel-1 data prior to the classification, 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.

Even though the time feature approach is able to compensate to a certain extent implications of extreme values (caused by land cover or by technical issues) they still affect the classification. Undetected clouds and cloud shadows as well as bright surfaces cause confusions and have impact on the **time interval algorithms**. These algorithms then present artefacts and data gaps which strongly increase for very short time periods. This experience led to the extension of the time window from 2 to 3 months in phase 2, compared to phase 1, and was also the reason for adding of the time window for the whole vegetation period from Mid-March to Mid-Oct supplementary to the already chosen ones for spring, summer and autumn. In both phases it turned out that feature-based algorithms are appropriate as they achieve high spatial consistency and very few data gaps thanks to the use of the entire time series as input and thus are able to minimize these artefacts. Thus, the use of extended time periods in combination with the time feature approach is a means to reduce unwanted effects.

Due to the high computational cost of the **feature calculation** for the whole raster, it has been investigated if the number of features can be reduced significantly without a significant loss of classification accuracy. The method of group based forward feature selection proved that the optimal accuracy can already be achieved by using a selection of best features (up to 25% of the initial feature set) in the tests for the Central site.

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-based feature selection as it has been performed in phase 2, could further reduce the processing cost without loss of accuracy.

As for the **crop type mask**, the tests in the Central site using the combined approach offers its full potential in joining the benefits of the S1 and those of the S2 time features. The ability of S1 time features to grasp texture and structure during the growing of the crops highly enhanced the crop type accuracy. Precondition is a suitable time window covering the whole vegetation period for the region to classify. Adaptions to regionally varying conditions should be done by stratification, taking biogeographic conditions like temperature, precipitation and altitude into account. This approach would minimize the confusion between crop types caused by shifted sowing and harvesting dates or shortened vegetation periods in areas with higher altitude. The focus of the stratification is the same than that of the crop type class nomenclature: homogeneous preconditions promote an accurate detection of the crop area respectively of the crop types and will lead to an accurate classification result.

From the tests results, it was found that the classes with high occurrence and number of samples, tend to reach high accuracies, whereas classes with low occurrence, show low performances. However, although this must be taken into account, also other aspects have an impact on differentiating crop types from each other. Distinct spectral characteristics, phenology and beginning and end of growing phase are also very important for crop type detection and for differentiation between the crop types. This was shown in phase 1 already, and is confirmed in WP41 and WP44 in both phases. In some cases, (as for class 21 and 22 in Central), the small number of samples is not that crucial when comparing the F1 scores.

The multisensory approach confirms the great performance of S2 for crop masking and crop types mapping. The combined S1+S2 tests provide the highest accuracies, although in case of the crop mask, the gain is marginal, not accounting for the factor that the benefits are increased when cloud conditions are worse. This might be the case proved in the benchmarking of the test for 2018 in Central in phase 2 for the crop types mapping. In the end, the cloud conditions must also be considered because it affects the time series density, and accordingly the time windows selection. As explained in grasslands, selecting the key phenological periods can enhance cost-efficiency of production, although especially in agriculture, due to the relevance of the S2 features, the periods should not be too short, to reduce the risk of lack of a sufficient number of scenes or quality of the imagery in the time series.

For the crop type classification an initial set of criteria to evaluate the best compositing method for crop detection and crop growth monitoring (CG) have been selected. The benchmark was performed on the test site Central (Germany/Switzerland/France) and test site WEST(Belgium) and showed promising accuracies as well as the high potential of time series and derived time features for crop mask extraction and crop type monitoring.

In Central tests, LUCAS data from 2018 constitute the main part of the **sample base** for the crop mask. This sample base has been complemented by samples for forest, grassland, water and urban areas taken from HRL2015 and by manual samples for specific land cover such as orchards. The best way of dealing with small/underrepresented crop type classes has still to be discussed as it will be a fundamental issue for a planned roll out covering large areas or even a Pan-European one. The SMOTE approach would be one option reduction of classes could be the other.

The availability and representation of crop and specifically crop type samples is essential for the model training and significantly impacts the performance and quality. When considering the crop types, more testing should be done when it comes to the differentiation of similar crop types. With a view to the planned implementation of a future agricultural HRL future it is highly recommended to analyze the regional diversity, local phenological conditions and crop types occurrences and to think about stratification.

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.

Concerning Agriculture classification an initial set of criteria to evaluate the best compositing method for crop recognition (cropland-CL, crop type-CT) and crop growth monitoring (CG) have been selected. The benchmark is performed on Central (Germany) and Belgium site and show promising accuracies and high potential of time series and derived time features for crop mask extraction and crop type monitoring. 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.

The availability and representation of crop and specifically crop types samples is essential for the model training and significantly impacts the performance and quality. When considering the crop types, more testing should be done when it comes to the differentiation of similar crop types, as well as regional

diversity for implementation of future agricultural HRL. In the case of agriculture, to adapt to local phenological conditions and crop types occurrences, stratification might be required.

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.

In phase 2, lessons learned from testing that were subsequently applied in the prototypes are related to the need of stratification to adapt to the different conditions, including altitudinal ranges, to improve the classifications and minimize mixing between some classes. Additional manual sampling has improved the classification as it is clear that the LUCAS 2018 points are not enough. The SMOTE algorithm also improved significantly the mapping of less frequent classes as shown in the Western site. The integration of auxiliary information (e.g., DEM) enhances the products quality significantly, especially in higher altitudes. The high OA (> 90%) for the crop mask classification looks very promising for future applications, e.g., like a future HRL on Cropland. Similarly, the promising results obtained in South-Africa for the main crop types in spite of the large diversity of crops.

3.3.5 New land cover products

In the frame of ECoLaSS, one thematic focus is laid on the testing and production of a prototype on New Land Cover (NLC). In this context, two different categories of prototypes are created. One of them focuses on a CLC-related classification whereas the other is a combination of the HRLs 2015 with the available Crop Masks produced during the first phase of ECoLaSS. Within this Deliverable, the methods for the CLC-related prototype are described in the following sections. However, since the HRL combined layer is not based on an extra classification but on a combination of already existing products, the methods applied in this context are not included in this current Deliverable but in the section on Experimental Setup in the referring prototype report (*D45.1b – Prototype Report New LCLU Products (Issue 2)*).

Work exposed in this section is almost entirely based on the annex document provided with the very recent CLC+ Backbone ITT (EEA, 2019). Main characteristics of the two main products of CLC+ Backbone can be summarized as follows:

- A raster product, pixel-based and derived from multi-temporal S-2 input imagery, at 10 by 10m spatial resolution, with a set of basic land cover classes;
- A vector product, with a 0.5ha MMU, object-based and derived from ancillary data's linear networks and multi-temporal image segmentation, whose attribute classes are derived from the zonal statistics taken from the previously mentioned raster and complemented by additional characteristics from VHR and S-1.

Four major steps are laid out in order to produce both products:

- First step – Level 1: creation of a geometric skeleton derived from persistent features (called “hard bones”, i.e. stable borders based on artificial or natural linear features – represented by polygons of roads, railways and rivers) in the landscape – such as roads, railways, rivers;
- Second step – Level 2: inside the skeleton underlined by Level 1 objects, Level 2 polygons are derived using image segmentation (called therefore “soft bones”) and based on mono-temporal VHR data as well as multi-temporal S-2 data (complemented with L-8 data if need be) – those soft bones represent spectrally and/or texturally homogeneous features in the shape of polygons with coherent temporal variation during the year, e.g.:

- Land cover units with a unique vegetation cover/surface property and homogeneous dynamics throughout the year;
- Identification of single field parcels – agricultural field structure can differ only in terms of land cover dynamics over time.
- Third step: production of an independent pixel-based land cover classification at 10m spatial resolution from multi-temporal S-2 dataset;
- Fourth step: Characterization of all Level 2 polygons using spectral, textural and backscatter characteristics from S-2, VHR, and S-1 datasets, using statistical information retrieved from the pixel-based land cover classification from the third step.

Details on the creation of final and intermediate product can be found in Table 3-80.

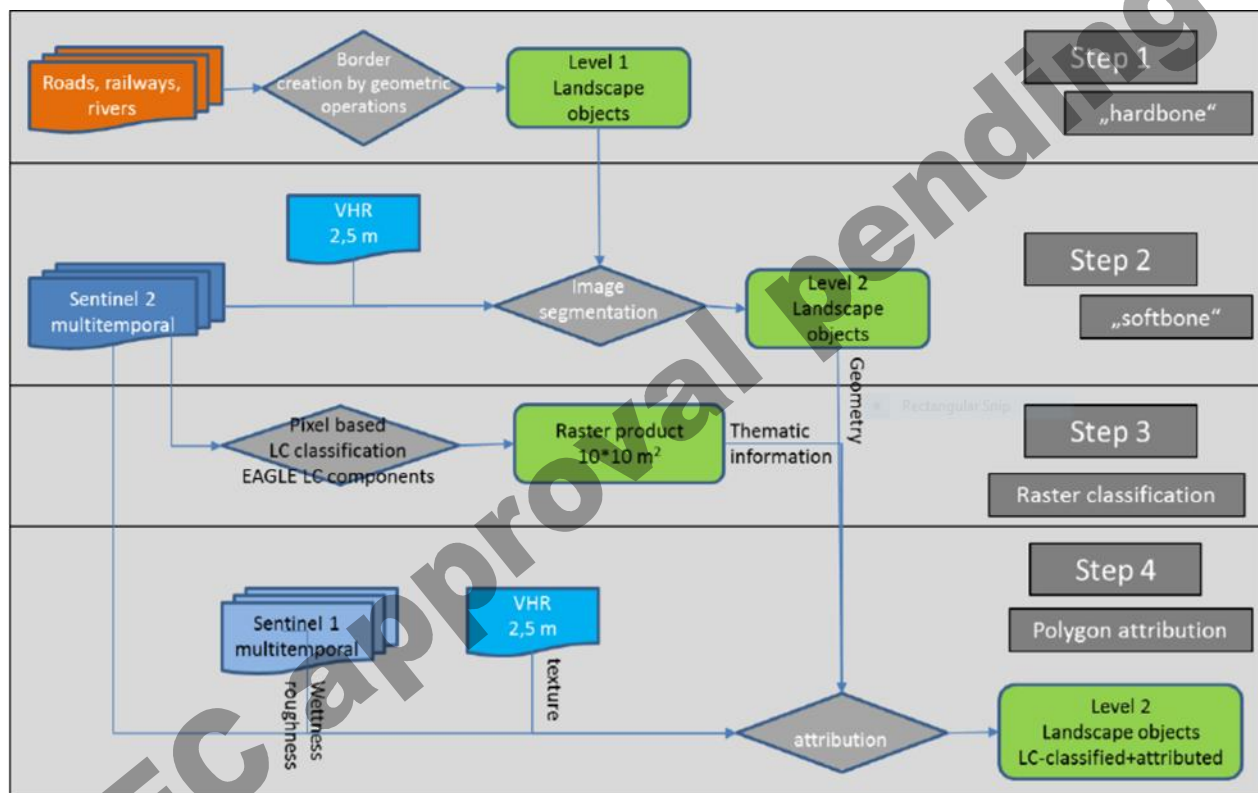


Figure 3-154: Proposed automated approach in the ITT for CLC+ Backbone (EEA, 2019).

Table 3-80:- Technical details for intermediate and final products created for the CLC+ Backbone.

	Hard bones	Soft Bones	Raster product	Vector product
Reference year	No older than 1-3 years	10/2017-09/2018 Could include 2017 and 2019 imagery		
Input data	For roads and Railways: <ul style="list-style-type: none"> - Open Street Map - National Reference Data For rivers and water bodies: <ul style="list-style-type: none"> - EU-Hydro - WISE WFD 	Sentinel-2 time series with VHR mono-temporal complement	Sentinel-2 time series Complemented over cloudy areas by Sentinel-1	Fusion of {Hard Bones and Soft Bones} Raster product Attributes from S-1 time series Attributes of VHR
MMU	-	0.5 ha	10m*10m pixel, 100m2	0.5 ha
MMW	-		10m	20m (more or less 10m)
Accuracy	<ul style="list-style-type: none"> - More or less 5m positional accuracy - Snapping tolerance 10m - No dangles - 95% of network completeness (from high ways to agricultural tracks) 	<ul style="list-style-type: none"> - Position accuracy of more or less 10m - Appropriate size: <ul style="list-style-type: none"> • Too large polygons: less than 10% of all polygons • Too small polygons: less than 15% of all polygons - Shift of border: 20m maximum 	<ul style="list-style-type: none"> - Land cover classes: 90% overall accuracy - Omission errors max 15% - Commission errors max 15% 	Same as Soft Bones
Nomenclature	<ul style="list-style-type: none"> - For roads and railways: all road types - For rivers: all rivers a drainage basin larger than 10km2 	None	<ol style="list-style-type: none"> 1. Sealed (Buildings and flat sealed surfaces) 2. Woody needle leaved trees 3. Woody Broadleaved Deciduous 4. Woody Broadleaved evergreen 5. Woody shrub 6. Permanent herbaceous (grassland) 7. Periodically herbaceous (arable land) 	<ol style="list-style-type: none"> 1. Sealed (Buildings and flat sealed surfaces) <ol style="list-style-type: none"> 11. Very high sealing degree (> 80%) 12. High sealing degree (50-80%) 2. Woody needle leaved trees <ol style="list-style-type: none"> 21. pure needle-leaved (> 75%) 22. dominantly needle leaved (50-75%) 3. Woody Broadleaved Deciduous <ol style="list-style-type: none"> 31. pure broadleaved > 75%: 311. pure deciduous >

			<p>8. Lichens and mosses 9. Sparsely vegetated 10. Non vegetated (Bare rocks, scree, sand, lichen, permanent bare soils) 11. Water 12. Snow and Ice</p>	<p>75% 312. pure evergreen > 75% 32. dominantly broadleaved 50-75% 4. Shrubland (> 50%) 5. Permanent herbaceous land (grassland, > 50%) 51. woody trees < 10% 52. woody trees 10-30% 53. woody trees 30-50% 6. Periodically herbaceous (arable land, > 50%) 7. Lichens and mosses land (> 50%) 8. Partly vegetated 81. intermediate vegetation cover 30-50% 82. low vegetation cover 10-30% 9. Non vegetated (Bare rocks, scree, sand, lichen, permanent bare soils, > 90%) 10. Water 11. Snow and Ice</p>
Attributes	- For roads and railways: information on the count of tunnels per line string	None		<ul style="list-style-type: none"> - The area percentage of the individual pixel-based land cover classes - From S-1: Wetness (5 classes), Roughness (5 classes) - From VHR: texture parameter - Spectral attributes: statistical mean and standard deviation for S-2 bands - Spectral indicators: NDVI, LAI (mean and variation across observation period)

3.3.5.1 Description of candidate methods

We explore here the fulfilment of each step detailed in the previous section.

“HARD BONES” CREATION

The **“hard bones” creation** uses the following datasets for permanent linear delineations, the so-called “hard bones” of the landscape:

- Open Street Map for road and railway shapes;
- EU-Hydro and WISE- for rivers and canals.

Those linear networks are meant to be represented in those hard bones layer as polygons. However, a discrimination between elements wider than 20m and narrower than 20m is introduced in the ITT. This kind of information, such as an average width, cannot be retrieved from the ancillary databases selected. Therefore, this action would be derived from remote sensing data – and for finer accuracy, would be derived from VHR images, not S-1 nor S-2 time series.

This work has been deemed to be outside the scope of ECoLaSS – since it not only calls for a heavy use of VHR images over both demonstration sites but also for a heavy manual work to assess this average width along roads, railways and rivers.

In order to stay as close as possible to the CLC+ requirements, selection among what is expected to be the widest road and railway kinds has been done in OSM, integrating in the hard bones layer:

- OSM railways class called ‘rails’;
- OSM roads classes, namely ‘motorway’, ‘motorway_link’, ‘trunk’, ‘trunk_link’, ‘primary’ and ‘primary_link’.

Data from EU-Hydro has been integrated, as well as main rivers from OSM. The WISE dataset, once downloaded, turned out to be corrupted for the year 2016. A 2012 dataset is available, but falls off the range of the ancillary data in term of temporal fit. The retained classes integrated into the hard bones layer are:

- OSM class named ‘water_a’ to retrieve water bodies in the shape of polygons;
- EU-Hydro classes named ‘canals_p’ and ‘rivers_net_p’ for polygons only.

However, the class ‘water_a’ also contains lakes – the OSM class ‘waterways’, which is only composed of purely linear objects, is used to remove all non-linear water objects. To complement this clean-up, the EU-Hydro ‘InLandWater’ is also involved in this intersection: elements falling outside of it are removed, leaving only polygons of linear water networks.

Therefore, methodological choices could be summarized as:

- OSM has been chosen over the South-West site (FR and ES) and the Central site (FR, DE, CH, AU) regarding the integration of roads and railways;
- OSM coupled with Eu-Hydro over both sites for water bodies and rivers.

Polygons are integrated into the hard bones layer in their current form, removed if no close linear network section exit in the vicinity or fused when possible if they fall below the expected MMU of 0.5ha.

Those resulting polygons are expected to be further subdivided in the second step using automated image segmentation.

“SOFT BONES” CREATION

The “**soft bones**” creation refers to the discrimination between objects that behaves temporally and spectrally differently during the year, which is encapsulated in a vector layer over the raster times series fed as input into the algorithm.

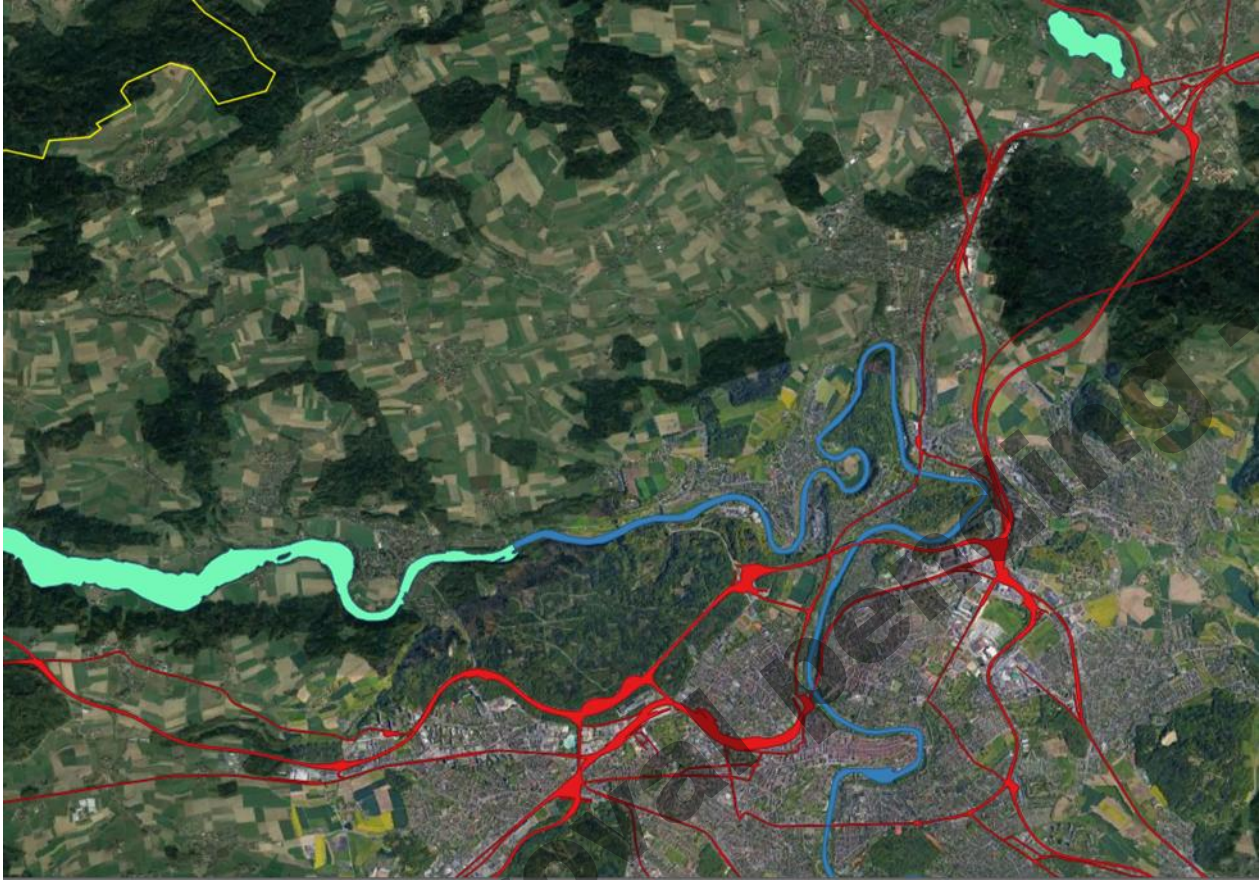


Figure 3-155: In red: OSM roads and railways; in dark blue: EU-Hydro; in light blue: OSM water classes.

This is a delicate step that requires to find a balance between too many segments and too few, in order to avoid a “salt and pepper” effect. Multi-scale approach can identify features at appropriate scales.

There is no general definition for the act of segmenting an image.

The characteristics of the input data used for the segmentation, the interpolated S-2 time series images, are detailed in the section 3.2, in order to avoid the potential disruption in the time series created by masked clouds and the associated lost of signal for a given date.

It has been decided to investigate several classical types of segmentations, based on the list of characteristics available in the Table 3-81:

- The SLIC algorithm;
- The Quickshift algorithm;
- The Watershed algorithm;
- The Large-Scale Mean Shift algorithm;
- The K-Means resulting from the Phenological intermediate products leading to the creation of the MPA layer and its derivatives.

- Simple Linear Iterative Clustering (SLIC) algorithm

This algorithm has been proposed to the community in 2010 (Achanta, et al., 2012) and is based on a local K -Means clustering (in the feature space that comprises color information as well as image location) which will generate a superpixel segmentation with K superpixels. SLIC is classified as a gradient ascent method and can be found implemented in the scikit-image library (van der Walt, et al., 2014).

A superpixel, concept proposed by Ren and Malik (Ren & Malik, 2003), is a local, coherent (in terms of similarity in color and spatial consistency) regrouping of pixels, which can also be defined as an aggregation of segments that “preserve most of the structure necessary for segmentation at the scale of interest”. The production of such an ensemble can be used as a pre-processing step for the real segmentation.

The SLIC algorithm randomly initializes cluster centers before redefining their location when pixels are associated to the nearest cluster center whose search region is overlapping its localisation. This step of assignment of pixels followed by redefinition of the centers is repeated until the error converges. A post-processing is required to enforce connectivity by assigning disjointed pixels to the closest superpixels.

Several parameters need to be fine-tuned for the algorithm to give the best results – some others are optional and are not relevant for our work here:

- Compactness – a float to balance color similarity and spatial proximity; higher values will give more weights to space proximity, making superpixels more cubic-like shaped.
 - Sigma – the width of Gaussian smoothing kernel executed as pre-processing for each dimension;
 - Max_iter – the maximal number of iterations for the K-means runs.
- Quickshift (QS) algorithm

The QS approach is a mode-seeking segmentation (Vedaldi & Soatto, 2008). It is an approximation of a kernelized mean-shift, meaning that instead of iteratively moving each point closer to a local mean, the algorithm forms a tree of links and moves each point representation into the feature space (of color and image location information) to the nearest neighbour in order to increase the kernel density estimate. No control over the size of segments is available. It is closely related to the SLIC algorithm. One of the advantages of this segmentation is the hierarchical computation executed at multiple scales simultaneously.

Here are the two main parameters, that are relevant for our soft bones:

- Ratio – a float to balance color similarity and spatial proximity; higher values will give more weights to color proximity, as opposed to the compactness parameter for SLIC;
- Sigma – the width of Gaussian smoothing kernel executed as pre-processing for each dimension;
- Kernel_size – the width of the Gaussian kernel used in smoothing the sample density; it controls the scale of the local approximation and higher values mean fewer clusters;
- Max_dist – the cut-off point for data distances; higher values also mean fewer clusters.

Table 3-81: - Characteristics of classical segmentation algorithms, available in open-source libraries.

Segmentation	Multi-spectral handling	Multi-temporal handling	Remarks	Open-Source availability	Multi-threading	Complexity	Investigated in ECoLaSS
Random Walker	Yes	Yes	Requires manual markers set by an operator	scikit-image	Yes	$O(N)$, where N the number of pixels	No
Active Contour	Yes	No	Requires manual points set by an operator	scikit-image	Yes	$O(N^2)$	No
Felzenswalb and Huttenlocher	Yes	No	Tends to oversegment	scikit-image eo-learn	Yes	$O(N \log(N))$	No
Simple Linear Iterative Clustering	Yes	Yes	Tends to oversegment; needs a post-processing step	scikit-image eo-learn	Yes	$O(K)$ where K is the number of expected superpixels	Yes
Quickshift	Yes	Yes	Tends to oversegment	scikit-image	Yes	$O(d N^2)$ where d is a small constant	Yes
Watershed	Yes	Yes	Tends to oversegment; Works better with manual markers set by an operator; Use of markers optional	scikit-image OTB	Yes Yes	$O(N \log(N))$	Yes
Chan-Vese Segmentation	No – grayscale only	No	-	scikit-image	Yes	$O(N)$	No
Morphological Geodesic Active Contours (MorphGAC)	No – grayscale only	?	Requires pre-processing: Inverse Gaussian Gradient to highlight contours	scikit-image	Yes	$< O(N)$	No
Morphological Active Contours without Edges (MorphACWE)	No – grayscale only	?	-	scikit-image	Yes	$< O(N)$	No
Large-Scale Meanshift Segmentation	Yes	Yes	Resource-consuming	OTB	Yes, partially	$O(N^2)$	Yes
Connected Components	No	Yes	Fractional land cover, depending on parameters to be chosen by operator	OTB	?	$O(E + V)$, where E is the number of edges and V of vertices	Yes

- Watershed algorithm

The traditional version of this algorithm is used mostly on gray-scale images (Beucher & Lantuejoul, 1979). It treats image like a topographic surface, using the gradient descent in order to artificially reproduce a flow gradually flooding the various regions of the image, usually starting from the minima, defined as sources. Those regions, forming catchment basins, define local geometrical structures of the image associated with one or several local extrema. The input data is usually a filtered version of the raw image. The algorithm strategy consists in treating the magnitude of pixels as a function f , describing height, assuming that higher (or lower) values of f (or $-f$) reveals the presence of natural boundaries on the original image. In the case of multi-spectral images, the considered height function is the gradient magnitude of the amplitude, i.e., the square root of the sum of squared bands.

The drawback of this algorithm lies in the production of a large quantity of regions, each associated with a local minimum – resulting in an over-segmentation. This can be partially alleviated by using a minimal watershed depth, where basins whose depths fall below this threshold are regrouped into one region.

In order to run smoothly over large areas, the algorithm implemented in OTB for the watershed segmentation tends to subdivide the AOI into smaller areas, and the segmentation is repeatedly run over those artificially isolated areas. The spectral discrimination is therefore only based on the spectral signature of pixels present in the considered smaller area.

Two parameters can be adjusted in the classical version of this algorithm:

- Depth threshold – expressed in percentage of the maximum depth of the image, which is here the maximal difference of reflectance;
- Flood level – float between 0 and 1, it is used to generate the merge tree from the initial segmentation.

- Large-Scale Mean Shift (LSMS) algorithm

This particular algorithm was selected during the production of the HRL 2015 Grassland for its robustness and multi-threaded execution mode. It is based on an iterative mode-seeking procedure, such as the QS method, focused on finding the local maximum of a density function, in order to form modes in color or intensity feature space of the considered image. This is a classical image segmentation method which produces irregularly shaped segment with no uniform size or minimal size and a post-processing is needed to enforce a minimal size, where cluster of pixels below the minimal size are regrouped with a nearby cluster whose spectral signature is the closest to the one of the considered segment. The complexity of the algorithm (in $O(N^2)$) is making it quite slow to implement.

A pre-processing step needs be applied on the image in order to filter the spectral noise: the meanshift smoothing. For any given pixel of the image, its value is replaced by the average spectral signature of its neighbour pixels (determined by the spatial radius for this pre-processing step in particular). This procedure is repeated until the maximum of iterations or a given smoothing threshold is reached.

This algorithm features numerous parameters that can be adjusted to tailor the segmentation result:

- Spatial radius – threshold on the spatial distance to consider pixels belonging to the same segments; a good value is half of the spatial radius parameter that was used in the meanshift smoothing step;
- Range radius – threshold on the spectral Euclidian distance distance (see second issue of WP31 [AD06]);
- Maximum number of iterations – number of iterations to finish the smoothing preprocessing;
- Minimal size - minimal size of a delineated region, with smaller clusters merged with the neighbouring cluster whose radiometry is the closest.

Following the CLC+ ITT requirements, this minimal size should be set to 50 pixels, corresponding to the 0.5ha MMU expected for the vector final product.

It should be noted that a multi-scale implementation was envisioned at the end of Phase I for the LSMS segmentation. However, this processing was supposed to be based on the pixel-based classification and the zonal statistics. The idea was to look at the repartition of classes inside a first large polygon (produced with a LSMS whose minimum size was 500 pixels, e.g.) and if multiple classes were to be represented, the polygon should then be subdivided using an intermediate scale segmentation input (e.g. from LSMS whose minimum size was 200 pixels) into smaller polygons, until either the smallest scale of the produced segmentations was reached or the representativity of classes was uniformed with one strongly dominant class.

This methodology was not compliant with the latest requirement of the CLC+ ITT and will therefore not be explored in Phase II. In fact, the ITT requires a finer discrimination than the classes representativity for a given polygon. For example, various states of forest, due to different canopy heights or to different tree ages, should be delineated into different polygons in the final segmentation, even when adjacent. Urban buildings and flat artificial areas should also be delineated into the segmentation.

- Phenological Layers

The phenological layers are generated based on the NDVI time series, where each pixel of the demonstration site is classified into an “activity” class, depending on the behaviour of its NDVI values during the year. The methodology is detailed in section previous sections and further discussed in the report of W41 [AD10].

This k-means classification on this phenological layers allows the regrouping of coherent ensembles in the landscape, just like a segmentation on the time series would do, with a particular focus on the vegetation, matching the type of high level typology that is set for the raster classification. A majority filter of 0.5ha (corresponding to 50 S-2 pixels) is applied and the resulting raster is transformed into a vector layer. The geometry thus created is used as a soft bones layer.

Table 3-82: - Targeted typology, used typology over the test sites and matching with LUCAS, CLC and other ancillary datasets

Last version available of the CLC+ nomenclature	Adopted nomenclature on SW and Central test sites	Matching typology in LUCAS	Matching typology in UA	Matching typology in Riparian Zone/Coastal Zone/ Natura 2000	Matching typology in CLC	Matching typology in other ancillary databases
1. Sealed (Buildings and flat sealed surfaces)	Sealed Areas	A00 Artificial Land	11100 Continuous Urban Fabrics	1.1.1.1 Continuous urban fabric	1.1.1 Continuous urban fabric	HRL 2015 IMP status layer IMD > 5%
				1.1.1.2 Dense urban fabric		
			11300 Isolated Structures	-	1.1.2 Discontinuous urban fabric	
			12100 Industrial, commercial, public, military and private units	1.1.1.3 Industrial or commercial units	1.2.1 Industrial or commercial units	
			12210 Fast transit roads and associated land	1.2.1.1 Road network and associated land	1.2.2 Road and rail networks and associated land	
			12220 Other roads and associated land			
			12230 Railways and associated land			
			12300 Port Areas	1.2.1.3 Port areas	1.2.3 Port areas	
2. Woody needle leaved trees	Coniferous	C20 Coniferous Woodland	31000 Forests	3.2 Coniferous forest	3.1.2 Coniferous forest	HRL 2015 FOR>30% DLT status layer
3. Woody Broadleaved Deciduous	Broadleaves	C10 Broadleaved Woodland		3.1.3.1 Other natural & semi natural broadleaved forest	3.1.1 Broad-leaved forest	HRL 2015 FOR>30% DLT status layer
			3.1.5.1 Highly artificial broadleaved plantations			

4. Woody Broadleaved evergreen	Evergreen	CXX9 Broadleaved evergreen forest		3.1.4.1 Broadleaved evergreen forest	3.2.3 Sclerophyllous vegetation	-	
5. Woody shrub	Clear cuts/Woody shrub	D10 Shrubland with Sparse Tree Cover	32000 Herbaceous vegetation associations	5 Heathland and scrub	3.2.4 Transitional woodland-shrub	-	
		D20 Shrubland without Tree Cover			3.2.2 Moors and Heathlands		
6. Permanent herbaceous (grassland)	Grasslands	E00 - Grassland	32000 Herbaceous vegetation associations	4 Grassland	3.2.1 Natural grasslands	HRL 2015 GRA status layer	
					2.3.1 Pastures		
7. Periodically herbaceous (arable land)	Annual Crops	B00 - Cropland	21000 Arable land (annual crops)	2.1.1.1 Non-irrigated arable land	2.1.1 Non-irrigated arable land	-	
				2.1.3.1 Irrigated arable land and rice fields	2.1.2 Permanently irrigated land		
				2.1.4.1 Complex patterns of irrigated and non-irrigated arable land			
	Permanent Crops		22000 Permanent crops	-	2.2.1 Vineyards	Open Street Maps for permanent crops (vineyards and orchards)	
					2.2.2 Fruit trees and berry plantations		
					2.2.3 Olive groves		
8. Lichens and mosses	Lichens and Mosses	F30 Lichens and Moss	33000 Open spaces with little or no vegetations	7.2.1.2 Unexploited peat bog	4.1.2 Peatbogs	-	
9. Sparsely vegetated	Determined using zonal statistics in the final fusion		-	33000 Open spaces with little or no vegetations	6.1.1.1 Sparsely vegetated areas	3.3.3 Sparsely vegetated areas	-
10. Non-vegetated (Bare rocks, scree, sand,	Sand	F20 Sand	33000 Open spaces with little or no	6.2.1.1 Beaches	3.3.1 Beaches	-	
				6.2.1.2 Dunes			

lichen, permanent bare soils)			vegetations	6.2.1.3 River banks		
	Rocks	F10 Rocks and Stones		6.2.2.1 Bare rocks and rock debris	3.3.2 Bare rocks	
	Burnt areas	F40 Other bare soil		6.2.2.2 Burnt areas (except burnt forest)	3.3.4 Burnt areas	
11. Water	Water	G00 Water Areas	50000 Water	9 Rivers and lakes	5.1.1 Water courses	HRL WaW 2015
					5.1.2 Water bodies	
				8 Lagoons, coastal wetlands and estuaries	5.2.1 Coastal lagoons	
					5.2.2 Estuaries	
12. Snow and Ice	Snow and Ice	G50 Glaciers, Permanent Snow	33000 Open spaces with little or no vegetations		5.2.3 Sea and ocean	
				6.2.2.3 Glaciers and perpetual snow	3.3.5 Glaciers and Perpetual Snow	

- EC approval pending

PIXEL-BASED CLASSIFICATION RASTER CREATION

Based on the results from the previous phase of tests, the classification used here is a random forest algorithm, taking advantage of the full time series of images for the considered year as input. Details of the algorithm have already been presented in section 3.2.

Several features are computed and then added to the raw pre-processed data:

- The algorithm computes a linear interpolation between all dates in order to fill the gaps left by the cloud masks or the no-data mask, as detailed for the phenological products in section 3.2;
- For each of those dates, the algorithm computes several spectral indices, the NDVI, the NDWI as well as the brightness index.

The nomenclature is based on the EAGLE land cover component concept:

1. Sealed (Buildings and flat sealed surfaces): all impervious and sealed surfaces, covered with features of a certain height above ground or without;
2. Woody needle leaved trees: trees belonging to the botanical group Gymnospermae;
3. Woody Broadleaved Deciduous: trees belonging to the botanical group Angiospermae, and that are leafless during a given period of the year;
4. Woody Broadleaved evergreen: trees also belonging to the botanical group Angiospermae, but that remain green all year long;
5. Woody shrub: woody plants in shrub growth, with a height usually less than 8m;
6. Permanent herbaceous (grasslands): a continuous vegetation cover throughout the year, without the occurrence of bare soil – those areas are either unmanaged or extensively managed;
7. Periodically herbaceous (arable land): arable areas characterized by at least one LC change between bare soils and vegetated surface during the considered year;
8. Lichens and mosses: any type of lichens and mosses, mostly found in northern European tundra;
9. Sparsely vegetated: unstable areas, with a mix between bare soils and vegetated surfaces, whose percentages should be comprised between 10 and 50%;
10. Non-vegetated (bare rocks, scree, sand, lichen, permanent bare soils): consolidated or unconsolidated materials with less than 10% of vegetation;
11. Water: all water bodies, including natural or artificial, salt or fresh, running or still;
12. Snow and Ice: areas covered almost permanently with snow (90% of the observation time) or permanently with ice (100% of the observation time).

In order to select the samples used as input for the classification, ancillary datasets have been used for each test site. The matching between those datasets and the nomenclature selected for the CLC+ Core products are listed in Table 3-82.

It should be noted that the “sparsely vegetated” class has no clear immediate match in the LUCAS databases, at least without a defined percentage of possible vegetation cover. Several classes (woody shrub, lichens and mosses, sparsely vegetated and non-vegetated) will be quite tricky to sample for the training of the classifier. The extraction of spectrally pure samples will be challenging for those classes, because samples at pan-European level will mostly be provided by CLC2018, with a spatial resolution of 25ha – a manual pre-selection will be required. Due to their restricted representativity, classes such as lichens and mosses, turned out to have very few points in the LUCAS dataset.

FUSION AND POST-PROCESSING: FUSION OF HARD BONES AND SOFT BONES

Both hard bones and soft bones layers are vector layers. In order to merge both, a polygon union is executed, which will preserve all polygons from both layers. Polygons will then be further subdivided, for example the road polygons from the hard bones will be further divided into the segmentation polygons.

A snapping threshold can be set in order to avoid many sliver areas. Several thresholds need to be tested depending on the quality of the segmentation and the datasets of linear networks.

The hard bones layer will be snapped over the segmentation in order to maintain the pixel-size matching between the segmentation and the S-2 time series, i.e. only the polygons from the hard bones should “move” to be set over the S-2 pixels grid.

FUSION AND POST-PROCESSING: OBTENTION OF A CLC+ BACKBONE PROTOTYPE

The fourth and last step consists in the attribution of the delineated objects from the second step, using:

- Spectral characteristics of the land cover from S-2 data;
- Textural characteristics from VHR data;
- Backscatter characteristics tailored for the whole geographical cover from S-1 data:
 - o Wetness: Vey wet, wet, intermediate, dry, very dry;
 - o Roughness: Very rough, rough, intermediate, smooth, very smooth.

It has been decided for this project to focus on the characteristics of the LC provided by S-2 data, since the use of a full coverage of VHR images over each demonstration site is outside the scope of this project, as well as the production of wetness and roughness categorical classes derived from S-1 backscatter time series, whose skills required for the conception at such scales are not to be found inside the consortium.

The attribution of land cover to polygons of the fusion of hard and soft bones, from the pixel-based classification raster, is done using zonal statistics, which is a vector operation that list the number of pixels present in a given polygon for each class of the raster classification. For each polygon, percentage of LC coverage per class is then determined based on this pixel count, and attribution is made:

- To the dominant LC class;
- To the 3 most dominant LC classes;
- To a particular class given a percentage of LC present in the polygon.

For each class of the final vector product and their associated subclasses, rules can be written, as detailed in Table 3-83. The woody trees classes are defined with:

- The woody – needle leaved trees;
- The woody – broadleaved, deciduous trees;
- The woody – broadleaved, evergreen trees.

The vegetation classes from the raster classification are the following:

- The woody trees;
- The woody – shrubs;
- Permanent herbaceous;
- Periodically herbaceous;
- Lichens and mosses.

Table 3-83: - List of rules to populate the vector layer created by the fusion of hard and soft bones, using the raster classification.

Raster classes	Vector classes	Vector subclasses	Rules
1 - Sealed (buildings and flat sealed surfaces)	Built-up land	11. Very high sealing degree	Sealed > 80%
		12. High sealing degree	Sealed between 50 and 80%
2 - Woody – needle leaved trees	Woodland – needle leaved trees	21. Pure needle leaved	Needle leaved trees > 75% Woody trees >= 50%
		22. Dominantly needle leaved	Needle leaved trees between 50 and 75% Woody trees >= 50%
3 - Woody - broadleaved, deciduous trees	Woodland – broadleaved trees	31. Pure broadleaved	Broadleaved trees > 75% Woody trees >= 50%
Woody – broadleaved, evergreen trees		311. Pure deciduous	Deciduous trees > 75% Woody trees >= 50%
		312. Pure evergreen	Evergreen trees > 75% Woody trees >= 50%
		32. Dominantly broadleaved	Broadleaved trees between 50 and 75% Woody trees >= 50%
Woody – shrubs	Shrubland	-	Shrubs > 50%
Permanent herbaceous (i.e. grasslands)	Permanent herbaceous land (i.e. grasslands)	51. Without trees	Permanent herbaceous > 50% Woody trees <= 10%
		52. With few trees	Permanent herbaceous > 50% Woody trees between 10 and 30%
		53. With many trees	Permanent herbaceous > 50% Woody trees between 30 and 50%
Periodically herbaceous (i.e. arable land)	Periodically herbaceous land (i.e. arable land)	-	Periodically herbaceous >= 50%
Lichens and mosses	Lichens and mosses land		Lichens and mosses >= 50%
Sparsely vegetated	Partly vegetated land	81. Intermediate vegetation cover	Total of vegetation classes between 30 and 50% Non-vegetated >= 50%
		82. Low vegetation cover	Total of vegetation classes between 10 and 30% Non-vegetated >= 50%
Non-vegetated (i.e. rock, screes, sand, lichen, permanent bare soil)	Non-vegetated land (i.e. rock, screes, sand, lichen, permanent bare soil)		Non-vegetated >= 90%
Water	Water		Water >= 50%
Snow and ice	Snow and ice		Snow and ice >= 50%

3.3.5.2 Benchmarking criteria

The intermediate steps, namely the creation of the hard bones and the soft bones, will be assessed visually by a photo-interpreter, in order to sort out candidates and refine the final selection, in particular for the segmentation methodologies.

The choice of the RF algorithm for the raster classification has been guided by the phases 1 and 2 results for all other thematic classification – therefore no attempt to benchmark it with another classifier has been done.

The most common metrics to evaluate the geometric quality of a segmentation are the Hoover metrics (Hoover, et al., 1996) – they can quantify cases of:

- correct detection;
- over-segmentation;
- under-segmentation;
- missed detection.

However, they require a ground truth as input – meaning a segmentation for all given classes should have been produced manually over each test site. Due to the time-consuming process, this kind of benchmarking was not undergone. Segmentations are visually evaluated, while also taking into account means of improvement, resource-consumption as well as other issues already pointed out in the CLC+ ITT.

3.3.5.3 Implementation and results of benchmarking

HARD BONES

A visual examination is realized over the two test sites by superposing the hard bones with a cloud-free S-2 image during the vegetation season peak, in order to highlight roads, railways and water bodies.

In Figure 3-156 and Figure 3-157, the network from hard bones can be seen superposed to ERSI Imagery maps.

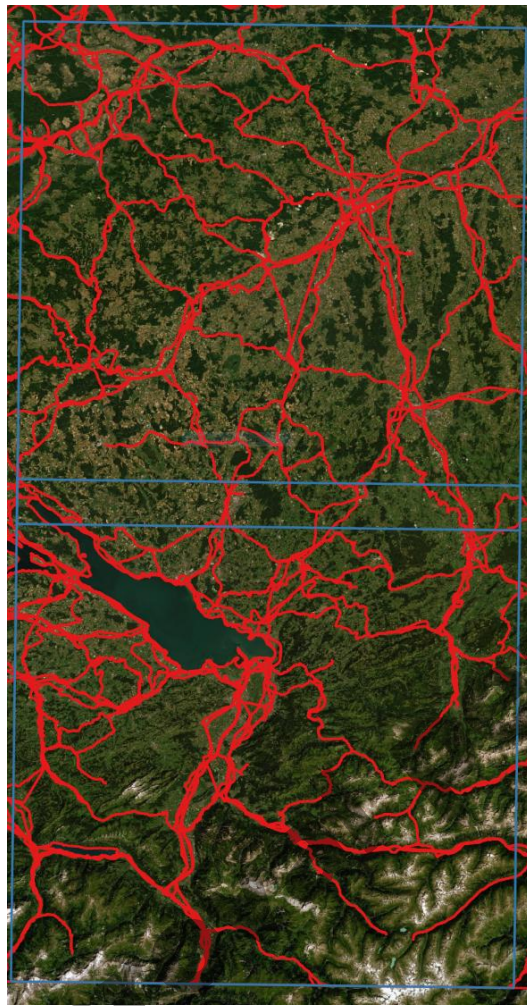


Figure 3-156 – Hard bones superposed over the S-2 tiles delineation (in blue) for the Central test site

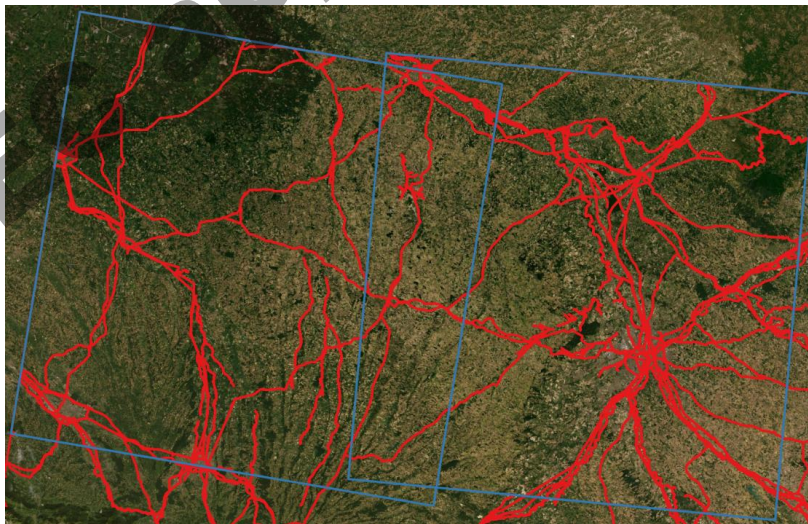


Figure 3-157 - Hard bones superposed over the S-2 tiles for the South-West tests site

SOFT BONES

The qualitative assessment starts with a first testing range based on logarithmic scales when possible. Depending on the selected best parameters, segmentations are run on a second testing range with a finer increment.

In order to increase contrast and reduce the amount of input data, the maximal values of each pixel, for each band, along the whole time series has been first computed – however, this statistical temporal metrics turned out to be contaminated by cloudy values, in particular over mountainous regions, where cloud detection algorithms experience issues to discriminate snow from clouds.

The mean computation of a manually selected ensemble of images free from those cloud mask issues was then fed to the statistical algorithm, in order to avoid anomaly that can be seen on Figure 3-158 and Figure 3-159.



Figure 3-158: Maximum values of all images over the year 2018 (tile T32TNT). On the lower right, lower values in dark grey are created by the gaps in swath trajectories. Dark grey squares come from the cloud detection algorithm of MAJA, while the light grey trace in the left corner is produced by undetected atmospheric veil.

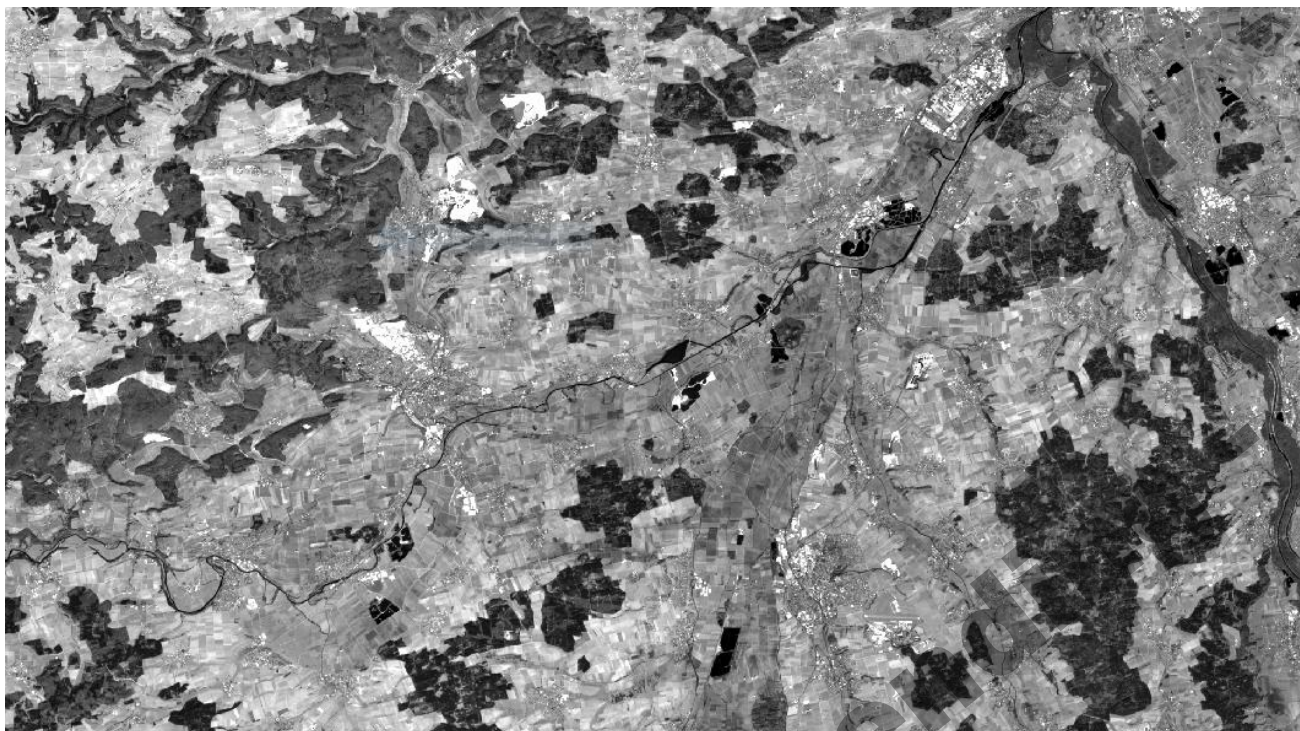


Figure 3-159: Mean values of a selection of the best scenes over the same area from the previous figure. Cloudy and snowy images have been removed from this restrained time series.

For the **LSMSS**, it is clear from the first testing range that higher values for the spatial and range radii produce to less segmented final vector layer – however it should be noted that this leads to longer and longer running times. The maximum number of iterations also considerably lengthens the process, even for a single image, but yields to very slightly better segments than the 5-iteration runs, mostly in the definition of water bodies and forest contours. The minimal size gives the best results and is kept as such in the second testing part, since it is in accordance with the MMU required by CLC+.

Table 3-84: - Parameters tested for the LSMS segmentations.

Parameters	First testing range	Best choice	Second testing range	Final choice
Spatial Radius	Smoothing: [1; 10; 100; 200]	200	[150; 200; 250]	150
	Segmentation: [0.5; 5; 50; 100]	100	[75; 100; 125]	75
Range Radius	Smoothing: [1; 10; 100; 200]	200	[150; 200; 250]	150
	Segmentation: [0.5; 5; 50; 100]	100	[75; 100; 125]	75
Minimal Size	[10; 25; 50]	50	50	50
Number of iterations	[5; 10; 20]	5	5	5

The best results, seen in Figure 3-160 and Figure 3-161, were obtained for high values in spatial and range radii, lengthening considerably the execution time, despite the built-in multi-threading mode that could take up to 56 processors in parallel – that unfortunately becomes impracticable for operational settings. This algorithm has been tested over a quarter of the T32TNT tile, and has already exhibited an incredibly resource-consuming behaviour, that is mainly due to the smoothing step, as well as the regrouping part to reach the set MMU – while the segmentation step in itself is rather quick. However, this smoothing step

ensures a spectral distribution of the reflectance values much closer to a Gaussian one and remains an imperative prerequisite.

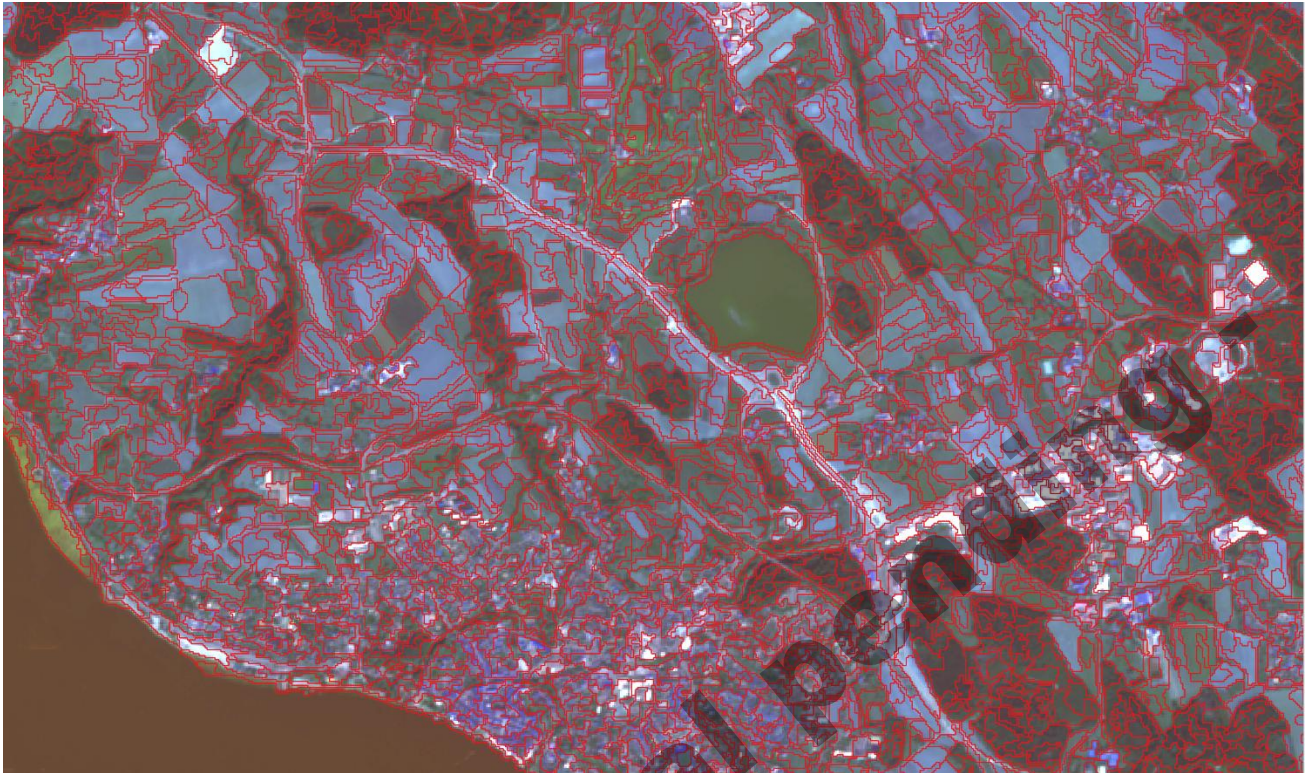


Figure 3-160: LSMSS as a vector layer with the best selected parameters dropped over the RGB bands (S-2 bands 2, 3 and 4) of the mean value raster of the selection of best scenes over the T32TNT.



Figure 3-161: For comparison, the LSMSS as a vector layer dropped over an ESRI Imagery VHR.

For **watershed** segmentation, a strong tendency to over-segment can be spotted for all choice of parameters, but some tiles fare better and coherent ensembles appear in the landscape for the first testing range. In the second one, with the flood level at 0.125, forests tend to be over-segmented, but agricultural parcels to be under-segmented, aggregating multiple fields together. The best results are reached at 0.0025 for the depth threshold and at 0.075 for the flood level, as seen in Table 3-85 but it should be noted that:

- Strange aggregations of agricultural fields are still visible;
- Strong over-segmentation is still present in the forest;
- There is a systematic over-segmentation of all roads;
- Multiples micro-polygons of less than 5 pixels are scattered over the landscape;
- There is no clear leveraging parameter on the minimal amount of pixels for the segments;
- There is a strong incoherence between tiles (produced automatically for multi-threading), i.e. the algorithm tends to over-segment or under-segment depending on the regions and the spectral signature of the given tiles.

The same area chosen for the LSMSS is displayed in Figure 3-162 and Figure 3-163. This last characteristic, however, calls for the dismissal of such algorithm. In fact, tiling will then clearly depend on the construction – and will not be reproducible over different material configuration in terms of memory and threads available. This should clearly be avoided in operational set-up. With the best set of parameters chosen, this discrepancy is less than visible than in the other set of parameters tested – yet this can still be spotted, as presented in Figure 3-164.

Table 3-85: Parameters tested for the classical watershed segmentation.

Parameters	First testing range	Best choice	Second testing range	Final choice
Depth threshold	[0.005; 0.01; 0.05; 0.1]	0.005	[0.0025; 0.005; 0.0075]	0.0025
Flood level	[0.01; 0.1; 0.2; 0.5]	0.1	[0.075; 0.1; 0.125]	0.0075

A final test has been carried out to try and tackle this discrepancy between tiles of the same image: as a pre-processing step, the smoothed image coming from the first part of the LSMS segmentation has also been used as input data. However, this does not resolve the under-segmentation of agricultural fields, even though a slight improvement of the micro-polygons could be seen, yet not sufficient.

This algorithm is therefore discarded for further testing after visual inspections.

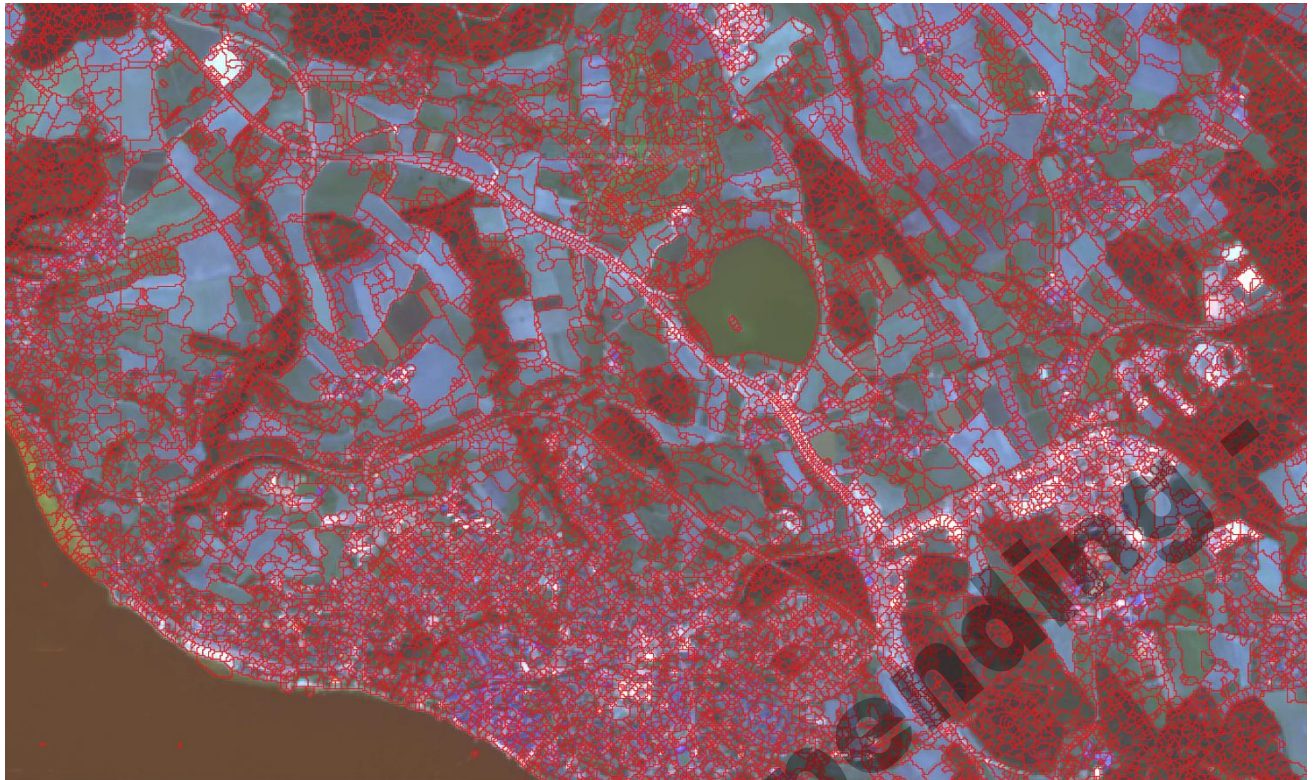


Figure 3-162: Watershed segmentation as a vector layer with the best selected parameters drapped over the RGB bands (S-2 bands 2, 3 and 4) of the mean value raster of the selection of best scenes over the T32TNT.



Figure 3-163: For comparison, the watershed segmentation as a vector layer drapped over the same ESRI Imagery VHR.



Figure 3-164: The separation between the two tiles generated by the watershed algorithm is in the middle of the image. On the left, over-segmentation of agricultural fields can be seen, while on the right, individual fields are quite well separated from other LC.

The SLIC algorithm leads to quick and rather visually coherent results, depending on the different parameters used, in particular the compactness. However, segmentation over water bodies is always over-segmented. Roads are clearly visible, but agricultural fields are also over-segmented, in particular between the borders of the fields and the inside of the field itself.

Table 3-86: Segmentation testing parameters

Parameters	First testing range	Final choice
Compactness	[0.01; 0.5; 1; 5; 10]	0.01
Sigma	$\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 10 \\ 10 \\ 10 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 10 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 10 \end{pmatrix}, \begin{pmatrix} 10 \\ 10 \\ 0 \end{pmatrix}, \begin{pmatrix} 10 \\ 10 \\ 1 \end{pmatrix} \right]$	$\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$
Number of iterations	[5; 50; 100]	100

The results are not really satisfying, unlike the LSMSS ones, but still better than the ones obtained with the watershed algorithm, as seen in Figure 3-165 and Figure 3-166.

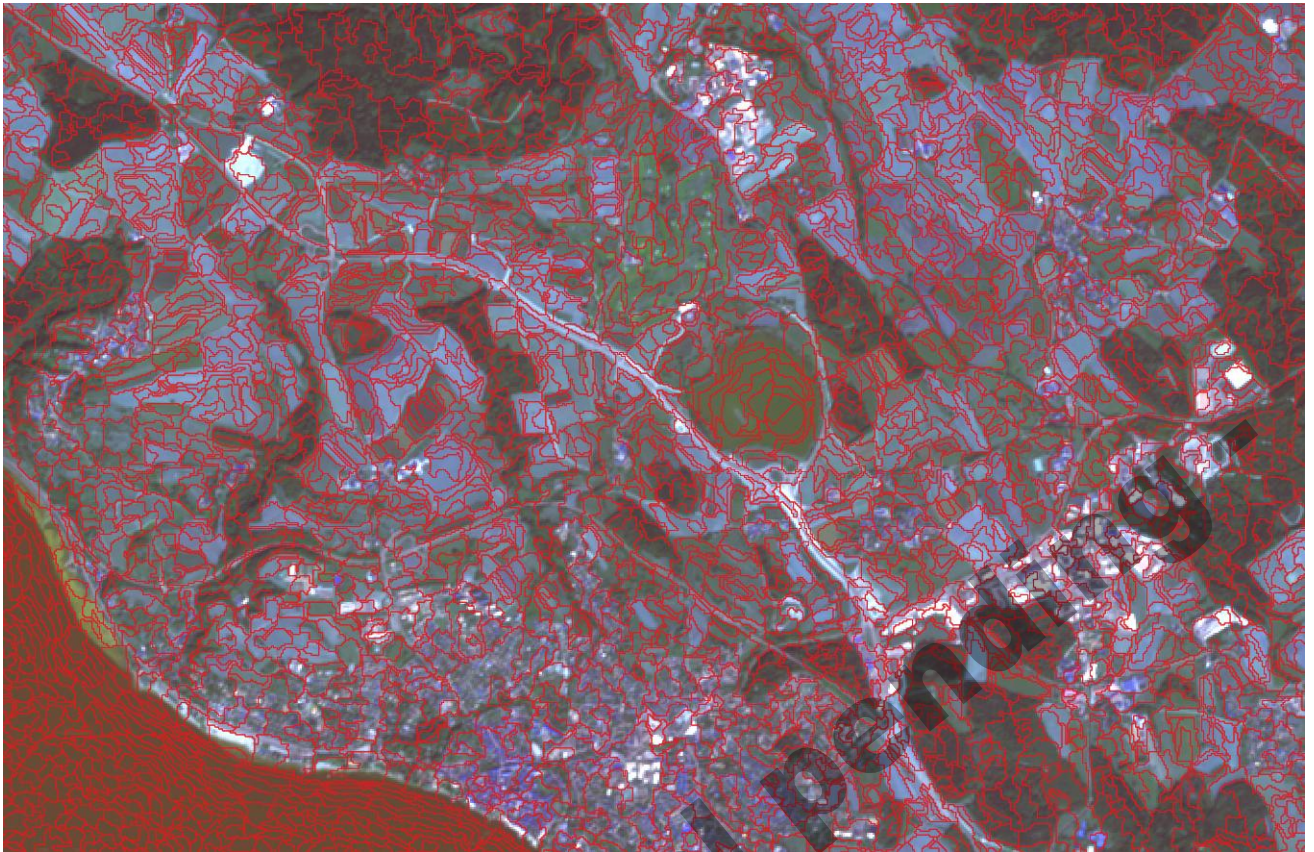


Figure 3-165: - SLIC segmentation as a vector layer with the best selected parameters drapped over the RGB bands (S-2 bands 2, 3 and 4) of the mean value raster of the selection of best scenes over the T32TNT.



Figure 3-166: For comparison, the SLIC segmentation as a vector layer drapped over the same ESRI Imagery VHR.

The phenological layers methodology is developed in the sections above and further detailed in the report of WP41 [AD10]. The methodology could be easily replicated from one year to the other, solving one of the issues raised by the CLC+ ITT. However, the overall aspect of the segmentation is not perfect, mostly in the urban areas. This could be improved by using a composed component algorithm targeting values of the NDVI indicating the presence of sealed areas. A few agricultural field are merged, due to the enforcement of the MMU, as can be seen in Figure 3-167 and Figure 3-168.

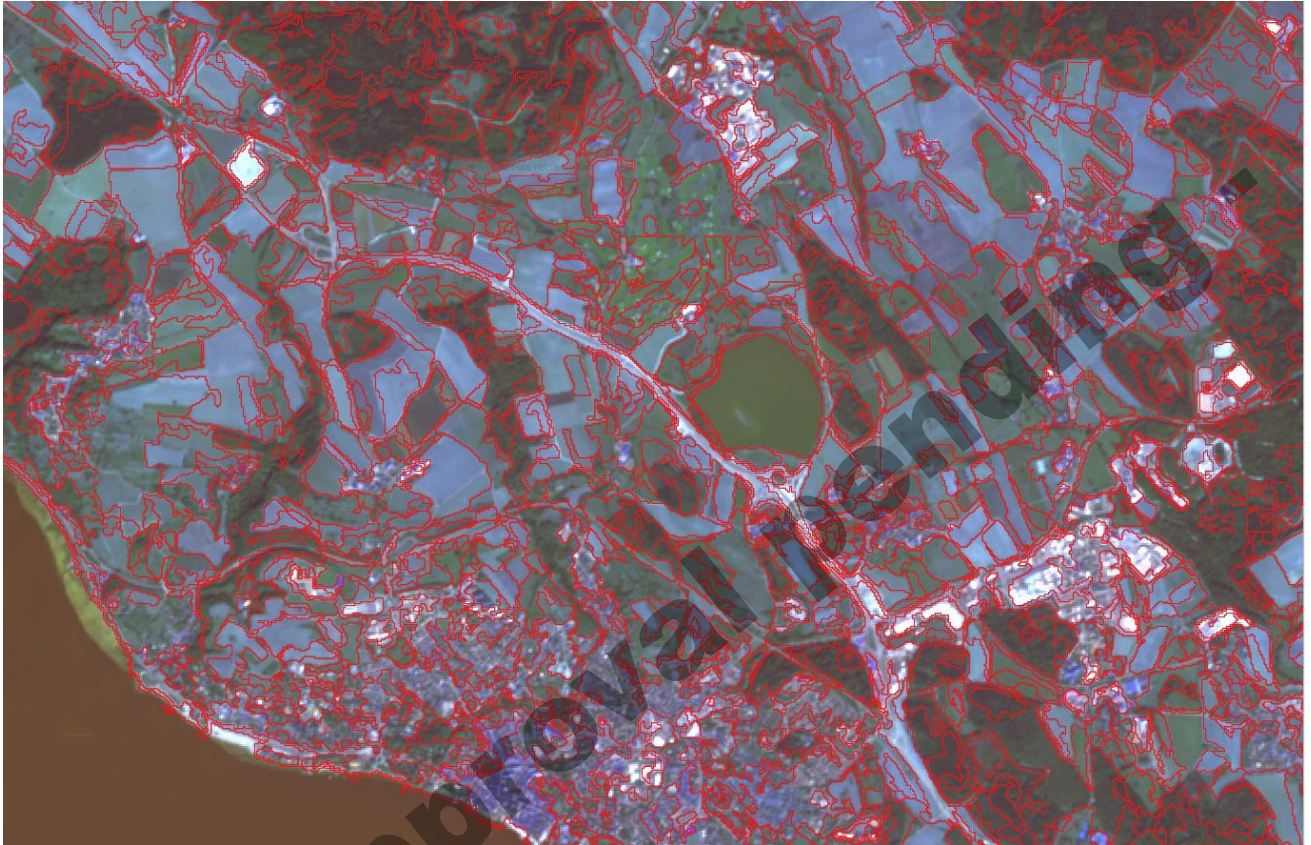


Figure 3-167: “Phenological segmentation” as a vector layer with the best selected parameters drapped over the RGB bands (S-2 bands 2, 3 and 4) of the mean value raster of the selection of best scenes over the T32TNT.



Figure 3-168: For comparison, the “phenological segmentation” as a vector layer draped over the same ESRI Imagery VHR.

The summarized results of those segmentations benchmarking can be found in Table 3-87.

Table 3-87: - Benchmarking for the segmentation algorithms.

Algorithms	Run time	Reproducibility on different hardware	Reproducibility from one year to the other	Required improvement if selected	Visual Appreciation
LSMSS	+++++	Ensured by the algorithm	Spectral and spatial smoothings rely on reflectance values, yet not as strongly as other algorithms	-	+++
Watershed	+	Not ensured	Parameters heavily rely on reflectance values range, thus depending on the region and weather	Enforcement of the MMU required	-
SLIC	++	Ensured by the algorithm	Compactness heavily relies on reflectance values range, thus depending on the region and weather	Enforcement of the MMU required	+
K-Means from phenological activities	+++	Ensured by the methodology	Ensured by the methodology	Enforcement of the MMU required	++

After careful reviews, the K-Means layer transformed as a vector layer is selected as the best trade-off between speed and correctness of the segmentation.

RANDOM FOREST CLASSIFICATION: TEST SITE SOUTH-WEST TILES (30TYP AND 31TCJ)

The repartition of the LUCAS points available, and used in the manual selection of sampling data, over the test site can be found in the Table 3-88.

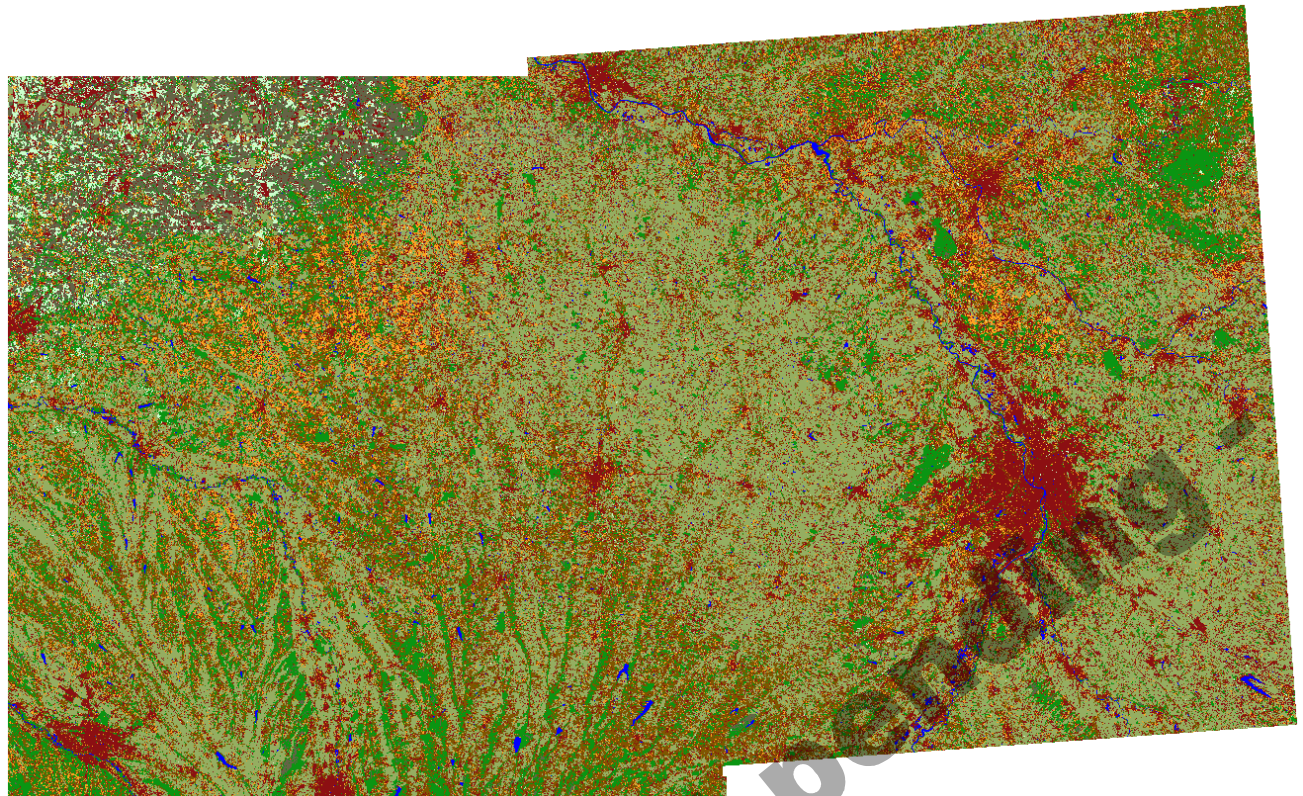
Table 3-88: - Available points in the LUCAS dataset from 2018, made available for ECoLaSS over the 31TCJ and 30TYP Sentinel-2 tiles.

Last version available of the nomenclature	Matching points over the test site in LUCAS	Matching polygons in CLC 2018
1. Sealed (Buildings and flat sealed surfaces)	110	1131
2. Woody needle leaved trees	39	375
3. Woody Broadleaved Deciduous	335	2440
4. Woody Broadleaved evergreen	0	0
5. Woody shrub	19	422
6. Permanent herbaceous (grassland)	370	1013
7. Periodically herbaceous (arable land)	633	1085
8. Lichens and mosses	0	-
9. Sparsely vegetated	-	0
10. Non-vegetated (Bare rocks, scree, sand, lichen, permanent bare soils)	29	2
11. Water	9	154
12. Snow and Ice	0	0

The best results are reached when at least 50 points are used per class, therefore excluding the LUCAS database as sole source for sampling selection. The use of random forest requires an equal number of points or polygons per class.

As seen in Table 3-88, woody shrub samples could be automatically selected from polygons in the 322 CLC class. However, since this class characterizes transitional land covers, at a minimal mapping unit of 25ha, the automated addition could potentially degrade the classification results if many evolutions have taken place in the landscape. It is also expected that the large spatial resolution may lead to the aggregation of mixed pixels, such as woodlands and bare soils, due to clear cuts, with potential shrub and forest regrowth resulting from older cuts, which would also lead to expected confusion in the classification results, due to the variability in the spectral and temporal signature.

There were not enough samples in the datasets of reference to create a class for the bare soils. However, when looking at the whole demonstration site, this class needs to be reintroduced. The classification can be found in the Figure 3-169 and the confusion matrix automatically generated in Table 3-89.



- Sealed areas
- Annual crops
- Permanent Crops
- Grasslands
- Woody shrub/Clear cuts
- Woody Broadleaved Deciduous
- Woody Needle leaved trees
- Water

Figure 3-169: Random Forest classification over T31TCJ and T30TYP

Permanent crops and annual crops were initially split, but it appears that some permanent crops could be linked to the forest classes (orchards e.g.) while other could be linked to woody shrubs (e.g. vineyards). This is resolved at the demonstration site scale in the report of WP45.

Table 3-89: Automatically generated confusion matrix for the test site in the South-West.

South-West test site		REFERENCE										
		Sealed Areas	Annual Crops	Permanent Crops	Grasslands	Woody Shrubs	Broadleaved Forest	Coniferous Forest	Water	Total	UA	
PRODUCT	Sealed Areas	11380	15	672	879	48	22	1	146	13163	0.86	
	Annual Crops	2367	84547	551	1727	34	104	1	0	89 331	0.95	
	Permanent Crops	1226	463	10277	1442	39	7	2	0	13456	0.76	
	Grasslands	1824	1599	1672	10535	87	117	0	0	15 834	0.67	
	Woody Shrubs	3616	392	21	19	11604	4	0	0	15 656	0.74	
	Broadleaved Forest	74	0	139	265	157	37454	300	0	38389	0.98	
	Coniferous Forest	20	0	3	2	1	1070	12424	0	13520	0.92	
	Water	211	177	17	17	1	0	1	17018	17442	0.98	
	Total	20178	87 193	13307	14886	11971	38 778	12 729	17164	217 791		
	PA	0.56	0.96	0.77	0.71	0.97	0.97	0.98	0.99			
											0.901	OA
											0.872	Kappa

RANDOM FOREST CLASSIFICATION TEST SITE: CENTRAL TILES (32TNT AND 32UNU)

The availability of data for sampling training can be seen in Table 3-90, with the repartition of the LUCAS points available, as well as the polygons from CLC2018.

Table 3-90: - Available points in the LUCAS dataset from 2018, made available for ECoLaSS over the T32TNT and T32UNU tiles.

Last version available of the nomenclature		Matching points over the test site in LUCAS	Matching polygons in CLC 2018
1.	Sealed (Buildings and flat sealed surfaces)	87	1859
2.	Woody needle leaved trees	117	1903
3.	Woody Broadleaved Deciduous	90	1043
4.	Woody Broadleaved evergreen	0	0
5.	Woody shrub	10	388
6.	Permanent herbaceous (grassland)	398	525
7.	Periodically herbaceous (arable land)	256	3612+363
8.	Lichens and mosses	0	-
9.	Sparsely vegetated	-	388
10.	Non vegetated (Bare rocks, scree, sand, lichen, permanent bare soils)	5	5+124+0
11.	Water	3	104
12.	Snow and Ice	0	21

The corresponding figure and table show the results of the tests carried out in the Central tiles.

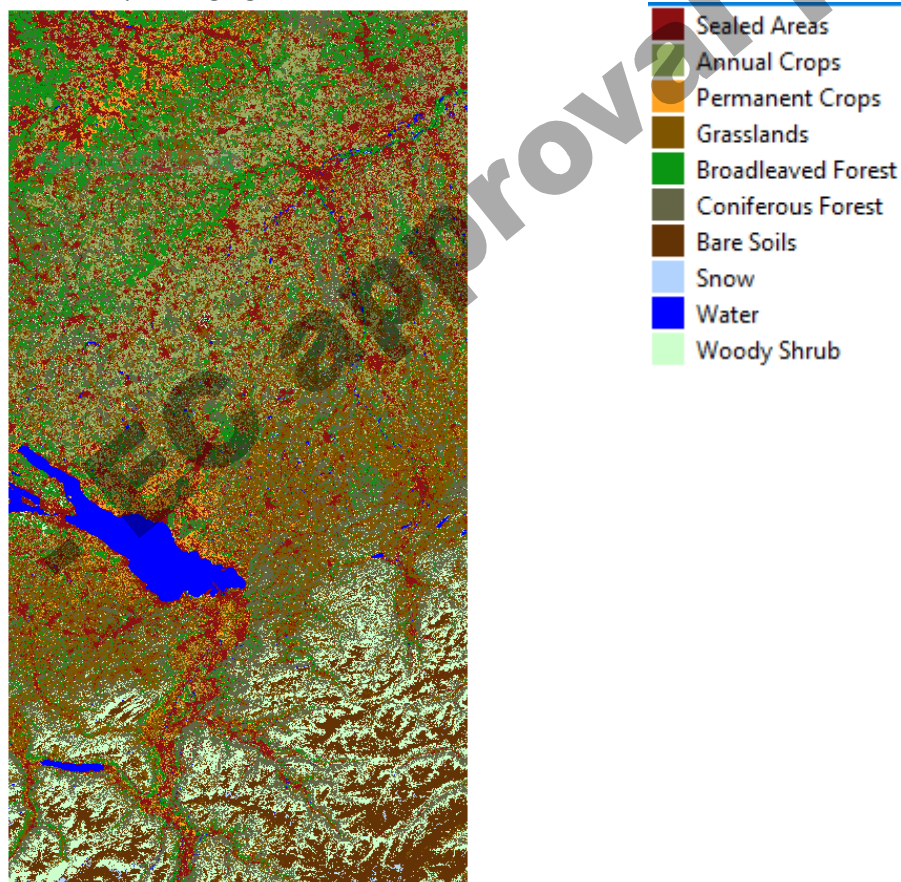


Figure 3-170: Raster classification from random forest algorithm over T32TNT and T32UNU

Table 3-91: Automatically generated confusion matrix for the test site in the Central test site.

Central test site		REFERENCE											
		Sealed Areas	Annual Crops	Permanent Crops	Grasslands	Broadleaved Forest	Coniferous Forest	Bare Soils	Snow	Water	Woody Shrub	Total	UA
PRODUCT	Sealed Areas	13127	53	336	454	61	17	23	23	8	12	14114	0.93
	Annual Crops	1679	29158	9	323	0	0	179	0	0	3	31351	0.93
	Permanent Crops	466	4	6523	278	128	46	0	0	0	41	7486	0.87
	Grasslands	620	967	990	34005	46	16	4598	0	0	541	41783	0.81
	Broadleaved Forest	225	61	201	175	46459	2128	0	0	0	565	49814	0.93
	Coniferous Forest	22	5	0	18	432	37983	0	0	0	285	38745	0.98
	Bare Soils	1165	20	0	22	0	5	63411	2129	148	9013	75913	0.83
	Snow	0	0	0	0	0	0	114	8345	0	0	8459	0.99
	Water	0	0	0	0	0	0	0	0	3139	0	3139	1
	Woody Shrub	1145	165	214	895	230	1063	1551	1	11	21716	26991	0.80
	Total	18449	30433	8273	36170	47356	41258	69876	10498	3306	32176		
PA		0.71	0.96	0.79	0.94	0.98	0.92	0.91	0.79	0.95	0.67		
												0.886	OA
												0.866	Kappa

3.3.5.4 Summary and conclusions

The segmentation is one of the major difficulties, that was already anticipated in the work document EEA/IDM/R0/17/003 (Kleeschulte, 2018). Aggregating large coherent ensembles that can present slight variations in their spectral signature while keeping the details small objects in the landscape is a major focus of this part of the methodology.

Several classes (woody shrub, lichens and mosses, sparsely vegetated and non-vegetated) will be quite tricky to sample for the training of the classifier. The extraction of spectrally pure samples will be challenging for those classes, in part because of the various land covers that can be categorized in them, and also because samples at pan-European level will mostly be provided by CLC2018, with a spatial resolution of 25ha.

- EC approval pending -

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 ECoSASS 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 Imperviousness, Forest and Grasslands, 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 intermediate product, the soft bones.

First, for Imperviousness, the analysis shows better results for a mono-temporal approach, the use of an active learning or SVM classifier and a 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. As shown on the demonstration sites in the WP42, the best approach is the combination of optical classification based on a selection of the best scenes (in order to avoid unwanted non-detected clouds that tarnish the results) with the use of temporal statistics for S-1 datasets.

Second, the potential of combining S-2 and S-1 data for the Forest delineation (tree cover and dominant leaf type mapping) is assessed by applying a random forest classifier to a number of experiments, using different combinations of sensors and time periods. Results of this analysis showed that 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.

However, this is always dependent on the data situation and increasing data volumes are naturally influencing the performance/cost ratio. Considering the lessons learned from phase 1, the integration of Sentinel-1 SAR into the TCM and DLT classification should not be completely discarded. Significant improvements in the tree cover detection could be achieved by the integration of SAR data (especially in agricultural areas and cloudy regions), whereas the added value of SAR time features for the leaf type discrimination is still insignificant. In view of the improved Tree Cover Density product at 10 m spatial resolution, which has been firstly tested and implemented in project phase 2, the ECoLaSS team has successfully demonstrated that median time features of the Sentinel-2 spectral bands are well suited for a consistent and seamless Tree Cover Density classification in high quality.

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 in phase 1 implemented in phase 2 are 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 and 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.

Finally, the New Land Cover classification benchmark, performed on the South West and Central sites, 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 test site and realizing a denser time series, to obtain more than just two seasons. In the second phase, the K-Means classification resulting from the Phenological intermediate products and leading to the creation of the MPA layer and its derivatives gives the best Softbone compromise between efficiency (Large scale Mean Shift segmentation method is low-efficient) and visual geometric accuracy. The raster classification, based on the results from phase 1 and phase 2 on all thematic classifications, has been chosen to be a random forest classifier for its efficiency and performance.

The ECoLaSS project follows a two-phased approach of two times 18 months duration. This deliverable comprises the second issue, containing all relevant updates concerning the benchmarking of input data for

classification as well as the time series classification methods and final results from the tests in WP33, aligned with the corresponding demos in Task 4 WPs.

- EC approval pending -

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. *Global Change Biology*, 2014.
- Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Susstrunk, S. (2012). Slc superpixels compared to state-of-the-art superpixel methods. *Pattern Analysis and Machine Intelligence, IEEE Transactions*, 34(11), 2274-2282.
- 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.
- Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, 9(5), 272.
- Ates, S. and Louahichi, M. (2012). Reflexions on Agro-pastoralists in the WANA region: challenges and future priorities. In: *New approaches for grassland research in a context of climate and socio-economic changes* by Z. Acar, A. López-Francos, C. Porqueddu (edtrs.). Publisher: Options Méditerranéennes, Série A (Séminaires Méditerranéens), No. 102 Zaragoza: CIHEAM / FAO / CITA-DGA.
- Bauer, M., B. Löffelholz & B. Wilson. 2007. Estimating and Mapping Impervious Surface Area by Regression Analysis of Landsat Imagery. In *Remote Sensing of Impervious Surfaces*.
- Belgiu M., Drăguț L. (2016). Random forest in remote sensing: A review of applications and future directions, *ISPRS Journal of Photogrammetry and Remote Sensing*, 114, 24–31.
- 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.
- Betbeder, J., Rapinel, S., Corgne, S., Pottier, E., and Hubert-Moy, L. (2015). TerraSAR-X dual-pol time-series for mapping of wetland vegetation. *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 107, 90-98.
- Beucher, S., & Lantuejoul, C. (1979). Use of watershed in Contour Detection. *International Workshop on image processing: Real-time Edge and Motion detection/stimation*. Rennes.
- 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.
- Breiman, L. (1984). *Classification and Regression Trees*. Chapman & Hall/CRC.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 4(1), 5-32.
- 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.
- 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.

- 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.
- Büttner G., Kosztra B., Maucha G., Pataki R. (2012): Implementation and achievements of CLC2006, ETC-LUSI, EEA, 65p.
- 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). *Kern methods for remote sensing data analysis*. John Wiley & Sons.
- Canu, S., Rosati, L., Fiori, M., Motroni, A., Filigheddu, R. and Farris, E. (2015). Bioclimate map of Sardinia (Italy). In: *Journal of Maps*, 11:5, 711-718, DOI: 10.1080/17445647.2014.988187.
- Carlinet, E., & Géraud, T. (2014). A Comparative Review of Component Tree Computation Algorithms. *IEEE Transactions on Image Processing*, 23(9), 3885-3895.
- Casas, J., Bonachela, S., Moyano, F., Fenoy, E. and Hernández, J. (2015). Agricultural Practices in the Mediterranean: A Case Study in Southern Spain. In: *The Mediterranean Diet*, Chapter 3. Elsevier Inc. 2015.
- Catorci, A., Ottaviani, G., Kosić, I. V., & Cesaretti, S. (2012). Effect of spatial and temporal patterns of stress and disturbance intensities in a sub-Mediterranean grassland. *Plant Biosystems-An International Journal Dealing with all Aspects of Plant Biology*, 146(2), 352-367.
- 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 30 m resolution: A POK-based operational approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 103(0):7 – 27. *Global Land Cover Mapping and Monitoring*.
- 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 Ressler, 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.
- Cong, N., Wang, T., Nan, H., Ma, Y., Wang, X., Myneni, R. B., & Piao, S. (2013). Changes in satellite-derived spring vegetation green-up date and its linkage to climate in China from 1982 to 2010: a multimethod analysis. *Global change biology*, 19(3), 881-891.
- Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote sensing of environment*, 37(1), 35-46
- Congalton, R.G. and Green, K. (2009) *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. 2nd Edition, Lewis Publishers, Boca Raton.
- Copernicus. (2019, 09 06). Retrieved from <https://land.copernicus.eu/pan-european/corine-land-cover>
- Copernicus. (2019, 09 06). Retrieved from <https://land.copernicus.eu/global/products/lc>
- 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.

- Cosentino, S. L., Porqueddu, C., Copani, V., Patanè, C., Testa, G., Scordia, D., & Melis, R. (2014). European grasslands overview: Mediterranean region. *The Future of European Grasslands*, 41
- Crawford, M., Tuia, D., & Ynag, H. (2013). Active Learning: Any Value for Classification of Remotely Sensed Data? *Proceedings of the IEEE*, 593-608.
- Crist, E. P. (1985). A TM tasseled cap equivalent transformation for reflectance factor data. *Remote Sensing of Environment*, 17(3), 301-306.
- Dams, J., Dujardin, J., Reggers, R., Bashir, I., Canters, F., and Batelaan, O. (2013). Mapping impervious surface change from remote sensing for hydrological modeling. *Journal of Hydrology*, 485, 84-95.
- 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 Analysis of 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 Beurs, K. M., & Henebry, G. M. (2010). Spatio-temporal statistical methods for modelling land surface phenology. In *Phenological research* (177-208). Springer, Dordrecht.
- 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.
- Defourny P., 2017. Land cover mapping and monitoring. In: *Handbook on remote sensing for agricultural statistics -FAO, GSARS*, p.21-58.
- Delincé J., 2015. Technical Report on Cost-Effectiveness of Remote Sensing for Agricultural Statistics in Developing and Emerging Economies. GSARS Technical Report: Rome. Available at: <http://gsars.org/en/technical-report-on-cost-effectiveness-of-remote-sensing-for-agricultural-statistics-in-developing-and-emerging-economies/>. Accessed 9 August 2017.
- 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.
- Duchemin, B., Maisongrande, P., Boulet, G., & Benhadj, I. (2008). A simple algorithm for yield estimates: Evaluation for semi-arid irrigated winter wheat monitored with green leaf area index. *Environmental Modelling & Software*, 23(7), 876-892.
- Dusseux, P., Vertès, F., Corpetti, T., Corgne, S., & Hubert-Moy, L. (2014). Agricultural practices in grasslands detected by spatial remote sensing. *Environmental monitoring and assessment*, 186(12), 8249-8265.
- EEA. (2019, June 04). Call for tender No EEA/DIS/R0/19/012. Annex 7, Technical specifications for implementation of a new land-monitoring concept based on EAGLE – D5: Design concept and CLC+ Backbone, technical specifications, CLC+ Core and CLC+ Instances draft specifications, including requirements review.
- Eklundh, L., & Olsson, L. (2003). Vegetation index trends for the African Sahel 1982–1999. *Geophysical Research Letters*, 30(8).
- 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 Services for Monitoring Dynamic Forest Disturbances to the Users – EOMonDis Project. Proceedings Book 7th edition of the International Scientific Conference ForestSAT 2016.
- Erasmî, 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.
- EUROSTAT (2016). Grassland areas, production and use. Lot 2. Methodological studies in the field of Agro-Environmental Indicators.
- Feranec, J., Jaffrain, G., Soukup, T., & Hazeu, G. (2010). Determining changes and flows in european landscapes 1990–2000 using corine land cover data. *Applied geography*, pp. 19-35.
- 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., Corbane, C., Ehrlich, D., Freire, S., Kemper, T., Maffenini, L., . . . Sabo, F. (2019). GHSL Data Package 2019.
- 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.
- Gallaun, H., Schardt, M., Linser, S., 2007. Remote sensing based forest map of Austria and derived environmental indicators. *Proceedings of the International Conference on Spatial Application Tools in Forestry (ForestSAT 2007)*. Montpellier.
- 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.
- Gallego, J., Peedell, S., & al., e. (n.d.). Using corine land cover to map population density. Towards Agri-environmental indicators, Topic report, pp. 92-103.
- 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.

- Gardi, C., Bosco, C., Rusco, E., & L., M. (2010). An analysis of the land use sustainability index (lusi) at territorial scale based on corine land cover. *Management of Environmental Quality: An International Journal*, 680-694.
- 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.
- Guarino, R. (2006). On the origin and evolution of the Mediterranean dry grasslands. *Berichte der Reinhold Tüxen Gesellschaft*, 18, 195-206.
- 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., Huc, M., Villa Pascual, D., Dedieu, G., 2010. A multi-temporal method for cloud detection, applied to FORMOSAT-2, VENμS, LANDSAT and SENTINEL-2 images. *Remote Sens. Environ.* 114, 1747–1755.
- 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., 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. 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.
- Hervieu, B. (2006). Agriculture: a strategic sector in the Mediterranean area. In: CIHEAM Analytic note N°6-March 2006.
- 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, pp. 1417–1434.
- Hoover, A., Jean-Baptiste, G., Jiang, X., Flynn, P., Bunke, H., Goldgof, D., . . . Fisher, R. (1996). An experimental comparison of range image segmentation algorithms. IEEE transactions on pattern analysis and machine intelligence, 18(7), 673-689.
- Horning, Ned. "Random Forests: An algorithm for image classification and generation of continuous fields data sets." Proceedings of the International Conference on Geoinformatics for Spatial Infrastructure Development in Earth and Allied Sciences, Osaka, Japan. Vol. 911. 2010.
- Huang, X., Zhang, L., & Li, P. (2007, May). Classification and Extraction of Spatial Features in Urban Areas Using High-Resolution Multispectral Imagery. IEEE Geoscience and Remote Sensing Letters, 4(2), 260-264.
- Hüttich, C., Gessner, U., Herold, M., Strohbach, B., Schmidt, M., Keil, M., & Dech, S. (2009). On the suitability of MODIS time series metrics to map vegetation types in dry savanna ecosystems: A case study in the Kalahari of NE Namibia. Remote sensing, 1(4), 620-643
- 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.
- Inglada, J. (2019, 09 06). Retrieved from <http://osr-cesbio.ups-tlse.fr/~oso/posts/2018-06-06-carte-s2-2017-vecteur/>
- Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D., & I., R. (2017a, January). Operational High Resolution Land Cover Map Production at the Country Scale Using Satellite Image Time Series. Remote Sensing, 95(9), 1-35. Retrieved 02 22, 2018, from <http://www.mdpi.com/2072-4292/9/1/95>
- Jacques, D., Kergoat, L., Hiernaux, P., Mougin, E. and Defourny, P. (2014). Monitoring dry vegetation masses in semi-arid areas with MODIS SWIR bands. In: Remote Sensing of Environment 153 (2014) 40–49.
- 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.
- Jouven, M., Lapeyronie, P., Moulin, C. H., & Bocquier, F. (2010). Rangeland utilization in Mediterranean farming systems. Animal, 4(10), 1746-1757.
- 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.
- Kleeschulte, S., Banko, G., Smith, G., Arnold, S., Scholz, J., Kosztra, B. and Maucha, G. (2018). Technical specifications for implementation of a new land-monitoring concept based on EAGLE - D4: Draft design concept and CLC-Backbone and CLC-Core technical specifications, including requirements review.
- Kottek, M., Grieser, J., Beck, C., Rudolf, B. and Rubel, F. (2006). World Map of the Köppen-Geiger climate classification updated. In: Meteorologische Zeitschrift, Vol. 15, No. 3, 259-263.
- Kulkarni, A. D. and Lowe, B. (2016). Random Forest Algorithm for Land Cover Classification. Computer Science Faculty Publications and Presentations, 2016.

- Lambert, Marie-Julie ; Waldner, François ; Defourny, Pierre, 2016. Cropland Mapping over Sahelian and Sudanian Agrosystems: A Knowledge-Based Approach Using PROBA-V Time Series at 100-m. In: Remote Sensing, Vol. 8(3), no.232, p. 1-23.
- 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.
- Lewiński S., Malinowski R., Rybicki M., Gromny E., Nowakowski A., Jenerowicz M., Krupiński M., Krupiński M., Krätzschmar E., Günther S., 2019. Automatic Land Cover Classification of Europe with Sentinel-2 imagery, 2019 Living Planet Symposium, 13-17 May 2019, MiCo - Milano Congressi, Milan, Italy.
- Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology, 49(4), 764-766.
- 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.
- Li, T., Ni, B., Wu, X., Gao, Q., Li, Q., & Sun, D. (2016). On random hyper-class random forest for visual classification. Neurocomputing, 172, 281-289.
- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.
- 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, 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.
- Lobell, D. B., Hammer, G. L., McLean, G., Messina, C., Roberts, M. J., & Schlenker, W. (2013). The critical role of extreme heat for maize production in the United States. Nature Climate Change, 3(5), 497
- 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.
- 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.
- López-Sánchez, J. M., Ballester-Berman, J. D., Navarro-Sanchez, V. D., & Vicente-Guijalba, F. (2012, July). Experimental validation of the interferometric coherence formulation in single-transmit mode. In 2012 IEEE International Geoscience and Remote Sensing Symposium (3114-3117). IEEE.

- Louhaichi, M., Johnson, M. D., Clark, P. E., & Johnson, D. E. (2012). Developing a coherent monitoring system for Mediterranean grasslands. *New Approaches for Grassland Research in a Context of Climatic and Socio-Economic Changes. Options Méditerranéennes Série A*, 102, 47-51.
- 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.
- Marinho, E., Vancutsem, C., Fasbender, D., Kayitakire, F., Pini, G., & Pekel, J. F. (2014). From Remotely Sensed Vegetation Onset to Sowing Dates: Aggregating Pixel-Level Detections into Village-Level Sowing Probabilities. *Remote Sensing*, 6(11), 10947-10965.
- Maranon, T. (1988). Agro-sylvo-pastoral systems in the Iberian Peninsula: dehesas and montados. *Rangelands*, 10(6), 255-258.
- Martone, M., Rizzoli, R., Wecklich, C., González, C., Bueso-Bello, J.L., Valdo, P., Schulze, D., Zink, M., Krieger, G., Moreira, A., 2018. The Global Forest/Non-Forest Map from TanDEM-X Interferometric SAR Data. *Remote Sensing of Environment*, vol. 205, pp. 352-373.
- Mas, J., & Flores, J. (2008). The application of artificial neural networks to the analysis of remotely sensed data. *International Journal of Remote Sensing*, 617-663.
- Matton, Nicolas ; Sepulcre Canto, Guadalupe ; Waldner, François ; Valero, Silvia ; Morin, David ; Inglada, Jordi ; Arias, Marcela ; Bontemps, Sophie ; Koetz, Benjamin ; Defourny, Pierre, 2015. An Automated Method for Annual Cropland Mapping along the Season for Various Globally-Distributed Agrosystems Using High Spatial and Temporal Resolution Time Series. In: *Remote Sensing*, Vol. 7, no.10, p. 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.
- Metzger, M.J., Bunce, R.G.H., Jongman, R.G.H., Mùcher, C.A., Watkins, J.W., 2005. A climatic stratification of the environment of Europe. *Global Ecology and Biogeography*, (Global Ecol. Biogeogr.) (2005) vol. 14, pp. 549–563.
- 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.
- Montserrat, P. E. D. R. O., & Fillat, F. E. D. E. R. I. C. O. (1990). The systems of grassland management in Spain. *Managed grasslands*, 17, 37-70.
- Moulin, S., Kergoat, L., Viovy, N., & Dedieu, G. (1997). Global-scale assessment of vegetation phenology using NOAA/AVHRR satellite measurements. *Journal of Climate*, 10(6), 1154-1170.
- 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.

- Möckel, T., Dalmayne, J., Prentice, H. C., Eklundh, L., Purschke, O., Schmidlein, S., & Hall, K. (2014). Classification of grassland successional stages using airborne hyperspectral imagery. *Remote Sensing*, 6(8), 7732-7761.
- 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.
- Nguy-Robertson, A. L., Peng, Y., Gitelson, A. A., Arkebauer, T. J., Pimstein, A., Herrmann, I. & Bonfil, D. J. (2014). Estimating green LAI in four crops: Potential of determining optimal spectral bands for a universal algorithm. *Agricultural and forest meteorology*, 192, 140-148.
- 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.
- Palacios-Orueta, A., Huesca, M., Whiting, M. L., Litago, J., Khanna, S., Garcia, M., & Ustin, S. L. (2012). Derivation of phenological metrics by function fitting to time-series of Spectral Shape Indexes AS1 and AS2: Mapping cotton phenological stages using MODIS time series. *Remote sensing of environment*, 126, 148-159.
- Peel, M. C., Finlayson, B. L., & McMahon, T. A. (2007). Updated world map of the Köppen-Geiger climate classification. *Hydrology and earth system sciences discussions*, 4(2), 439-473.
- Pekel, J. F., Ceccato, P., Vancutsem, C., Cressman, K., Vanbogaert, E., & Defourny, P. (2011). Development and Application of Multi-Temporal Colorimetric Transformation to Monitor Vegetation in the Desert Locust Habitat. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2(4), 318-326.
- 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., Huadong, G., Blaës, X., Ferri, S., Gueguen, L., Halkia, M., . . . Zanchetta, L. (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(5), 1939-1404.
- 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., Dudik, 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.
- Plantureux, S., Bernués, A., Huguenin-Elie, O., Hovstad, K., Isselstein, J., McCracken, D., ... & Vackar, D. (2016). Ecosystem service indicators for grasslands in relation to ecoclimatic regions and landuse. The Multiple Roles of Grassland in the European Bioeconomy. European Grassland Federation (EGF), Trondheim, Norway, 524-547.
- Pontius, R.G. and Millones, M. (2011) Death to Kappa: Birth of Quantity Disagreement and Allocation Disagreement for Accuracy Assessment. *International Journal of Remote Sensing*, 32, 4407-4429. <http://dx.doi.org/10.1080/01431161.2011.552923>.
- Porqueddu, C., Ates, S., Louhaichi, M., Kyriazopoulos, A. P., Moreno, G., Pozo, A., ... & Nichols, P. G. H. (2016). Grasslands in 'Old World'and 'New World'Mediterranean-climate zones: past trends, current status and future research priorities. *Grass and Forage Science*, 71(1), 1-35.
- Porqueddu, C., Melis, R. A. M., Franca, A., Sanna, F., Hadjigeorgiou, I., & Casasús Pueyo, I. (2017). The role of grasslands in the less favoured areas of Mediterranean Europe
- 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.
- 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.
- 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. and Defourny, P. (2010). Automated image-to-map discrepancy detection using iterative trimming. *Photogramm. Eng. Remote Sens.*, 76, 173–181.
- 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., 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.
- Reed, B. C., Brown, J. F., VanderZee, D., Loveland, T. R., Merchant, J. W., & Ohlen, D. O. (1994). Measuring phenological variability from satellite imagery. *Journal of vegetation science*, 5(5), 703-714.
- White, M. A., Thornton, P. E., & Running, S. W. (1997). A continental phenology model for monitoring vegetation responses to interannual climatic variability. *Global biogeochemical cycles*, 11(2), 217-234.
- Ren, X., & Malik, J. (2003). Learning a classification model for segmentation. *Ninth IEEE International Conference on Computer Vision*, (p. 10).
- 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.
- Rivas-Martínez, S., Rivas Saenz, S. and Penas, A. (2011). Worldwide bioclimatic classification system. In: *Global Geobotany*, Vol n°1. December 2011. Pp.1-634+4maps.
- 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 Photogrammetry 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.
- Roggero, P., Bagella, S., Salis, L., Marrosu, G., Rossetti, I. Fanni, S. and Caria, M. (2013). Effects of long-term management practices on grassland plant assemblages in Mediterranean cork oak silvo-pastoral systems. In: *Plant Ecology, An International Journal*. ISSN 1385-0237, Volume 214, Number 4, *Plant Ecol* (2013) 214:621-631. DOI 10.1007/s11258-013-0194-x.
- Rouse Jr, J., Haas, R., and Schell, J. (1974). Monitoring vegetation systems in the great plains with erts. *NASA Special Publication*, 351, 309.
- 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.
- Sabo, F., Corbane, C., Politis, P., Pesaresi, M., & Kemper, T. (2019). Update and improvement of the European Settlement map. *Joint Urban Remote Sensing Event (JURSE)* (pp. 1-4). IEEE.

- Sakamoto, T., Wardlow, B. D., Gitelson, A. A., Verma, S. B., Suyker, A. E., & Arkebauer, T. J. (2010). A two-step filtering approach for detecting maize and soybean phenology with time-series MODIS data. *Remote Sensing of Environment*, 114 (10), 2146-2159.
- Salis, L., Sitzia, M., Fanni, S., Bagella, S., Zanzu, N., & Roggero, P. P. (2011). Relationships between grassland management, soil and pasture characteristics in piedmont Mediterranean grazing systems. In 16th Meeting of the FAO CIHEAM Mountain Pastures Network.
- 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., Hiltcher, 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.
- Siedentop, S., & Meinel, G. (n.d.). Corine land cover 2000 in nation-wide and regional monitoring of urban land use and land consumption.
- 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. *Remote 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. (2001): One-class classification. PhD thesis, Delft University of Technology.
- Tax, M.J.D. and P. W. Duin (2004). Support Vector Data Description, *Machine Learning*, 54, 45–66.
- Todorovic, M. (2016). Climate Change and the Mediterranean agriculture: expected impacts, possible solutions and the way forward. In: *CIHEAM Watch Letter n°37*, 2016.
- 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.
- 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.
- 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.
- Valero, S., Morin, D., Inglada, J., Sepulcre, G., Arias, M., Hagolle, O. & Koetz, B. (2016). Production of a dynamic cropland mask by processing remote sensing image series at high temporal and spatial resolutions. *Remote Sensing*, 8(1), 55.
- 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.
- Vanden Borre, J., Paelinckx, D., Mûcher, 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.
- Van der Walt, S., Schönberger, J., Nunez-Iglesias, J., Boulogne, F., Warner, J., Yager, N., . . . contributors, s.-i. (2014). *scikit-image: image processing in Python*. PeerJ.
- Van Tricht, K., Gobin, A., Gilliams, S., & I., P. (2018). Synergistic use of radar sentinel-1 and optical sentinel-2 imagery for crop mapping: A case study for belgium. *Remote Sensing*.

- Varmuza, K., & Filzmoser, P. (2016). Introduction to multivariate statistical analysis in chemometrics. CRC press
- Vedaldi, A., & Soatto, S. (2008). Quick shift and kernel methods for mode seeking. In European conference on computer vision (pp. 705-718). Springer.
- Verhegghen, A. (2013). Global land surface vegetation phenology using 13 years of SPOT VEGETATION daily observations (Doctoral dissertation, UCL-Université Catholique de Louvain).
- 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.
- VITO. (2019, 09 06). Retrieved from <https://lcviewer.vito.be/>.
- Vrahnakis, M. (2016): Mediterranean Type Ecosystems (MTEs): a brief introduction. Technological Institute of Thessaly, Greece.
- Waldner, François ; Sepulcre Canto, Guadalupe ; Defourny, Pierre, 2015. Automated annual cropland mapping using knowledge-based temporal features. In: *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 110, p. 1-13.
- Waldner, F., De Abelleira, 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. Heinzl, 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.
- White, J. D., Running, S. W., Nemani, R., Keane, R. E., & Ryan, K. C. (1997). Measurement and remote sensing of LAI in Rocky Mountain montane ecosystems. *Canadian Journal of Forest Research*, 27(11), 1714-1727.
- 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.
- Wulder, M., & Franklin, S. (2012). Remote sensing of forest environments: concepts and case studies. Springer Science & Business Media.
- 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., 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.
- 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.
- 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.
- Zeng, L., Wardlow, B. D., Wang, R., Shan, J., Tadesse, T., Hayes, M. J., & Li, D. (2016). A hybrid approach for detecting corn and soybean phenology with time-series MODIS data. *Remote sensing of environment*, 181, 237-250.
- 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.
- 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. (2017). Change detection using landsat time series: A review of frequencies, preprocessing, algorithms, and applications. *ISPRS Journal of Photogrammetry and Remote Sensing*, 130, 2017.
- 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.
- 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.
- Zolotokrylin, A.N.(2012). Droughts: causes, distribution and consequences. In: *Natural Disasters – Vol II*.