JOINT LAND COVER AND CROP TYPE MAPPING USING MULTI-TEMPORAL SENTINEL-2 DATA FROM VARIOUS ENVIRONMENTAL ZONES IN GREECE

C. Karakizi *, Z. Kandylakis, A. D. Vaiopoulos, K. Karantzalos

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

KEY WORDS: Classification, Time Series, Temporal Features, Datacubes, Stratification

ABSTRACT:

In this work, we elaborate on the gained insights from various classification experiments towards detailed land cover mapping over four representative regions of different environmental characteristics in Greece. In particular, the proposed methodology exploits Sentinel-2 data at an annual basis, for the joint classification of 35 land cover and crop type classes. A number of pre-processing steps were employed on the satellite data, in order to address atmospheric and geometric effects, as well as clouds and pertinent shadows. Several classification set-ups were designed and performed using either time series of spectral features or temporal features. The latter consisted of statistical metrics, derived from the spectral time series, and therefore were significantly reduced in dimension. Experiments using the Random Forest algorithm were performed by building several per-tile models, as well as cross-regional models based on training data from all considered regions/tiles. Overall classification accuracy rates exceeded 90% for most experiments. Further analysis on the experimental results highlighted that crop types were classified more accurately when using the spectral time series features, compared to the temporal ones. Classification accuracy for non-crop classes proved much less affected by the type of employed features. The inclusion of auxiliary data layers was beneficial in all cases, both for overall and for per-class accuracy metrics. Qualitative evaluation on the predicted maps further affirmed the efficiency of the developed methodology.

1. INTRODUCTION

Open availability of high spatial and temporal resolution multi-spectral data, like Sentinel-2 and Landsat 8, has significantly increased the capabilities for various mapping applications, by using image time series. Many recent studies have exploited successfully such data time series, along with machine learning techniques, for land cover and crop type classification tasks (Defourny et al., 2019; Inglada et al., 2017, 2016; Stoian et al., 2019). Nevertheless, the high dimensionality of these multi-temporal multispectral observations poses significant challenges regarding the seamless exploitation and integration of numerous images, but also the efficient management and computationally demanding processing of these massive volumes of data.

In this regard, relevant studies (Pflugmacher et al., 2019; Waldner et al., 2017; Zhang et al., 2020) have exploited temporal metrics derived from the time series, as classification features of reduced dimensions. The use of these features offers also the advantage of independency from the temporal shifts, observed in dynamic classes across extensive or different study areas (e.g., discrepancies of crop calendars) and could tackle possible redundancy issues for the non-dynamic stable classes, too. Still, there is limited literature regarding the relative contribution of these advantages, in comparison to the higher frequency of observations in a time series, which is crucial in order to capture the dynamic phenological patterns of crops (Griffiths et al., 2019; Karakizi et al., 2020).

To this end, in this work we present experiments of joint land cover and crop type mapping, using multi-temporal Sentinel-2 data, over four study areas in Greece, presenting various environmental characteristics. We discuss the contribution of spectral, temporal and auxiliary features, on the classification accuracy of 35 categories, including 18 crop types. In addition we assess different stratification set-ups for training and testing the models and we also perform a thorough qualitative evaluation on the produced land cover and crop type maps.

2. MATERIALS AND METHODS

2.1 Study Areas

The experiments were applied on four different Sentinel-2 tiles: 35TLF, 34TEK, 34SEJ and 34SEG, with a total extent of about 47,000 km^2. These tiles cover part of the three main climate zones present in Greece, according to the climatic stratification of the environment of Europe (Metzger, 2018; Metzger et al., 2005) as presented in Figure 1. More specifically, tile 35TLF covers a part of northern Greece, and a small part of neighboring Bulgaria. Regarding its climatic stratification, it belongs to the Mediterranean North zone. It is covered mainly by broad-leaved forests in its northern part, while on the southern part lays an agricultural plain, bordering the Aegean Sea. Arable crops like maize, cotton, cereals, tobacco and sunflowers are mostly cultivated in this area. A big water body, i.e., Lake Vistonida, covering an area of about 50 km^2, plays an important role for the local climate and along with river Nestos on the western part of the tile, creates a significant fresh-water ecosystem in the area.

Tile 34TEK corresponds to an inland region in the north-western part of the country, mostly belonging to the Mediterranean Mountains zone, as high altitude (~2000m) masses and multiple smaller hills are scattered across the region. Covers of natural vegetation like broad-leaved and coniferous forests, natural grasslands, sclerophyllous vegetation and barelands are dominant in the mountainous areas. Agriculture is practiced in small plateaus, with the main crops being cereals, maize, grass fodders but also stone fruit and nut trees. Two of

* Corresponding author

https://doi.org/10.5194/isprs-archives-XLIII-B3-2021-319-2021

the largest lignite mines of the country, covering about 100 km^2 are also located near Ptolemaida city. Two vast water bodies can also be found in this tile, the Orestiada lake and the Polyfytos artificial lake, a result of damming the Aliakmonas river, the longest watercourse in the country.

Tile 34SEJ is located in central Greece and presents a combination of all three Mediterranean zones of the climatic stratification. This tile consists of highly heterogeneous landscapes with 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 like coniferous and broad-leaved forests, sclerophyllous vegetation and barelands. On the eastern part of the tile, lies a significant part of the Thessalian Plain, the largest agricultural zone of the country. Agricultural land consists mainly of cotton, maize, cereals, and grass fodders. The study area also includes several vast water bodies, like the man-made lakes Plastira and Kremaston, and part of the Amvrakikos Gulf.

Tile 34SEG covers the biggest part of western Peloponnese and belongs for most of its extent to the Mediterranean South zone, while bordering the sea to the west and south. Thermo-Mediterranean brushes and sclerophyllous shrubs cover an extensive part of this region from lower to higher altitudes. Mount Taygetos and mount Mainalo can be found on the eastern part of the tile, covered mostly by coniferous forests and sclerophyllous vegetation. Regarding agriculture, olive groves are dominant and vastly cultivated in this area. In addition, grape vines, citrus trees, potatoes, and vegetables, many of which under cover (greenhouses), are crop types that can be also found in this region. The large lignite mine of Megalopolis is located in the eastern part of the tile, too.

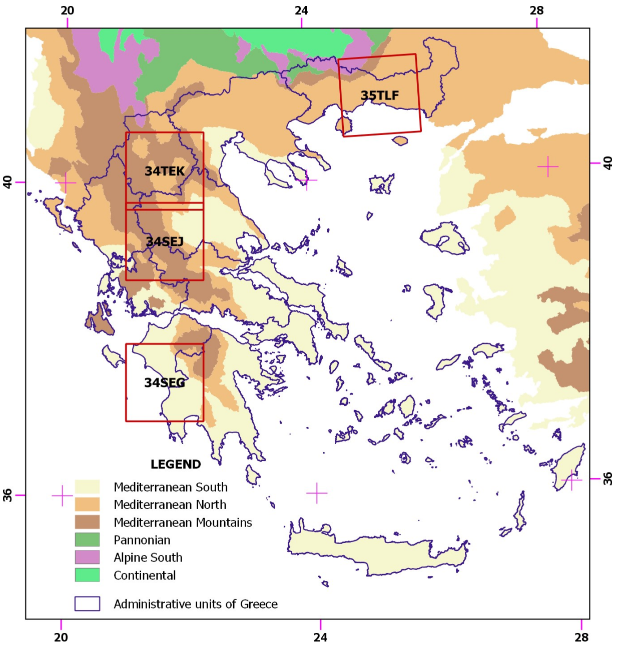

Figure 1. The selected Sentinel-2 tiles overlaid on the Environmental Stratification of Europe map (Metzger, 2018; Metzger et al., 2005).

For all studied tiles, urban regions cover relatively small areas with most cities recording extents of less than 100 km^2, except for cities of Alexandroupolis (650 km^2) in tile 35TLF, Trikala (600 km^2) in tile 34SEJ and Kalamata (450 km^2) in 34SEG. Sparse urban fabric regions, corresponding to small villages are scattered across all study areas.

2.2 Classification Nomenclature and Reference Data

One of the main contributions of this work is the high thematic analysis of the classes' nomenclature. We aim at a detailed land cover and crop type product with 35 classes, including 18 crop types. In order to design the land cover classes nomenclature, we took into consideration the nomenclatures of related land cover and/or crop type products and classification systems, such as CORINE Land Cover (CLC), Greek land cover databases (Ilots and Subilots), the FAO Land Cover Classification System (LCCS), the land parcel information system (LPIS), the LUCAS survey, the Eagle framework and other national land cover products (Inglada et al., 2017). Nevertheless, we based the formation of the classification nomenclature on the definition of spectrally and contextually homogeneous classes at the respective spatial resolution of target (10m). Further parameters taken into account were class representativeness, relative extent, and importance for the area of study. Crops' nomenclature in particular, was designed by also taking into consideration agronomical and spectral distinctness or similarity (grouping of classes, e.g. cereals) between different types, but also the cultivation extent and importance in the agricultural economy of Greece.

In this context, the classification nomenclature included the following land cover categories: Dense urban fabric (DUF), Sparse urban fabric (SUF), Industrial/Commercial units (ICU), Road/Asphalt networks (RAN), Photovoltaic units (PHV), Broad-leaved (BLF) and Coniferous (CNF) forests, Natural grasslands (NGR), Dense (DSV) and Sparse sclerophyllous vegetation (SSV), Sparsely vegetated areas (SVA), Beaches/Dunes/Mines (BDM), Bare rocks (RCK), Wetlands (WTL), Water courses (WCR), Water bodies (WBD) and Coastal water (CWT). Crop-type classes included: Olive groves (OLG), Grape vineyards (VNY), Citrus trees (CTR), Pome trees (POM), Stone fruit trees (STN), Nut Trees (NUT), Rice (RIC), Cereals (CRL), Cotton (CTN), Maize (MAI), Tobacco (TBC), Sunflowers (SUN), Legumes (LEG), Potatoes (POT), Vegetables (VEG), Grass fodders (FDR), Greenhouses (GRH) and Fallow (FLW).

Reference data for the classification categories were created through an intensive image interpretation procedure, by photo-interpretation experts. Digitization was performed on very high resolution imagery, like Google Earth and Bing satellite data, Planetscope data and orthophotos of the National Cadastre and Mapping Agency. Various geospatial datasets were used as ancillary information, e.g., geospatial data from field campaigns conducted by our laboratory, national geospatial information layers, Copernicus Global Land Service products and high-resolution satellite datasets. Particularly for arable crop classes, geospatial data from the Greek Paying Agency and LPIS were used for the creation of reference data, following some processing steps towards the exclusion of erroneous ("outliers") records. At the same time, adhering to good practice for the creation of reference datasets, sample size per class was kept relative to each class's respective occurrence in the study area.

2.3 Sentinel-2 Data and Preprocessing

Sentinel-2 L1C data of less than 20% cloud coverage for the year 2019 were acquired from ESA Sci-Hub for the four above mentioned Sentinel-2 tiles. Several pre-processing steps were applied, towards the seamless exploitation of the studied time series of multispectral images. The initial step was atmospheric correction, using the Sen2Cor processor, followed by a BRDF (Bi-directional Reflectance Distribution Function) correction using a semi-empirical kernel-based as described in Su et al. (2009). Furthermore, cloud and shadow screening was performed, applying the F-mask algorithm (Zhu and Woodcock, 2012) and interpolated values were produced for gap-filling, using the median function. Following that, multi-temporal mis-registration errors, which are often present in Sentinel-2 scenes (Yan et al., 2018), were addressed applying the AROP (Automated Registration Orthorectification Package) method (Gao et al., 2009). As a final step, the medium (20m) resolution bands of Sentinel-2 were sharpened from the, spectrally nearest, high (10m) spatial resolution band, based on the High-Pass Filter (HPF) fusion algorithm.

2.4 Classification Algorithm, Features and Set-Ups

Recent literature has demonstrated the efficiency of shallow machine learning frameworks, compared to deeper architectures, in terms of a better trade-off between computational cost and accuracy, especially in the case of operational mapping applications (Karakizi et al., 2018; Stoian et al., 2019). For this study we implemented a Random Forest (RF) classifier with Python 3.7.8, using scikit-learn and 100 trees. All experiments were executed on a server running Ubuntu 20.04, with an Intel(R) Core(TM) i7-5820K CPU at 3.30GHz and 48 GBs of RAM.

We adopted two approaches regarding the applied classification features, i.e., one based on time series of spectral features and one based on temporal features, i.e., statistical derivations extracted from the spectral time series per pixel. The selection of the specific spectral and temporal features was based on our previous research efforts (Karakizi et al., 2020, 2018) and on the related literature as analyzed in the Introduction section. The time series (TS) sets were formed for each tile by stacking seven spectral bands, i.e., Blue (Band 2), Green (Band 3), Red (Band 4), Red-Edge (Band 5), NIR (Band 8), SWIR 1 (Band 11), SWIR 2 (Band 12) and three spectral indices, i.e., NDVI, NDWI and NDBI. The TS datacubes consisted of: 290 layers for tile 35TLF, 310 for 34SEJ and 350 for 34SEG and 34TEK, based on the available acquisitions, adhering to less than 20% cloud coverage for each tile. The temporal features (TF) sets were formed for each tile by stacking nine temporal metrics i.e., the 10th, 25th, 50th, 75th, 90th percentiles, minimum, maximum, mean values, and standard deviation for all ten spectral bands and indices. For each tile, the TF datacubes had the same number of features, i.e., 90.

Additionally, the contribution of three auxiliary features was assessed for both sets (TS & TF). Recent studies using multi-temporal satellite data towards land cover mapping, have documented on the enhancement of the classification result, when appending ancillary data like elevation, textural, geographic, bioclimatic and thematic/contextual information (Chen et al., 2015; Pflugmacher et al., 2019; Verde et al., 2020; Zhang et al., 2020; Zhu et al., 2016). To this end, we exploited a digital elevation model (EU-DEM) and the CLC product for the year 2018, from the Copernicus Land Monitoring Service, and the Environmental Zones product from the climatic stratification of Europe (Metzger, 2018).

Classification models were built both on a per-tile/region basis and by combining regions' training data into cross-regional models. Several experiments were performed, applying the TS or TF approach and different combinations of the auxiliary data.

3. EXPERIMENTAL RESULTS AND DISCUSSION

3.1 Validation Framework and Accuracy Metrics

In a recent work (Karakizi et al., 2020) we advocated the use of spatially independent training and testing sets, in order to avoid over-estimation for the accuracy metrics. To this end, for each tile, the reference data were split at the polygon level, using approximately 65% for training and keeping 35% for testing. The validation of the classification experiments was quantitatively implemented by forming confusion matrices at the pixel level, expressing the agreement between predictions and testing data labels.

The standard accuracy metrics of overall accuracy (OA), user's and producer's accuracy (UA and PA) were calculated. Per class F-measure (F1) scores were also calculated as the harmonic mean between UA and PA. In addition, the average F1 from all classification categories (avF1) was calculated, as a single-number indicator of each experiment's efficiency, comparable to OA, but without the bias of class sample size. To assist our analysis regarding crop and non-crop classes, F1 scores were also averaged only for crop classes (avF1crops) and only for, non-crop, general land cover classes (avF1non-crops). Qualitative evaluation was also performed by thorough, intensive observation of the produced maps.

3.2 Quantitative Comparative Analysis

Results from experiments using the TS and TF features, form the basis of the comparative analysis for the per-tile experiments. Additional experiments using auxiliary data and cross-regional/single classifiers are also presented and discussed. For the sake of brevity we do not present results from all combinations of auxiliary information, since the exploitation of the full-set had the best performance. Nevertheless, several insights and remarks from those experiments are discussed in the following paragraphs

3.2.1 Spectral Time Series VS Temporal Features: Figure 2 presents OA, avF1, avF1crops and avF1non-crops rates, for the experiments performed on a per-tile base, with various features' set-ups. In bars of blue hue, rates for the TS experiments are presented, while in bars of red hue, rates for the TF ones. For both TS and TF darker shades represent full-set experiments, which include auxiliary data. OA exceeded 80% and F1 69% in all cases. For all tiles, TS features gave better classification results compared to TF. OA differences between the two sets were at most 3%; nevertheless, the avF1crops metric exhibited the largest differences, up to 8%, with every case in favor of TS. The latter highlights the fact that the higher number of features present on the TS set, due to the higher frequency of observations, was critical for a more accurate detection of the crop classes. The classification of different crop types corresponds to a more complex task than the classification of non-crop land cover classes, whose spectral behavior is less variant across the year. This is also apparent in the comparison between the avF1crops and avF1non-crops rates, with the first one in the range of 56-87% and the second in the range of 82-94%. In support of that, avF1non-crops rates appear significantly less affected, from using either TS or TF features for the classification procedure.

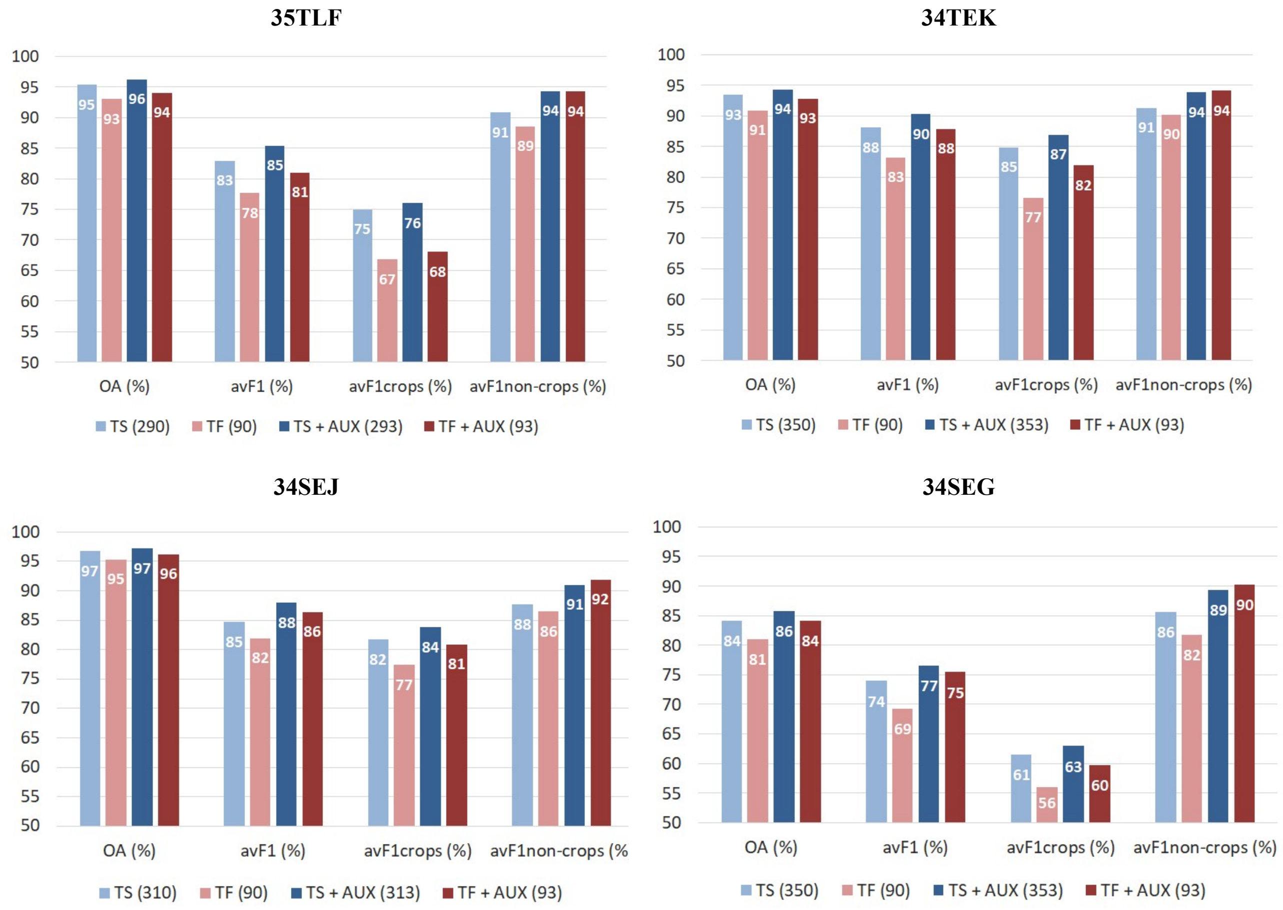

Figure 2. Accuracy metrics for the performed per-tile experiments using different sets of features (TS: spectral time series, TF: temporal features, TS+AUX: TS and auxiliary data layers, TF + AUX: TF and auxiliary data layers).

3.2.2 Contribution of Auxiliary Data: The contribution of auxiliary information on the classification was beneficial in all cases. Improvements in rates ranged from 1-5% and 3-8% for avF1crops and avF1non-crops respectively, while for OA reached up to 3%. As expected, the influence of appending three more features on the TF datasets of 90 layers, was greater compared to the impact on the TS datasets, composed of 290-350 layers. In this sense, for experiments using auxiliary data, the disparity in accuracy metrics observed between the TS and TF approach is smaller, in comparison to experiments that did not utilize auxiliary data.

Further analysis on additional experiments, by including or excluding layers of the auxiliary data from the classification process, highlighted certain remarks, regarding their contribution in accurately classifying each class and preventing classification errors between spectrally similar classes. In particular, elevation information from the EU-DEM product, in most cases, considerably improved the classification results for natural vegetation classes and crop classes that present spectral similarities, such as NGR and FDR, OLG and SSV or DSV, and for high brightness artificial classes (DUF, SUF, ICU) and bareland classes (SVA, BDM, RCK). Mixing and classification errors between those high brightness artificial and bareland classes were also significantly reduced, by incorporating thematic/contextual information from the CLC18 product. The geographic/bioclimatic information derived from the Environmental Zones product had a slight positive effect mainly for certain crop classes.

3.2.3 Variation in Classification Performance of Different Regions: Variation in classification performance between the studied tiles can be observed in Figure 2. The most obvious note is that tile 34SEG presented the lowest rates for all metrics. OA for all other three tiles exceeded 90%. The highest rates were achieved for tile 34SEJ, upwards of 77% for all metrics and set-ups. Comparing between 34SEJ and 34SEG, albeit more data were available for the latter, i.e., 35 multi-spectral images instead of 31 for the former, 34SEG exhibited the lowest accuracy rates and 34SEJ the highest. This can be directly attributed to differences in sample size of the reference data for these tiles. In particular, 34SEJ had the largest amount of reference data (~1.6 million pixels) since this region represents a highly heterogeneous landscape, as mentioned in subsection 2.1. On the contrary, tile 34SEG had the lowest amount of reference data (~0.8 million pixels), since half of its extent is covered by the sea. Furthermore, as also documented in subsection 2.1, regarding agriculture, the cultivation of OLG prevails, and many other crops like POM, STN, NUT, RIC and LEG are cultivated in a relatively limited extent and so a smaller number of reference data is available for these categories. Tiles 35TLF and 34TEK presented a similar total sample size of reference data, around 1.5 million pixels. In the cases of 35TLF and 34TEK, OA rates for TS experiments were over 93% and for TF over 91%, while avF1 over 83% and 78% respectively.

3.2.4 Single VS Per-tile Models: In order to examine the impact of cross-region class variability in the training process of the classifier, additional models were built by combining TF data from all four tiles. Similar experiments based on TS could not be performed, since the time series cubes had different feature counts, dependant on the available acquisitions for each tile, while the temporal distributions of acquisitions were also different across the four tiles studied.

Table 1 shows accuracy metrics for the testing assessment performed separately on each tile, when using the model trained on the same tile (per-tile), and when using the cross-regional model (single), trained with data from all tiles. In all these cases the features' set-up is TF+AUX, i.e. temporal features plus all available auxiliary features. Experiments using models trained and tested on the same tile, presented slightly higher accuracy rates for tiles 35TLF, 34TEK and 34SEJ. This can be attributed to the fact that, the classifier is trained and tested using the same spectro-temporal cube across one region/tile. However, that was not the case when testing the performance on tile 34SEG, since per-class average metrics (avF1, avF1crops, avF1non-crops) were higher when using the single classifier. As described on the previous section, tile 34SEG had a limited number of reference data, especially for certain crop classes, compared to all other tiles. In this sense, incorporating more training data from the other regions proved beneficial for many categories. Enhancements in accuracy when using the single model, were also recorded for the other tiles too, for certain low (per-tile) sample size categories, like GRH in 35TLF, TBC and VEG in 34TEK and ICU in 34SEJ.

	35TLF		**34TEK**		**34SEJ**		**34SEG**	
Model / Metric	*per-tile*	*single*	*per-tile*	*single*	*per-tile*	*single*	*per-tile*	*single*
OA	94.43	94.12	92.71	92.26	96.20	95.69	84.18	83.24
avF1	81.34	80.15	87.70	84.93	86.33	84.79	75.43	76.39
avF1 crops	68.42	67.52	81.69	79.60	81.10	78.27	59.71	61.47
avF1non-crops	94.25	92.78	94.08	90.61	91.57	91.32	90.22	90.43

Table 1. Comparison of the models trained on a per-tile basis, with the single model trained with data from all regions, using the TF+AUX set-up.

Further insights on the relative performance of each class can be derived by the F1 rates presented per class in Table 2 for selected experiments. In particular, the first two "Experiments" columns show the F1 scores for the single model using training and testing data from all tiles, applying the TF and TF+AUX set-ups. The total sample size percentage from all regions is also presented for each class. Relative rates in Table 2 indicate that the contribution of the auxiliary data is positive in the OA and F1 scores of all classes, in agreement with the corresponding analysis for the per-tile experiments (subsection 3.2.2). In particular, the F1 scores of artificial classes DUF, SUF and ICU were considerably improved (8-21%), when using auxiliary data. F1 scores for natural vegetation classes, NGR and SSV, were also increased up to 15%, while bareland class BDM exhibited an 8% increment. Several crop type classes like VNY, STN, NUT, VEG and FLW also yielded increased F1 rates by over 5%, while fruit tree classes CTR and POM showcased a significant improvement, of approximately 20%. Regarding sample size influence on the cross-regional experiments, most classes of high sample size percentage e.g., BLF, CRL, CTN, CNF, MAI, presented F1 scores of over 90% for both set-ups. On the contrary, low sample size for certain classes, like POM, NUT and VNY was associated with lower F1 scores. Nevertheless, this correlation was not consistent across all classes studied, with exceptions, e.g., RAN, PHV, WCR, exhibiting low sample size percentages, but very high F1 scores, most probably related to their distinct spectral characteristics.

The third and fourth "Experiments" columns in Table 2, record the F1 scores when applying the cross-regional (single) model and when applying the per-tile trained model, in order to predict the 34SEG testing data, with the TF+AUX set-up. The sample size percentage for tile 34SEG is also presented for each class. As already mentioned in previous paragraphs, certain low-sample size classes in 34SEG (e.g., STN, RIC, POM, ICU, RAN, SVA), were detected more accurately when applying the single model. Nonetheless, exceptionally low sample size crop classes (LEG, POM, STN) hold very poor classification rates for both cases.

Class Code	**Total Sample Size perc.**	**Experiment:**				**34SEG: Sample Size perc.**
		M: single **T:** All tiles **F:** TF	**M:** single **T:** All tiles **F:** TF +AUX	**M:** single **T:** 34SEG **F:** TF +AUX	**M:** per-tile **T:** 34SEG **F:** TF +AUX	
		F1 scores (%)				
DUF	0.30	71.07	86.37	79.81	80.41	0.41
SUF	0.37	78.67	86.91	91.27	91.67	1.15
ICU	0.33	70.94	92.22	81.92	74.97	0.34
RAN	0.37	91.82	95.85	97.39	95.40	0.45
PHV	0.19	97.26	97.67	99.37	99.53	0.32
BLF	17.37	98.73	99.30	96.68	98.53	2.95
CNF	6.60	96.43	98.25	98.08	97.71	6.67
NGR	0.71	71.82	86.44	75.79	75.53	0.54
DSV	5.07	88.07	90.20	84.49	84.27	14.10
SSV	3.72	81.59	86.58	83.32	84.09	13.75
SVA	1.65	88.07	91.68	91.37	89.91	1.24
BDM	1.21	87.02	94.92	97.10	96.74	2.46
RCK	0.62	89.36	92.87	97.03	96.64	1.59
WTL	2.19	92.30	98.77	97.08	98.50	1.25
WCR	0.23	87.31	89.12	71.91	75.16	0.17
WBD	3.08	98.91	99.61	94.78	94.73	1.08
CWT	3.04	99.32	100.00	100.00	99.98	2.29
OLG	3.64	85.96	90.40	91.79	92.74	16.74
VNY	0.97	67.58	72.86	73.67	76.50	1.77
CTR	0.34	63.76	80.89	78.38	82.42	1.58
POM	0.28	43.62	65.21	4.45	1.35	0.05
STN	0.64	79.66	85.75	41.78	0.00	0.05
NUT	1.34	52.52	59.74	24.80	28.70	0.82
RIC	0.48	98.92	99.28	96.55	85.79	0.09
CRL	11.55	92.71	93.37	50.01	54.93	4.50
CTN	9.25	97.42	97.93	90.06	89.17	3.28
MAI	6.60	98.32	98.49	95.14	95.12	2.60
TBC	1.26	85.35	87.83			
SUN	2.14	88.30	91.22			
LEG	2.03	62.79	67.59	0.00	0.00	0.05
POT	0.80	81.18	86.14	87.82	89.53	1.90
VEG	1.78	65.38	72.81	74.13	76.23	3.67
FDR	7.22	84.86	87.68	61.92	64.10	5.41
GRH	0.58	82.11	85.81	84.52	85.51	3.17
FLW	2.04	47.49	55.36	28.51	33.29	3.55
OA (%)		90.05	92.44	83.24	84.18	
avF1 (%)		81.90	87.57	76.39	75.43	

Table 2. OA and per-class F1 rates for selected experiments (M: model, T: testing data source, F: features' set-up).

3.2.5 Feature importance: In order to validate the contribution of all classification features, we assessed the impurity-based feature importance (Gini index), computed as the normalized total reduction of the criterion brought by each feature, when building the RF classifiers for all experiments. For both TS and TF datasets, the use of spectral indices, and especially NDVI, had a significant contribution in the classification task, independently of the date for TS and the statistical metric for TF. NIR, Red, Red-Edge and SWIR bands followed, while Green and Blue bands were in most cases at the bottom of the ranking.

Regarding the TS models, for tiles 34SEJ and 34TEK, spectral features from images acquired during the summer were higher on the ranking list, for 34SEG features corresponding to spring and autumn proved more important, while for the northernmost tile, 35TLF, spectral features corresponding to all four seasons were found in the top places. Variance in importance, depending on the season the imagery was acquired, can be associated with inter-regional differences in the dynamic classes' behavior along the year, but also on the presence and extent of classification categories in each area. For example, tiles 34SEJ and 34TEK present large extents of arable crops, thus the relatively higher importance of images acquired during the summer can be attributed to the dynamic behavior of agricultural classes during summertime.

Regarding the TF models, the statistical metrics of standard deviation, median, 25%, mean, 10% and 75% were found to be the most important across all spectral features. This indicated that extreme values, i.e., min and max, were not very efficient for discriminating between classes. This can be attributed to, the fact that such values in some cases constitute outliers. The auxiliary data of EU-DEM and CLC18, were for most experiments, in top places regarding their GINI importance, as also documented for similar ancillary information in recent literature (Pflugmacher et al., 2019; Verde et al., 2020). Nevertheless, this remark should be interpreted with caution, since all other competing features are either multi-temporal observations (TS), or derived from said observations (TF), thus presenting high correlations between each other.

3.3 Qualitative Validation on the Map

Numerous maps were produced by applying the developed classifiers. Qualitative evaluation was performed by thorough observation and comparison of the produced land-cover and crop type maps with very high resolution reference imagery. Examples from maps for the southernmost (34SEG) and northernmost (35TLF) tiles are presented hereby.

In Figure 3, a part of tile 34SEG is presented, depicting the city of Kalamata and surrounding areas. The map produced by applying the per-tile classifier is showcased in Figure 3b, while the map resulting from the cross-regional single classifier is presented in Figure 3c. The input data used are the same in both cases, i.e., the TF+AUX set-up. OA rates for both experiments were around 84% and both maps appear quite accurate, when compared to Bing satellite imagery superimposed with the reference data, showcased in Figure 3a. However, slight differences can be observed between the two maps. In particular, on the western part of the map, lays the airport of Kalamata. This area is correctly classified as the artificial classes ICU and RAN in Figure 3c, while in Figure 3b numerous pixels are erroneously depicted as the bareland class BDM. This remark is further supported by the quantitative analysis of the classification results (Table 2), since these three classes hold higher F1 rates when using the single, instead of the per-tile, model.

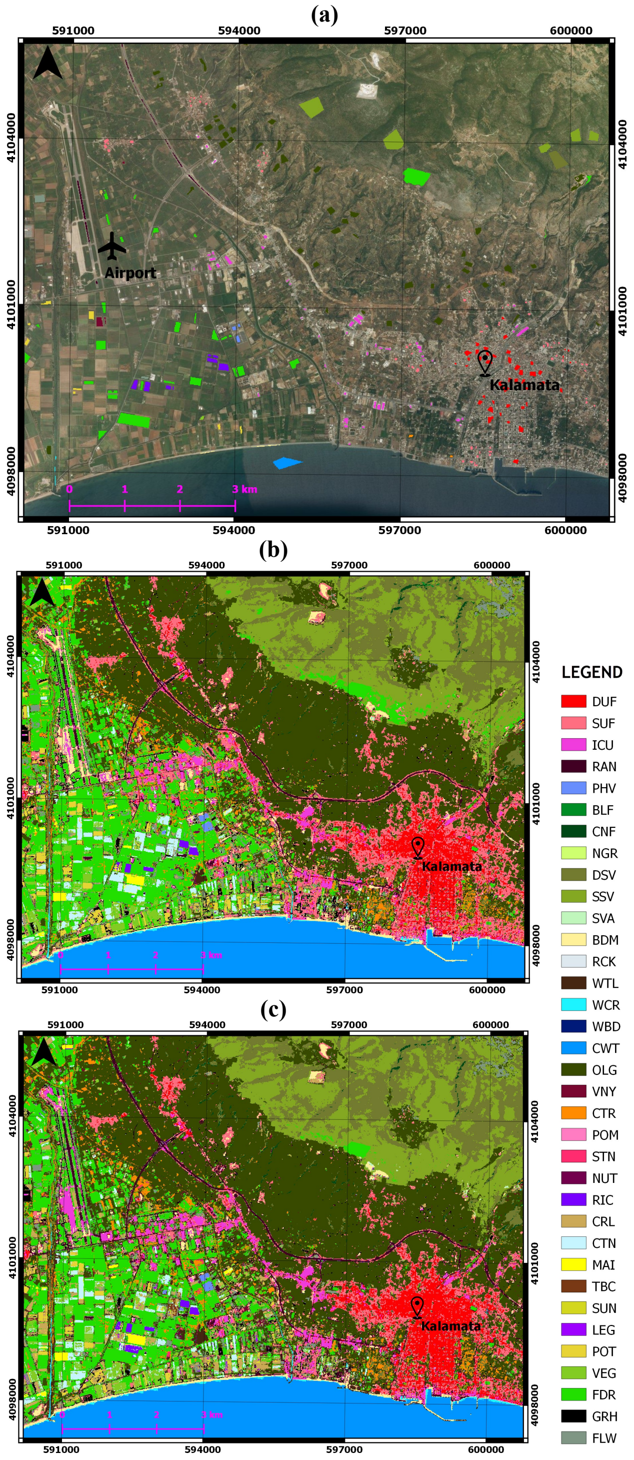

Figure 3. Part of the 34SEG tile, near the city of Kalamata: (a) Reference data over Bing Satellite imagery, the maps produced by applying: (b) the per-tile model and (c) the single model, for the TF+AUX set-up.

In Figure 4 a part of tile 35TLF is presented, depicting the Nestos river delta, and surrounding agricultural areas, where maize and rice are dominant. Presented in the subfigures are maps produced by applying the per-tile classifier with the

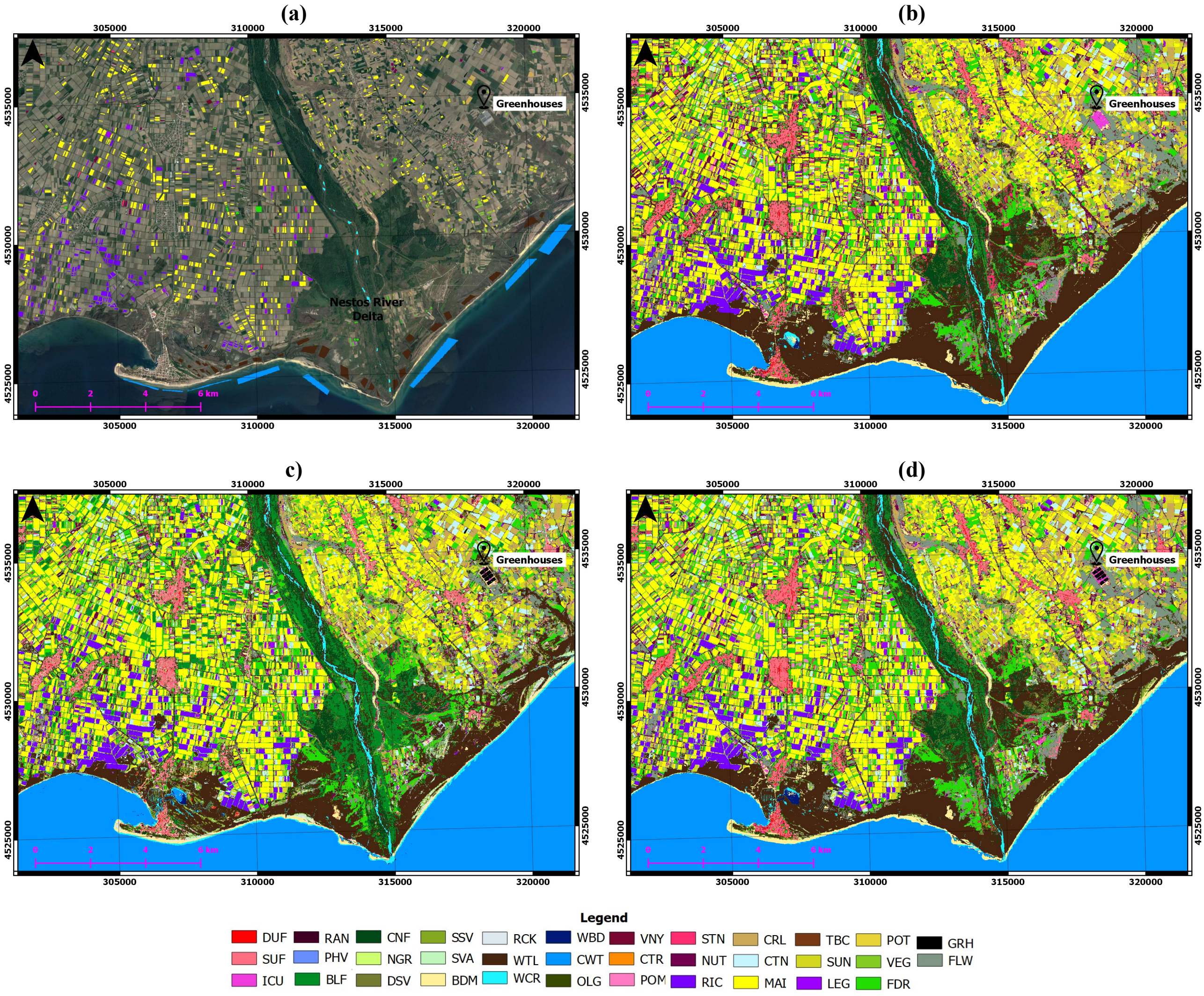

Figure 4. Part of the 35TLF tile, near the Nestos river delta: (a) Reference data over Google Satellite imagery, the maps produced by applying: (b) the per-tile classifier with the TF+AUX set-up, (c) the single classifier with the TF set-up and (d) the single classifier with the TF + AUX set-up.

TF+AUX set-up (Figure 4b), the single classifier with the TF set-up (Figure 4c) and the single classifier with the TF+AUX set-up (Figure 4d). OA for these experiments exceeded 93%.

Comparative analysis of Figures 4c and 4d, regarding the impact of auxiliary data, further supports the corresponding quantitative analysis insights derived from Table 2. In particular, artificial classes (SUF, DUF) appear more accurately detected for villages and small towns in Figure 4d, while Figure 4c displays certain mis-classifications cases towards the class BDM. Agricultural parcels do not present many discrepancies, apart from certain fruit tree parcels that are incorrectly labeled as BLF in Figure 4c. Wetland areas also appear to be more consistently detected when using the auxiliary data. The per-tile model predictions (Figure 4b) in comparison to the single model ones (Figure 4c), present slightly enhanced results regarding the classification of the agricultural parcels and wetland areas. Nonetheless, as also mentioned in subsection 3.2.4, low sample size class GRH is more accurately detected when applying the single model. In this case, the block of greenhouses on the eastern part of the image is classified as ICU using the per-tile model, while it is correctly detected as GRH from the cross-regional single model.

4. CONCLUSIONS

In this paper we presented a methodology for joint land cover and crop type mapping using multi-temporal Sentinel-2 data over four regions belonging to different environmental zones in Greece. The classification task was performed using the RF algorithm towards the classification of 35 land cover classes, including 18 crop-type categories. High OA rates of over 90% were achieved for most experiments. Quantitative comparative analysis on the experimental results, indicated that temporal features, which have a considerably reduced dimensionality, gave results of slightly lower (up to 3%) OA rates, in comparison to the time series spectral features that were approximately triple to quadruple in dimension. Although general land cover classes were insignificantly affected by the type of classification features, crop categories' classification was noticeably better when using the time series approach. In this sense, a higher frequency of observations was proven significant when classifying numerous crop types. The inclusion of elevation and thematic/contextual auxiliary data to the classification procedure was beneficial in all cases. Further analysis on per-class accuracy rates, highlighted that both sample size and spectral distinctness affected the performance of each classification category. Aggregation of all available training data, towards producing a single, cross-regional model, resulted in marginally lower OA rates (<1%) compared to per-tile trained models, but improved the classification accuracy of low (per-tile) sample size classes. Qualitative evaluation on predicted maps affirmed the quantitative evaluation and established the efficiency of the developed methodology.

REFERENCES

Chen, Jun, Chen, Jin, Liao, A., Cao, X., Chen, L., Chen, X., He, C., Han, G., Peng, S., Lu, M., Zhang, W., Tong, X., Mills, J., 2015. Global land cover mapping at 30m resolution: A POK-based operational approach. *ISPRS Journal of Photogrammetry and Remote Sensing, Global Land Cover Mapping and Monitoring* 103, 7–27. https://doi.org/10.1016/j.isprsjprs.2014.09.002

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

Gao, F., Masek, J.G., Wolfe, R.E., 2009. Automated registration and orthorectification package for Landsat and Landsat-like data processing. *JARS* 3, 033515. https://doi.org/10.1117/1.3104620

Griffiths, P., Nendel, C., Hostert, P., 2019. Intra-annual reflectance composites from Sentinel-2 and Landsat for national-scale crop and land cover mapping. *Remote Sensing of Environment* 220, 135–151. https://doi.org/10.1016/j.rse.2018.10.031

Inglada, J., Vincent, A., Arias, M., Marais-Sicre, C., 2016. Improved Early Crop Type Identification By Joint Use of High Temporal Resolution SAR And Optical Image Time Series. *Remote Sensing* 8, 362. https://doi.org/10.3390/rs8050362

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.

Karakizi, C., Karantzalos, K., Vakalopoulou, M., Antoniou, G., 2018. Detailed Land Cover Mapping from Multitemporal Landsat-8 Data of Different Cloud Cover. *Remote Sensing* 10, 1214. https://doi.org/10.3390/rs10081214

Karakizi, C., Tsiotas, I.A., Kandylakis, Z., Vaiopoulos, A., Karantzalos, K., 2020. ASSESSING THE CONTRIBUTION OF SPECTRAL AND TEMPORAL FEATURES FOR ANNUAL LAND COVER AND CROP TYPE MAPPING. I*nt. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* XLIII-B3-2020, 1555–1562. https://doi.org/10.5194/isprs-archives-XLIII-B3-2020-1555-2020

Metzger, M.J., 2018. The Environmental Stratification of Europe.

Metzger, M.J., Bunce, R.G.H., Jongman, R.H.G., Mücher, C.A., Watkins, J.W., 2005. A climatic stratification of the environment of Europe. *Global Ecology and Biogeography* 14, 549–563. https://doi.org/10.1111/j.1466-822X.2005.00190.x

Pflugmacher, D., Rabe, A., Peters, M., Hostert, P., 2019. Mapping pan-European land cover using Landsat spectral-temporal metrics and the European LUCAS survey. *Remote Sensing of Environment* 221, 583–595. https://doi.org/10.1016/j.rse.2018.12.001

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.

Su, L., Huang, Y., Chopping, M.J., Rango, A., Martonchik, J.V., 2009. An empirical study on the utility of BRDF model parameters and topographic parameters for mapping vegetation in a semi-arid region with MISR imagery. *International Journal of Remote Sensing* 30, 3463–3483. https://doi.org/10.1080/01431160802562230

Verde, N., Kokkoris, I.P., Georgiadis, C., Kaimaris, D., Dimopoulos, P., Mitsopoulos, I., Mallinis, G., 2020. National Scale Land Cover Classification for Ecosystem Services Mapping and Assessment, Using Multitemporal Copernicus EO Data and Google Earth Engine. *Remote Sensing* 12, 3303. https://doi.org/10.3390/rs12203303

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

Yan, L., Roy, D.P., Li, Z., Zhang, H.K., Huang, H., 2018. Sentinel-2A multi-temporal misregistration characterization and an orbit-based sub-pixel registration methodology. *Remote Sensing of Environment* 215, 495–506. https://doi.org/10.1016/j.rse.2018.04.021

Zhang, M., Huang, H., Li, Z., Hackman, K.O., Liu, C., Andriamiarisoa, R.L., Ny Aina Nomenjanahary Raherivelo, T., Li, Y., Gong, P., 2020. Automatic High-Resolution Land Cover Production in Madagascar Using Sentinel-2 Time Series, Tile-Based Image Classification and Google Earth Engine. *Remote Sensing* 12, 3663. https://doi.org/10.3390/rs12213663

Zhu, Z., Gallant, A.L., Woodcock, C.E., Pengra, B., Olofsson, P., Loveland, T.R., Jin, S., Dahal, D., Yang, L., Auch, R.F., 2016. Optimizing selection of training and auxiliary data for operational land cover classification for the LCMAP initiative. *ISPRS Journal of Photogrammetry and Remote Sensing* 122, 206–221. https://doi.org/10.1016/j.isprsjprs.2016.11.004

Zhu, Z., Woodcock, C.E., 2012. Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sensing of Environment* 118, 83–94. https://doi.org/10.1016/j.rse.2011.10.028