ASSESSING THE CONTRIBUTION OF SPECTRAL AND TEMPORAL FEATURES FOR ANNUAL LAND COVER AND CROP TYPE MAPPING

C. Karakizi *, I. A. Tsiotas, Z. Kandylakis, A. Vaiopoulos, K. Karantzalos

Remote Sensing Laboratory, National Technical University of Athens, Heroon Polytechniou 9, 15780 Zographos, Greece
chrkarakizi@central.ntua.gr

KEY WORDS: Classification, Mapping, Sentinel-2, Datacubes, Time-Series, Evaluation

ABSTRACT:

Freely available satellite image time-series are currently the most exploited data towards land cover mapping. In this work we assess the contribution of spectral and temporal features for the detailed, i.e., with more than thirty classes, land cover and crop type mapping based on annual Sentinel-2 data. As a baseline we employed a datacube consisting of spectral features, i.e., spectral bands and indices from one tile of Sentinel-2A data for the year 2016. Then we formed two different datacubes of reduced dimensions, containing either spectrotemporal or temporal features and performed the same experiments in order to assess their contribution. For the second dataset only spectral features that fulfilled certain temporal conditions were retained, reducing by 40% the initial datacube dimensionality. The third dataset was formed only of temporal features resulting to a reduction of 50%. A random forest classifier was employed for the classification procedure and standard accuracy metrics for the validation. All experiments resulted into high overall accuracy rates of over 90% while rates for average F-score metric exceeded 78% in all cases. The quantitative and qualitative validation indicated that the baseline dataset modestly outperformed the other two of spectrotemporal and temporal features. Insights regarding the influence of spectral differentiation among classes and the impact of their sample size, on the per-class performance are further discussed. The importance of spatial independency for training and testing sets was also demonstrated highlighting the need of following best practises during validation in order to deliver a realistic estimation of the produced map accuracy.

1. INTRODUCTION

Accurate and regularly-updated mapping along with change analysis arises as essential for several scientific communities, but also for public and regional authorities in terms of supporting decision making, planning, sustainable development and natural resources management. At the same time open data policies both in the USA and EU, are delivering an unprecedented volume of satellite data with increasing levels of resolution. Currently, the availability of Landsat-8 and Sentinel-2 data significantly expands the capabilities of timely, accurate and detailed land cover mapping from time series of cost-free satellite data. The use of multi-temporal data for land cover mapping tasks at annual basis has become currently the standard method, since in this way the valuable information of the phenological variations of different land-cover types can be exploited (Cihlar, 2000; Inglada et al., 2017; Jia et al., 2014; Karakizi et al., 2018b; Xie et al., 2019)

Mapping crop areas and classifying different crop types, arises as an even more complex problem requiring the use of data with higher spatial, spectral and temporal resolution. In recent studies the demanding task of crop-type mapping has been addressed with the use of satellite imagery combined with machine learning frameworks (Defourny et al., 2019; Inglada et al., 2015; Karakizi et al., 2018a; Lira Melo de Oliveira Santos et al., 2019). The use of machine learning techniques like Support Vector Machines (SVM) and Random Forests (RF) have gained rapid recognition for land cover and crop type classification studies (Defourny et al., 2019; Inglada et al., 2017, 2015; Karakizi et al., 2018b, 2018a; Xie et al., 2019; Zhai et al., 2018). A different approach, deep learning and Neural Networks (NN), is currently one of the fastest-growing trends in remote sensing. Compared to more shallow architectures (like SVM and RF) for the task of land cover mapping, the NN based analysis has given results of similar accuracy. However, it is accompanied by a significantly increased computational cost related with the complexity of the training procedure combined with the high dimensionality of image time series (Karakizi et al., 2018b; Khatami et al., 2016; Stoian et al., 2019).

Apart from the optimal selection of a classifier that balances highly accurate results and lower computational needs, the choice of classification features also plays an important role in the efficiency of the classification framework. Spectral features, namely multispectral satellite bands and derived spectral indices, have been established as the main set of input features for land cover and crop type classification in the recent literature (Defourny et al., 2019; Inglada et al., 2017, 2015; Karakizi et al., 2018b, 2018a). High dimensionality issues and redundancy problems of unchanged regions related with multispectral time series are often tackled by deriving temporal metrics from the multi-temporal data (Zhai et al., 2018). Multi-temporal metrics are statistical derivations of the time series per pixel, e.g., maximum, minimum, median and other percentile values of the spectral bands.

Temporal features are frequently used and recommended for crop-areas mapping (Defourny et al., 2019; Song et al., 2017; Valero et al., 2016; Waldner et al., 2017) since those type of metrics capture vegetation phenology but are generally insensitive to the timing of phenological differences across large areas (Egorov et al., 2018). Temporal metrics can also be employed to produce spectrotemporal features. This is achieved by defining key dates of remote sensing stages for the dynamic classes (vegetation, crops etc.) and then deriving the reflectance values of the spectral bands on those specific dates. As a result, the dataset to be classified is independent of the calendar sequence, i.e., the sequence of acquisition dates, for the intra-annual changes and thus grants better handling of the dynamic

cropland diversity and the agro-climatic gradient across the landscape (Matton et al., 2015).

Towards this direction, in this paper we assess the contribution of spectral and temporal features for the challenging task of detailed land cover and crop type mapping based on multi-temporal Sentinel-2 data from the year 2016. Aiming to reduce the computational costs of a hardware-hungry baseline dataset consisting of spectral features, we formed two additional datasets with reduced dimensions. The reduced datasets were formed employing spectrotemporal and temporal features derived from the initial baseline dataset. The same experimental set-up based on a RF classifier was applied to all three datasets, to ensure an objective comparison. A comprehensive quantitative and qualitative evaluation was performed for all experiments, while additional aspects like the spatial independency between training and testing sets were also discussed in order to deliver a reliable estimation of the produced map accuracy.

2. MATERIALS AND METHODS

2.1 Study Area

The study area is located in central continental Greece and corresponds to the Sentinel-2 tile 34SEJ (Figure 1). It covers an extent of about 10,000 km^2 and includes parts of four administrative regions of the country. We selected this region as the study area since it presents highly heterogeneous landscapes and varying terrain relief including plains but also several mountain masses. The Pindus mountain range crosses the study area from north-west to south-east, presenting covers of natural vegetation such as broadleaved and coniferous forests, natural grasslands, sclerophyllous vegetation and barelands. On the east of Pindus lies a big part of the Thessalian Plain, one of the most important agricultural regions of Greece. Agricultural land consists mainly of cotton, maize, cereals, clover and other grass fodders. Urban areas like the cities of Trikala and Karditsa and smaller towns and villages are scattered across the plain. The study area also includes several water bodies, e.g, man-made lakes Plastira, Pournariou, Kremaston, while on the south-west border of the study area, lies the salty-water lagoon of Messolonghi.

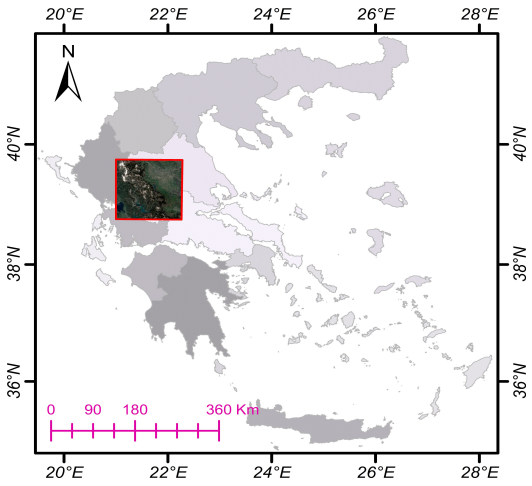

Figure 1. The study area (in red box) corresponding to the Sentinel-2 tile 34SEJ (in a natural color composite of March 2016) and its location on the Greek territory

2.2 Sentinel-2 Data Selection and Preprocessing

Sentinel-2A L1C data of less than 10% cloud coverage (10 dates) for year 2016 was downloaded from the ESA Sci-Hub for the tile 34SEJ. An atmospheric correction procedure followed using the Sen2Cor processor, in order to produce surface reflectance products of level 2C. Afterwards, BRDF correction was carried out, where pixel reflectance values were adjusted as if the satellite's location was at the nadir concurrently with image acquisition. Then, cloud and shadow screening was conducted with the F-mask algorithm and interpolated values using the previous and following cloud/shadow-free dates, were produced for cloudy and invalid pixels. Furthermore, geolocation shift errors (1-2 pixels) that may occur between different dates/images of the same tile were also resolved. As a final step, the medium (20m) and low (60m) resolution bands of Sentinel-2 were sharpened from the nearest high (10m) spatial resolution band.

2.3 Class Nomenclature and Reference Data

Concerning the class nomenclature, the aim was to highly analyse land cover of the study area including detailed information for crop types. To this end, we used the nomenclature of CORINE Land Cover (CLC) third level with several modifications for general land cover classes, while geospatial data from Rapid Field Visits (RFVs) of the Greek Paying Agency (OPEKEPE) defined the thematic analysis of arable land classes. An intensive annotation procedure was carried out to produce reference data. Two image interpretation experts manually annotated polygons for 31 different land cover and crop type classes using a variety of datasets including Sentinel-2 images for the year 2016, Google Earth and Bing Satellite very high resolution imaging data, CLC2012 product and crop's geospatial data from Greek Paying Agency. Since representative training samples are one of the most critical components in supervised classification, the experts studied the area thoroughly and noted as many variations of each class as possible (Karakizi et al., 2018b). At the same time, the good practice of keeping sample size per class relative to each class's respective occurrence in the study area was also considered when creating the reference data. However, this practice was impossible to be achieved for arable land classes, since reference data for those -hard to photo interpret- crop classes were limited by the availability of RFV's geospatial data.

2.4 Classification Algorithm and Features

A Random Forest (RF) classifier was used for the implementation of the proposed methodology. RF classifier as further analysed in the Introduction section, is considered a robust machine learning classification method that has been successfully applied in recent similar studies (Defourny et al., 2019; Inglada et al., 2017, 2015; Xie et al., 2019). The choice of parameters for the RF classifier has been proven not very sensitive for land cover tasks (Pelletier et al., 2016). For this study we implemented an RF classifier with Python 3.7.0 using scikit-learn and 100 trees. All experiments were executed on a server running Ubuntu 18.04, with an Intel(R) Core(TM) i7-5820K CPU at 3.30GHz and 48 GBs of RAM.

Based on our previous research efforts (Karakizi et al., 2018b, 2018a) and on the related bibliography (Defourny et al., 2019; Inglada et al., 2017, 2015) we chose six Sentinel-2 spectral bands, i.e., Blue (Band 2), Green (Band 3), Red (Band 4), Red-

Edge (Band 6), NIR (Band 8), SWIR (Band 11) along with three spectral indices, namely NDVI, NDWI and NDBI to serve as the classification features. The baseline dataset, **Dataset A**, was formed by stacking the 10-dates' spectral features into a single cube of 90 layers in total.

Dataset B was formed by deriving the spectral features from dates/images of Dataset A that fulfilled specific temporal metric conditions. Distinct remote sensing stages, especially for crop-areas, have been reported to be defined by key dates linked with spectral features' statistical derivations, like maximum (max) and minimum (min) value of NDVI and Red band (Lambert et al., 2016; Matton et al., 2015). We experimented with multiple combinations of temporal metrics as conditions, like min, max and median of the available spectral bands and the final, most efficient set, comprised of: maxRed, minRed, medianRed, maxNDVI, minNDVI, medianNDVI. Dataset B was formed by extracting the values of the nine spectral features from Dataset A, from the respective date that fulfilled the condition posed by each temporal metric, per pixel. As a result, Dataset B consisted of 54 layers in total.

Dataset C was formed by calculating temporal metrics on the Dataset A time-series and using directly those metrics as input features for the classification. We experimented with multiple combinations of temporal metrics as features and the final, most efficient set, comprised of min, max, median, mean, standard deviation of all nine spectral bands and indices. By this way Dataset C consisted of 45 layers in total.

3. EXPERIMENTAL RESULTS AND DISCUSSION

In this section a thorough quantitative and qualitative analysis is presented based on the results obtained after applying the same classification framework on Datasets A, B and C.

3.1 Accuracy Metrics

All models were trained and validated using exactly the same training and testing areas, in order to allow valid comparison between datasets. The independency of the validation procedure was secured by splitting the reference data at the polygon level, using 65% for training and 35% for testing. An additional validation procedure for Dataset A, was conducted by splitting the reference data randomly at pixel level, holding the same ratio for training and testing, towards quantitatively assessing the impact of the spatial correlation on accuracy metrics' overestimation.

The validation of the land cover classification experiments was quantitatively implemented forming confusion matrices at the pixel level. The standard accuracy metrics of Overall Accuracy (OA), User's and Producer's Accuracy (UA & PA) were calculated. Per class F-measure (F1) scores were also calculated as the combined metric of the harmonic mean between UA and PA. Qualitative evaluation was also performed by throughout intensive observation of the produced land-cover and crop-type maps.

3.2 Quantitative Comparative Analysis

In Table 1, the resulting F1 rates are presented per class and dataset, while their average rates and the OA of each experiment are presented in the last two rows.

In particular, lower rates (less than 40%) are marked with double underline and highest rate per class (per row) between the different datasets with bold. Scores on the last two rows of Table 1 indicate that the proposed approach achieved results of very high OA (<90%) for all datasets. The average F1 rates from all classes ranged between 78%-83% for the different experiments.

Code	Class	F1 SCORE (%)		
		Dataset A	Dataset B	Dataset C
DUF	Dense urban fabric	**71,05**	68,42	68,89
SUF	Sparse urban fabric	**70,07**	67,68	68,31
ICU	Industrial commercial units	**66,54**	60,01	65,79
RAN	Road/Asphalt networks	64,44	59,61	**65,62**
MES	Mineral extraction sites	**67,39**	63,95	64,80
PHV	Photovoltaic units	**86,62**	83,32	84,69
GRH	Greenhouses	**97,85**	96,19	97,70
BLF	Broad-leaved forest	**97,48**	96,38	97,38
CNF	Coniferous forest	**100,00**	**100,00**	99,50
NGR	Natural grasslands	**91,90**	87,50	88,24
DSV	Dense scleroph. vegetation	**96,00**	95,30	95,50
SSV	Sparse scleroph. vegetation	**87,64**	86,47	86,43
VNY	Vineyards	83,86	**85,29**	79,98
OLG	Olive groves	**78,58**	70,78	77,34
FRT	Fruit trees	**80,54**	61,62	72,18
WHT	Wheat	**92,74**	90,45	92,23
BRL	Barley	**55,50**	32,42	39,02
OAT	Oat	**31,28**	24,08	18,49
MAI	Maize	**94,53**	89,67	91,83
CTN	Cotton	**94,50**	89,47	94,05
TBC	Tobacco	96,00	91,78	**96,43**
CLV	Clover	**91,06**	83,49	87,72
GRF	Other grass fodders	62,86	67,02	**68,43**
FLW	Fallow	**62,66**	52,23	54,84
SVA	Sparsely vegetated areas	**88,27**	86,60	85,37
BDS	Beaches, dunes, sands	**78,56**	77,28	77,08
RCK	Bare rocks	**91,37**	89,24	89,90
MRS	Marshes	**94,21**	90,78	92,12
WCR	Water courses	**90,46**	90,29	90,42
WBD	Water bodies	**100,00**	97,44	**100,00**
CWT	Coastal water	**100,00**	95,94	98,89
	Average F1 (%)	**82,71**	78,41	80,30
	OA (%)	**93,55**	90,70	91,54

Table 1. OA and per class F1 rates for the RF experiments on Datasets A, B and C. Highest rate per row is marked with bold and low rates (<40%) with double underline

Classification on Dataset A, of 90 spectral features, holds the highest rates (with bold) for average F1 and OA, and for most individual classes' F1 scores, compared to the performances of the other two experiments. Dataset C of 45 temporal features performed, in almost every case, better that Dataset B, of 54 spectrotemporal features. Comparing Dataset A and C, the latter comes with half the dimensionality of the first. However, accuracy rates did not differ much, i.e., less than 3% for average F1 and OA. For the experiments in all three datasets, F1 score for the majority of classes exceeded 70%. For all three datasets, very high F1 rates of over 90% were achieved for classes *Greenhouses*, *Broad-leaved forest*, *Coniferous forest*, *Dense sclerophyllous vegetation*, *Wheat* and wetland (MRS) and water classes (WCR, WBD, CWT). Sub-classes of the cereal crop family, name *Barley* and *Oat* were the only cases of very low rates of under 40%.

In general, individual classes' performance followed the same pattern, concerning accuracy rates derived from the three different datasets. To better illustrate the latter and at the same time assess the classes' sample size contribution, on the per class accuracy produced by the classification model, Figure 2 is provided and analysed. On the three coloured bars, the PA rates are presented per dataset for every studied class. Producer's Accuracy (PA) is an indicator of the map accuracy from the point of view of the map/model maker (the producer). It expresses the proportion of ground truth pixels that were correctly classified by the classification model. The black line with markers in Figure 2 indicates the proportion of sampling size per class, derived as the ratio between sample size of each class and total sample size for all classes. Sample size is examined since RF, like other supervised machine learning algorithms, tend to promote prediction accuracy of prevailing classes in terms of representativeness which in most cases leads to lower accuracies in less represented classes.

Upon more careful examination of Figure 2 it becomes apparent that the majority of classes, at least 20 out of 31, reached PA scores of over 70% for all 3 datasets. Concerning sample size, the top-five better represented classes, i.e., *Coniferous forest*, *Broad-leaved forest*, *Cotton*, *Clover* and *Maize*, achieved in most cases very high PA rates of more than 85%. However PA rates of over 85% were also achieved for very low sample size classes (<1%) like *Greenhouses*, *Marshes* and *Water courses*, suggesting that distinct spectral characteristics of specific classes eliminate the sample size bias. This is indeed highlighted by very high performance (>90%) in both F1 (Table 1) and PA rates for all cover classes related with the presence of water (WCR, WBD, CWT). On the other hand for classes with less distinctive spectral behavior and thus having more competitive classes of similar behavior, like *Industrial commercial units* and *Road and asphalt networks*, low sampling size seems to have affected performance against the other man-made classes.

Another important remark from Figure 2 is that the greatest differences in PA scores between datasets A, B and C occurred for crop classes (VNY-FLW). This is also the case for the greatest differences in F1 scores (Table 1). For crop classes other than *Vineyards*, *Tobacco* and *Other grass fodders*, classification on Dataset A ensured highest per class accuracy rates. Sample size seems to have played an important role for those classes and especially the arable ones, that their reference data creation, as explained in section 2.3., was limited by the availability of geospatial data provided by the Greek Paying Agency. In this case, classes *Fruit trees*, *Barley*, *Oat* and *Fallow* suffered both from low representativeness (<1%) and lower per class accuracy rates. Between those, *Barley* and *Oat* presented the lowest per class accuracy rates (in most cases <40%) and this fact can be also attributed to the presence of another crop-class of the same family, namely *Wheat* that had a significantly bigger sample size than the other two cereal classes. Apart from low sample size, for classes *Fruit trees* and *Fallow* lower accuracy rates could be also attributed to the fact that they include different covers or crop sub-types within the same class/category.

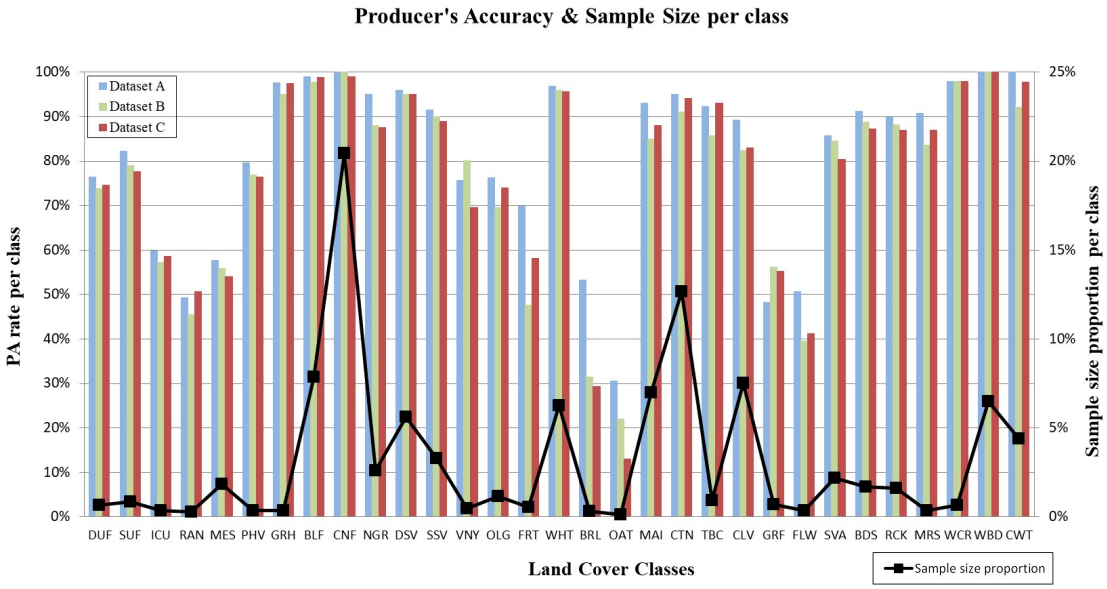

Figure 2. The resulting per-class PA rates (coloured vertical bars) for the RF experiments on Datasets A, B and C. The proportion of each class in the reference dataset, i.e., each class's sample size, is also presented with the black line with markers

3.3 Omission Errors per Class

In order to have a view of classification mixings between classes, we hereby present and discuss Table 2. Table 2 was calculated by dividing the produced confusion matrix for the most successful experiment, that of Dataset A, with the sum of reference testing pixels (ground truth) for each row. As a result, the table should be read per row (reference labels), so as the diagonal of the matrix expresses the PA percentage of each row/class, and all other cells the omission percentage for that class from classification errors towards all other column-classes (produced labels). Black line borders have been designed in order to visually group classes of the same general land cover family.

Artificial classes that consist of man-made surfaces achieved, in overall, medium to very high PA rates (49–98%) and in most cases they were confused between each other. *Photovoltaic units* and *Greenhouses*, characterised by materials of distinct spectral behaviour, held high rates for PA (>80%) and very low omission errors accordingly. The two classes of urban fabric (DUF, SUF) presented mixings of about 15% between each other, as expected. Omission errors towards those two classes, adding up to approx. 30%, presented the *class Road /Asphalt networks* with a PA of 49%. The artificial terrain class *Mineral extraction sites*, which is the only non-sealed man-made cover category in our nomenclature, achieved a PA rate of 58%, while presenting a high omission error rate (34%) to class *Beaches, dunes, sands*, of similar cover materials. *Industrial commercial units* class had omission errors towards both artificial (DUF, SUF) and bareland (BDS, RCK) classes of high brightness, presenting a PA rate of 60%.

Concerning areas of natural vegetation, i.e., forests (BLF, CNF), natural grasslands (NGR) and shrubland (DSV, SSV), PA reached rates over 92%. Thus, those classes presented very few omission errors to other classes.

For the 12 classes of the agriculture family PA rates presented a large range (31%-97%). Permanent crops (VNY, OLG, FRT) were classified with PA rates of about 70%-75% presenting omission errors mainly to other crop classes but also natural vegetation classes and especially towards *Sparse sclerophyllous vegetation*. Arable land classes *Wheat*, *Maize*, *Cotton*, *Tobacco* and *Clover* all held very high PA rates (>89%) and presented insignificant omission errors mainly to other crop classes. *Barley* and *Oat* classes of the cereal family, presented lower PA rates, since 36% and 25% respectively, of their reference data pixels used for testing, were classified as *Wheat*. As also discussed in the previous section, cereal class *Wheat* had a significantly bigger sample size than the other two cereals. Additionally *Other grass fodders* class also lost 49% of its ground truth pixels to *Wheat*. *Fallow* class, not corresponding to a single crop category, presented lower PA rates and omission errors mainly to *Wheat* (14%), *Cotton* (6%), *Natural grasslands* (9%) and *Sparse urban fabric* (11%). The latter misclassification case can be attributed to similar spectral behaviour at this scale (10m) since *Fallow* and *Sparse urban fabric* classes include both vegetated and non-vegetated covers in an irregular way.

Regarding bareland classes (SVA, RCK, BDS) very few omission cases were recorded and so PA rates exceeded 85%. Unsurprisingly *Beaches, dunes, sands* class presented omission errors of 4% to *Water courses* class, since its reference data also included sand along river banks. *Bare rocks* class had a 7% of its reference testing pixels classified as *Mineral extraction sites* class, strengthening the spectral relationship, mentioned in previous paragraphs, between those different-family but similar-material classes.

Wetland (MRS) and water classes' (WCR, WBD, CWT) reference testing pixels were classified with very high accuracy rates of over 91%, highlighting the spectral distinctness of water-related covers especially when exploiting the infrared spectrum.

Confusion Matrix: Dataset A - PA and Omission Errors (%)

	DUF	SUF	ICU	RAN	MES	PHV	GRH	BLF	CNF	NGR	DSV	SSV	VNY	OLG	FRT	WHT	BRL	OAT	MAI	CTN	TBC	CLV	GRF	FLW	SVA	BDS	RCK	MRS	WCR	WBD	CWT
DUF	**76**	16	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0
SUF	12	**82**	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0
ICU	5	4	**60**	0	17	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	3	0	0	0	0
RAN	16	15	0	**49**	0	6	0	1	0	0	1	3	0	1	0	0	0	0	2	1	0	1	0	0	2	0	1	1	0	0	0
MES	1	0	3	0	**58**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	34	3	0	0	0	0
PHV	1	2	0	0	0	**80**	0	0	0	3	0	5	0	2	0	3	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0
GRH	0	1	0	0	0	0	**98**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
BLF	0	0	0	0	0	0	0	**99**	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CNF	0	0	0	0	0	0	0	0	**100**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NGR	0	0	0	0	0	0	0	2	0	**95**	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
DSV	0	0	0	0	0	0	0	3	0	0	**96**	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SSV	0	2	0	0	0	0	0	0	0	0	3	**92**	0	1	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0
VNY	0	5	0	0	0	0	0	0	0	0	0	6	**76**	2	0	0	0	0	1	9	0	0	0	0	0	0	0	0	0	0	0
OLG	0	0	0	0	0	0	0	0	0	5	0	15	0	**76**	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0
FRT	0	2	0	0	0	0	0	1	0	0	3	5	1	13	**70**	0	0	0	1	1	0	3	0	0	0	0	0	0	0	0	0
WHT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**97**	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0
BRL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36	**53**	9	0	2	0	0	0	0	0	0	0	0	0	0	0
OAT	0	4	0	0	0	0	0	0	0	14	0	8	0	0	0	25	1	**31**	11	0	0	2	0	2	1	0	0	0	0	0	0
MAI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**93**	2	0	4	0	0	0	0	0	0	0	0	0
CTN	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	**95**	0	1	0	0	0	0	0	0	0	0	0
TBC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	2	**92**	1	0	0	0	0	0	0	0	0	0
CLV	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	0	6	0	**89**	0	0	0	0	0	0	0	0	0
GRF	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	49	0	0	0	0	0	0	**48**	2	0	0	0	0	0	0	0
FLW	0	11	0	0	0	0	0	0	0	9	0	5	0	0	0	14	0	0	0	6	0	0	0	**51**	2	0	0	0	0	0	0
SVA	0	2	0	0	0	0	0	0	0	5	0	5	0	1	0	0	0	0	0	1	0	0	0	0	**86**	0	1	0	0	0	0
BDS	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**91**	2	0	4	0	0
RCK	1	0	1	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	**90**	0	0	0	0
MRS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**91**	9	0	0
WCR	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**98**	0	0
WBD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**100**	0
CWT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**100**

Table 2. Confusion matrix for the RF experiment on Dataset A. Diagonal values (bold and boxed) represent the PA (recall) values for each class, while off-diagonal ones the omission errors. Omission errors greater than 3% are marked with red text and red filling

3.4 Overestimating Accuracy with Spatial Dependency on Training and Testing Sets

Machine learning and remote sensing best practices for validating classification outputs suggest splitting testing and training datasets in a manner that ensures the spatial independency of the two sets. This is also the case for land cover and crop type mapping applications (Defourny et al., 2019; Inglada et al., 2017; Stoian et al., 2019). This splitting strategy was followed also in this paper for the core experiments presented and analysed in the previous sub-sections.

However, in order to quantitatively assess the over-estimation effect on accuracy metrics, when training and testing data are not spatially independent, an additional experiment was carried out on Dataset A by splitting the reference data at pixel level instead of polygon level. Table 3 presents OA, average PA, UA and F1 for the auxiliary pixel split experiment on Dataset A in comparison with the original experiment on Dataset A, using polygon splitting for training and testing.

Accuracy metric	**Dataset A** (*split at polygon level*)	**Dataset A** (*split at pixel level*)	**Impact** *of spatial dependency*
OA	93,55%	98,70%	↑ 5,15%
PA	81,00%	95,29%	↑ 14,29%
UA	86,24%	97,48%	↑ 11,24%
F1	82,71%	96,37%	↑ 13,66%

Table 3. Accuracy metrics derived with different splitting strategies and their impact in estimated map accuracy

Table 3 indicates that all metrics exceeded the remarkably high rate of 95% when splitting at pixel level presenting significant raises compared to the main polygon-split experiment. In particular, OA for the complementary experiment exceeded 98% presenting an increase of more than 5%, while average per class metrics of PA, UA and F1 presented rises of over 10%. Rises were quite expected since splitting at pixel level involves considering pixels of the same polygon for both training and testing, which would probably be very similar spectrally.

After all, producing land cover maps covering large extents, such as those of one satellite scene, with not-automatically produced reference data from other LC products, is usually characterized by a small proportion of reference data compared to the whole area to be classified. In such a way the impact of spatial correlation when validating, is not to be ignored in order to provide the users an accuracy estimation that is closer to reality for the whole map.

3.5 Map Validation and Discussing Challenging-to-map Areas

The final step of the proposed methodology is map production. Three maps were produced, one for each dataset, applying the same RF framework. For predicting the map products the whole reference dataset was utilized when training the RF classifier, aiming at maximum exploitation of all available information. The general visual inspection on the produced maps did not reveal coarse differences between the different products. The most successful experiment i.e., the one based on Dataset A, is presented in Figure 3 (left hand side). Additionally, zoomed-in areas from all three maps are presented for visual comparison in Figure 3 (right hand side). This allows the direct comparison of the performance especially in challenging-to-map areas.

On the eastern part of the map (Figure 3, left) one can observe that the urban and suburban areas, i.e., cities of Trikala and Karditsa as well as many villages scattered across the Thessalian plain have been correctly detected and mapped (indicative classes: DUF, SUF, ICU) by the proposed classification framework. Forest classes (BLF, CNF) were correctly reported on upland and mountainous areas, while Mediterranean-type evergreen sclerophyllous vegetation classes (DSV, SSV) were mapped on lower terrains.

Regarding the qualitative evaluation over agricultural areas, the Thessalian plain was expectedly classified as various crop classes. A vast dominance of classes *Cotton*, *Wheat* and *Maize* cultivations is evident on the map which agrees with the agricultural information provided in subsection 2.1. Additionally, large inland water areas such as the man-made lakes of the study area, the salty-water lagoon of Messolonghi and its surrounding wetlands have been successfully detected as the respective classes (WBD, CWT, MRS).

Higher altitude regions of mountain massifs, mainly represented by the spur of Pindus, seem in most cases correctly illustrated with the presence of *Natural grasslands* and bareland classes (SVA, RCK). However a closer look on bareland areas (whitish colors) even at this scale, reveals certain omission errors towards man-made classes (redish/pink colors) that are further illustrated on the zoomed-in parts in the right side of Figure 3. In particular the example of mount Lakmos (alt. 2100m) of the Pindus range is presented here. This area consists of bareland covers and usually snow cover across the year. Therefore it is depicted in the imagery data with high brightness values across the spectrum, which is also the case for many other alpine regions.

By examining the zoomed-in parts from the maps of Dataset A, B and C, classification errors are apparent for all. The zoomed-in map of Dataset A present the fewest commission errors to -other than bareland classes- among the three maps. However, there are still considerable regions covered with urban and suburban classes (DUF, SUF) but also class *Cotton* of the crop family. The artificial classes' mixings with bareland can be attributed to high brightness values for both types. As far as class *Cotton* is concerned, its cultivation practices require high humidity levels on the soil, so it is possible that alpine vegetated regions covered by moss and lichen were confused with this class. More classification errors can be observed for Dataset B and C. Larger bareland areas are erroneously classified to even more in number artificial classes, namely *Dense urban fabric*, *Sparse urban fabric*, *Mineral extraction sites* and *Greenhouses*. These remarks are in accordance with per class accuracies, since classification in Dataset A resulted in *Sparsely vegetated areas* PA rates of 86%, 85%, 80% and that of *Bare rocks* of 90%, 88%, 87% for datasets A, B and C respectively.

In overall, visual inspection on the produced three maps agrees with the high classification OA rates (>90%) achieved by all datasets. Additionally, analysis on the challenging-to-map example, indicated also qualitatively Dataset A as the most successfully classified dataset.

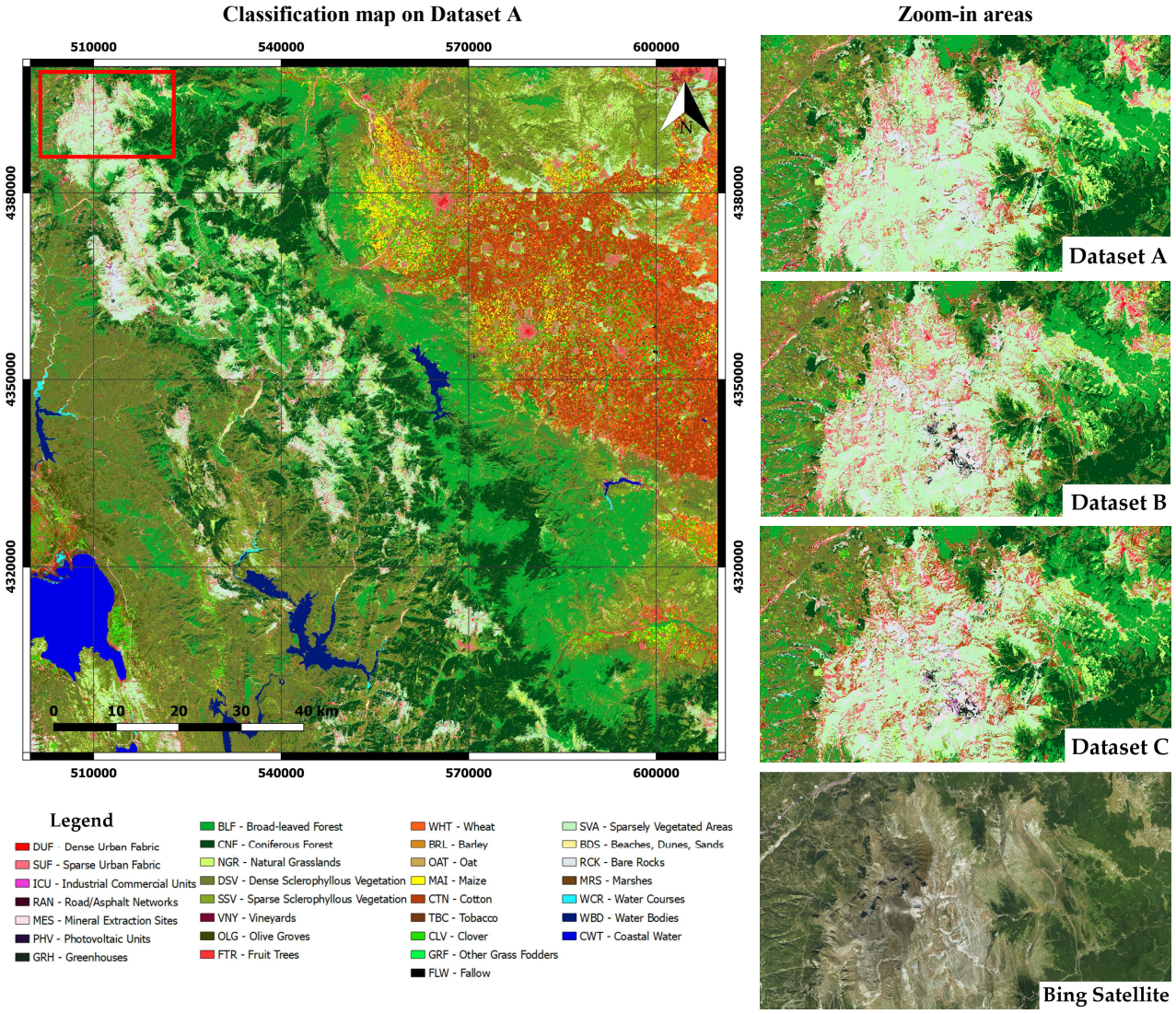

Figure 3. Classification map of the study area from experiment on Dataset A (left) and zoom-in region of mount Lakmos from the maps of all three datasets along with Bing Satellite imagery of the same region (right)

4. CONCLUSIONS

In this paper we assessed the contribution of spectral and temporal features for detailed land cover and crop type mapping using annual Sentinel-2 data for the year 2016. Three datasets were created and benchmarked using the same RF classification framework. The first baseline dataset (Dataset A) consisted of 90 spectral features, while the other two were formed with spectrotemporal (Dataset B) and temporal features (Dataset C), reducing the initial datacube dimensionality by 40% and 50% respectively. Quantitative evaluation demonstrated the efficiency of the proposed methodology by achieving results of high overall accuracy of over 90% in all experiments. Comparative quantitative and qualitative validation highlighted that in most cases Dataset A yielded modestly better results than the other two datasets that employed temporal metrics.

The choice of using temporal metrics for this work has been encouraged by recent similar studies and especially crop mapping ones, with the aim of allowing independency from the crop calendar and dealing with the cropland diversity across landscapes (Defourny et al., 2019; Egorov et al., 2018; Waldner et al.,2017). In this work, the study area consisted of one Sentinel-2 tile and thus cropland diversity and agro-climatic gradient did not ranged significantly. To this extent, the classification experiment based only on temporal metrics (Dataset C) presented lower rates by a 2-2,5% for OA and average F1, compared to the initial experiment on full-set spectral features of Dataset A. Nevertheless relative low differences in accuracy for experiments performed, encourage the use of temporal metrics for similar mapping tasks and especially for large geographical areas, contributing to reduced data dimensionality and thus less demands on hardware resources.

Further analysis on per class classification errors indicated that sample size along with the distinct spectral characteristics of each class had a significant impact in accuracy rates. Especially for arable crop classes, that present high spectral similarities among each other, reference data availability emerges as a crucial prerequisite for accurate classification. Towards this end, the free/open access to geospatial reference datasets like the annual declarations on the Land Parcel Identification System (LPIS, for EU) should significantly contribute to accurate and detailed land cover mapping at an annual basis.

REFERENCES

Cihlar, J., 2000. Land cover mapping of large areas from satellites: status and research priorities. *International Journal of Remote Sensing* 21, 1093–1114. https://doi.org/10.1080/014311600210092

Defourny, P., Bontemps, S., Bellemans, N., Cara, C., Dedieu, G., Guzzonato, E., Hagolle, O., Inglada, J., Nicola, L., Rabaute, T., Savinaud, M., Udroiu, C., Valero, S., Bégué, A., Dejoux, J.-F., El Harti, A., Ezzahar, J., Kussul, N., Labbassi, K., Lebourgeois, V., Miao, Z., Newby, T., Nyamugama, A., Salh, N., Shelestov, A., Simonneaux, V., Traore, P.S., Traore, S.S., Koetz, B., 2019. Near real-time agriculture monitoring at national scale at parcel resolution: Performance assessment of the Sen2-Agri automated system in various cropping systems around the world. *Remote Sensing of Environment* 221, 551–568. https://doi.org/10.1016/j.rse.2018.11.007

Egorov, A.V., Roy, D.P., Zhang, H.K., Hansen, M.C., Kommareddy, A., 2018. Demonstration of percent tree cover mapping using Landsat Analysis Ready Data (ARD) and sensitivity with respect to Landsat ARD processing level. *Remote Sensing* 10, 209.

Inglada, J., Arias, M., Tardy, B., Hagolle, O., Valero, S., Morin, D., Dedieu, G., Sepulcre, G., Bontemps, S., Defourny, P., Koetz, B., 2015. Assessment of an Operational System for Crop Type Map Production Using High Temporal and Spatial Resolution Satellite Optical Imagery. *Remote Sensing* 7, 12356–12379. https://doi.org/10.3390/rs70912356

Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D., Rodes, I., 2017. Operational high resolution land cover map production at the country scale using satellite image time series. *Remote Sensing* 9, 95.

Jia, K., Liang, S., Zhang, N., Wei, X., Gu, X., Zhao, X., Yao, Y., Xie, X., 2014. Land cover classification of finer resolution remote sensing data integrating temporal features from time series coarser resolution data. *ISPRS Journal of Photogrammetry and Remote Sensing* 93, 49–55. https://doi.org/10.1016/j.isprsjprs.2014.04.004

Karakizi, C., Antoniou, G., Karantzalos, K., 2018a. Towards Joint Land Cover and Crop Type Mapping with Numerous Classes, in: IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium. Presented at the *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2980–2983. https://doi.org/10.1109/IGARSS.2018.8517473

Karakizi, C., Karantzalos, K., Vakalopoulou, M., Antoniou, G., 2018b. Detailed land cover mapping from multitemporal landsat-8 data of different cloud cover. *Remote Sensing* 10, 1214.

Khatami, R., Mountrakis, G., Stehman, S.V., 2016. A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research. *Remote Sensing of Environment* 177, 89–100. https://doi.org/10.1016/j.rse.2016.02.028

Lambert, M.-J., Waldner, F., Defourny, P., 2016. Cropland Mapping over Sahelian and Sudanian Agrosystems: A Knowledge-Based Approach Using PROBA-V Time Series at 100-m. *Remote Sensing* 8, 232. https://doi.org/10.3390/rs8030232

Lira Melo de Oliveira Santos, C., Augusto Camargo Lamparelli, R., Kelly Dantas Araújo Figueiredo, G., Dupuy, S., Boury, J., Luciano, A.C. dos S., Torres, R. da S., le Maire, G., 2019. Classification of Crops, Pastures, and Tree Plantations along the Season with Multi-Sensor Image Time Series in a Subtropical Agricultural Region. *Remote Sensing* 11, 334. https://doi.org/10.3390/rs11030334

Matton, N., Canto, G., Waldner, F., Valero, S., Morin, D., Inglada, J., Arias, M., Bontemps, S., Koetz, B., Defourny, P., 2015. An Automated Method for Annual Cropland Mapping along the Season for Various Globally-Distributed Agrosystems Using High Spatial and Temporal Resolution Time Series. *Remote Sensing* 7, 13208–13232. https://doi.org/10.3390/rs71013208

Pelletier, C., Valero, S., Inglada, J., Champion, N., Dedieu, G., 2016. Assessing the robustness of Random Forests to map land cover with high resolution satellite image time series over large areas. *Remote Sensing of Environment* 187, 156–168. https://doi.org/10.1016/j.rse.2016.10.010

Song, X.-P., Potapov, P.V., Krylov, A., King, L., Di Bella, C.M., Hudson, A., Khan, A., Adusei, B., Stehman, S.V., Hansen, M.C., 2017. National-scale soybean mapping and area estimation in the United States using medium resolution satellite imagery and field survey. *Remote Sensing of Environment* 190, 383–395. https://doi.org/10.1016/j.rse.2017.01.008

Stoian, A., Poulain, V., Inglada, J., Poughon, V., Derksen, D., 2019. Land cover maps production with high resolution satellite image time series and convolutional neural networks: Adaptations and limits for operational systems. *Remote Sensing* 11, 1986.

Valero, S., Morin, D., Inglada, J., Sepulcre, G., Arias, M., Hagolle, O., Dedieu, G., Bontemps, S., Defourny, P., Koetz, B., 2016. Production of a Dynamic Cropland Mask by Processing Remote Sensing Image Series at High Temporal and Spatial Resolutions. *Remote Sensing* 8, 55. https://doi.org/10.3390/rs8010055

Waldner, F., Hansen, M.C., Potapov, P.V., Löw, F., Newby, T., Ferreira, S., Defourny, P., 2017. National-scale cropland mapping based on spectral-temporal features and outdated land cover information. *PLoS ONE* 12, e0181911. https://doi.org/10.1371/journal.pone.0181911

Xie, S., Liu, L., Zhang, X., Yang, J., Chen, X., Gao, Y., 2019. Automatic Land-Cover Mapping using Landsat Time-Series Data based on Google Earth Engine. *Remote Sensing* 11, 3023. https://doi.org/10.3390/rs11243023

Zhai, Y., Qu, Z., Hao, L., 2018. Land Cover Classification Using Integrated Spectral, Temporal, and Spatial Features Derived from Remotely Sensed Images. *Remote Sensing* 10, 383. https://doi.org/10.3390/rs10030383