Aggregation of Sentinel-2 time series classifications as a solution for multitemporal analysis

Stanisław Lewiński*[a], Artur Nowakowski[a], Radek Malinowski[a], Marcin Rybicki[a], Ewa Kukawska[a], Michał Krupiński[a]

[a] Space Research Centre, Polish Academy of Sciences, Bartycka 18A, Warsaw 00-716, Poland

* stlewinski@cbk.waw.pl, phone +48 22 4966286; fax +48 22 8403131; cbk.waw.pl

ABSTRACT

The general aim of this work was to elaborate efficient and reliable aggregation method that could be used for creating a land cover map at a global scale from multitemporal satellite imagery. The study described in this paper presents methods for combining results of land cover/land use classifications performed on single-date Sentinel-2 images acquired at different time periods. For that purpose different aggregation methods were proposed and tested on study sites spread on different continents. The initial classifications were performed with Random Forest classifier on individual Sentinel-2 images from a time series. In the following step the resulting land cover maps were aggregated pixel by pixel using three different combinations of information on the number of occurrences of a certain land cover class within a time series and the posterior probability of particular classes resulting from the Random Forest classification. From the proposed methods two are shown superior and in most cases were able to reach or outperform the accuracy of the best individual classifications of single-date images. Moreover, the aggregations results are very stable when used on data with varying cloudiness. They also enable to reduce considerably the number of cloudy pixels in the resulting land cover map what is significant advantage for mapping areas with frequent cloud coverage.

Keywords: Decision fusion, land cover, machine learning, multitemporal analysis, Random Forest, Sentinel-2

1. INTRODUCTION

The constantly increasing availability and accessibility of multitemporal remote sensing observations such as those delivered by different Sentinel or Landsat missions encourage and facilitate analysis of selected phenomena by considering differences in their spectral response at different time of a year (e.g. seasons). Multitemporal data is being used frequently for land cover/land use mapping [1–3], in which case many relatively dynamic classes with predominant vegetation cover such as agricultural areas, woody vegetation or grasslands are analyzed. Mapping of such classes may particularly benefit from an analysis of an image time series due to the phonological changes taking place, along the growing season, in the plants they are composed of. However, the analysis and fusion of multitemporal data may be performed in different forms and at different levels of the image time series analysis. In most research works all image data from a time series are analyses and processed simultaneously to find multitemporal characteristics of land cover classes that will enable their differentiation [1,4,5]. In some other studies, however, only specific information highlighting most evident phonological characteristics in vegetation growing cycle are applied [2,3,6]. In both cases spectral bands and their derivatives like spectral indices (e.g. NDVI), textural features and transformations (e.g. PCA) may be used.

In our study, however, we propose a different approach to multitemporal data analysis. With the general aim of producing a reliable land cover map based on multitemporal observations, we suggest that instead of fusing all the available image data from a time series it may be worth to combine results of separate classifications of individual images. Therefore, the more specific objective of the study was to find an efficient way for combining the results derived from classifications of individual images from a time series as alternative method to the above-mentioned approaches. By our approach we aimed at increasing thematic accuracy by optimizing land cover recognition and minimizing the number of cloudy pixels in the resulting land cover map. The study presented in this paper was conducted as a part of the European Space Agency (ESA)

founded Sentinel-2 Global Land Cover project (S2GLC) (under Scientific Exploitation of Operational Missions S2-4Sci Land and Water - Study 3: Classification) that aims at developing procedures and providing recommendations for deriving land cover classifications on a global scale from the Sentinle-2 data [7].

2. MATERIALS

2.1 Study area

The work presented in this paper was performed on selected test sites, originating from the S2GLC project, representing different landscape and climate conditions and characterized by various combination of land cover types and degree of cloudiness. These sites include one site in Colombia, two sites in Germany and one site in Namibia (Figure 1). Each of the analyzed test sites covers area of 110 x 110 km^2 that is the size of an individual tile of Sentinel-2 data whose extent and naming is based on the Military Grid Reference System (UTM proj.) with additional extension of 5 km on all sides [8].

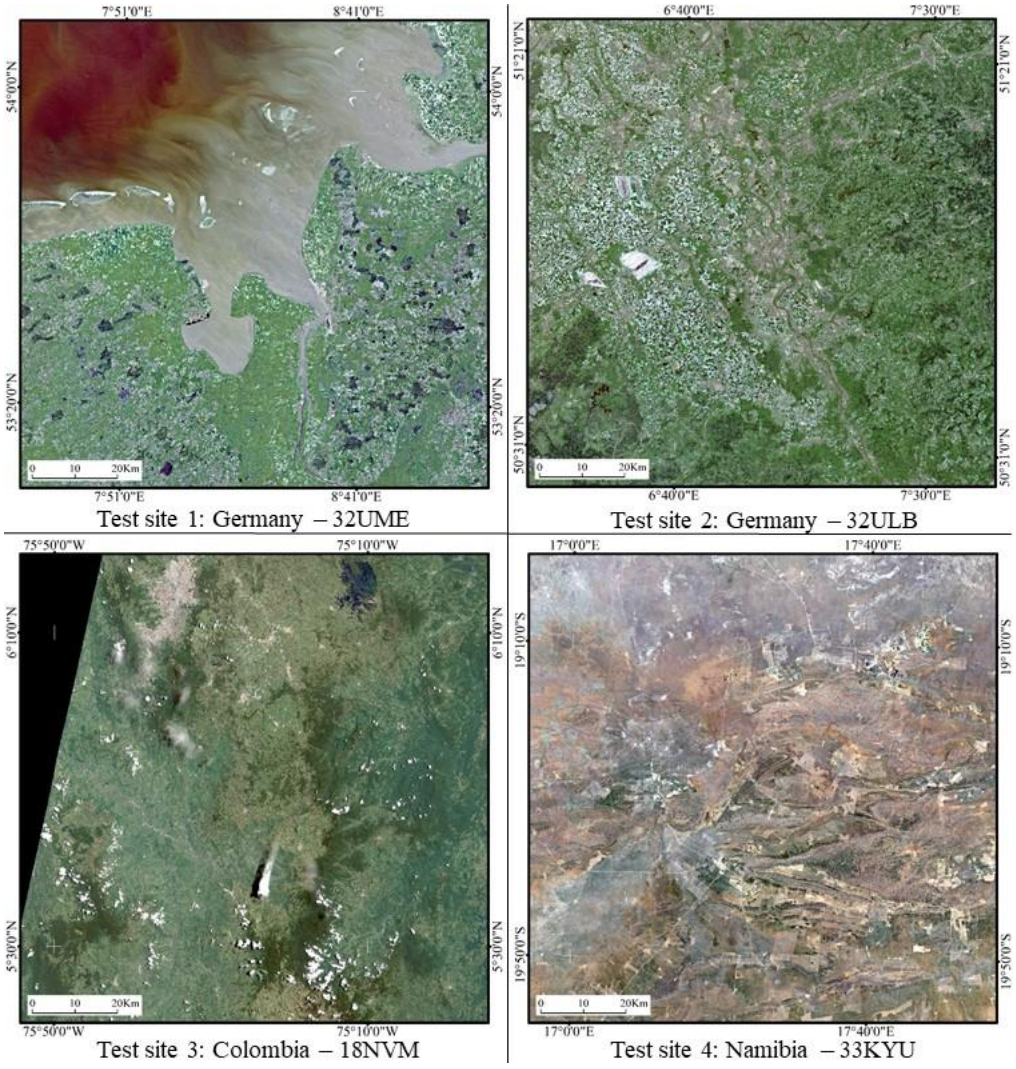

Figure 1. Overview of the study test sites shown on Sentinel-2 images in 4-3-2 band compositions (natural colors) from May 2016 (Germany and Namibia) and September 2017 (Colombia).

2.2 Satellite imagery

To illustrate the performance of the proposed methods we used Sentinel-2 data whose characteristics enable operational land cover mapping on regional, national and a global scale. The first satellite from the constellation, called Sentinel-2A, was launched in June 2015 as part of the Copernicus Program [9] and Sentinel-2B was launched in March 2017. Some aspects make Sentinel-2 quite distinct from other multispectral imaging systems with medium and high spatial resolution operating in the optical visible and reflective infrared spectrum like e.g. Landsat or SPOT. Firstly, the Sentinel-2 deliver imagery from as many as 13 distinctive and relatively narrow spectral bands, from visible to short-wave infrared spectrum. Secondly, data from the Sentinel mission are provided under a fully free and open access policy. These characteristics make Sentinel-2 data the only freely available, high resolution, satellite imagery. Moreover, with the very high temporal resolution of 5 days from a constellation of the two satellites the Sentinel-2 data have become extremely valuable data source and an important tool for frequent and systematic monitoring of the environment. The aspect of frequent revisit period is especially valuable in case when areas with frequent cloud cover are under investigation.
The analyzed image time series consist of eight to thirteen Sentinel-2 images, depending on a test site, that were acquired during 2016 year and represent different periods of a growing season and varying amount of cloud cover from nearly none up to almost 100 % cover.

2.3 Reference dataset

Reference data used for training the classifier and for the purpose of accuracy assessment were collected by means of visual interpretation of high spatial resolution images available in Google Earth™ and the Sentinel-2 data themselves.
Training data were collected by drawing polygons on the reference images that represent the classes of interest. By drawing polygons a great number of pixel samples were collected from which the appropriate amount was used for training the classifier. For validation purpose an independent set of samples was collected by random selection of predefined number of pixels from within the analyzed test areas. The land cover classes of the selected sample pixels were then visually checked on reference images and assigned to be used for accuracy assessment of the individual classifications of Sentinal-2 images and the final aggregation results.
For both cases of training and validation data the effort was made to use images from Google Earth™ that closely matched the acquisition dates of Sentinel-2 images from the used time series. This approach was adopted in order to avoid incorrect class assignment resulting from changes taking place between acquisition of the reference and Sentinel-2 data.

3. METHODS

The overall workflow of the performed study is presented in Figure 2 and shows all processes important for the aggregation of land cover classifications.

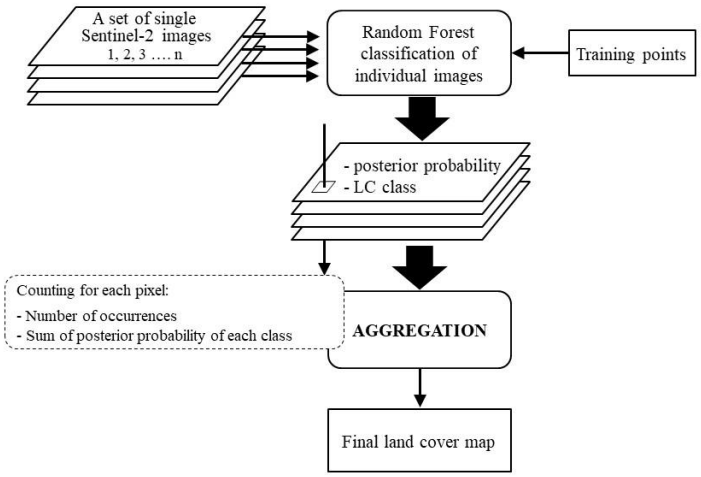

Figure 2. Overall study workflow of aggregation of individual land cover classifications.

3.1 Pre-processing

To perform tests on the aggregation methods all images selected for the analysis were classified separately on a pixel level. These classifications were preceded with pre-processing of Sentinel images disseminated by ESA as Level-1C product with atmospheric correction and resampling of all Sentinel spectral bands to a unified pixel size. The atmospheric correction was performed with Sen2Cor processor that allowed conversion from Top- (Level-1C) to Bottom-Of-Atmosphere (Level-2A) reflectance values [8]. During the correction process additional datasets is created that include among other things information on the extent of different type of clouds and cloud shadows. This dataset was used in the classification process to mask out image pixels and reference data covered by clouds and their shadows. Resampling of spectral bands of Sentinel-2 images was made with the nearest neighbor method to the common output pixel size of 10 m. All Sentinel-2 bands resulting from the atmospheric correction were used in the subsequent classification process.

3.2 Classification of image data

Classification of Sentinel-2 images was performed with a Random Forest (RF) classifier that was selected as the one most suitable for the project needs. RF is an ensemble classifier which train a set of classification trees using bootstrap aggregating (bagging) [10]. Each classification tree in the forest is trained using a training subset generated from an original training set by sampling with replacement. A decision rule in each node of the trained tree is defined based on randomly selected features. The final classification decision is made based on classification results from all decision trees using the plurality voting. The advantages of RF are the high accuracy of its classifications, efficient computation on large dataset and auto-selection of the features used [4,11]. All images from the analyzed time series were classified using the same parameters. The number of trees was set to 50 and the minimum number of samples per node was calculated automatically based on misclassification probability. The number of features randomly selected at each node equals three that corresponds to an approximate value of the square root of the total number of all features. The number of training points was set to 1000 with a minimum count of 50 to be used in case of very cloudy images and limited occurrence of a given class. The land cover classes analyzed are presented in Table 1 and their occurrences differ between the study areas depending on the landscape and climate conditions. This list of classes originates from a broader legend of land cover/land use developed for the need of S2GLC project [7].
The RF classification delivers two forms of outputs and both are used in the aggregation process. The first is the actual map with land cover/land use class labels assigned. The second is a raster dataset presenting information on the number of trees voting for the final class found in the ensemble for each classified pixel divided by total number of trees [12]. This additional estimate is frequently referred to as the posterior probability of a given class [13].

3.3 Aggregation approaches

The aggregation strategy proposed here might be compared to techniques of decision fusion, geographic stacking, ensemble classification, multiple classifier system or staked generalization [4,14–17]. However, the distinct difference is that instead of analyzing the same image with different classification techniques (i.e. classifiers) and combining them into a single result by applying some kind of meta-classifier we use in our approach the same method (i.e. classifier) to classify all individual images from an image time series. This step is followed by the final aggregation of the individual classifications, which process however, might be performed in different ways. Thus, the closest comparison may be done with the decision-level data fusion in which case not a spectral value of individual pixels from the input datasets (pixel-level) or features derived from them (features-level) are fused and analyzed together but only the final output of the analysis (e.g. land cover classifications) [18–20].
In this paper we propose and compare performance of three different methods of aggregation. The first two methods utilize both the classification results and posterior probability data which are included in the process of the final class estimation as a kind of weighting factors. In the first approach, named by us for the need of this study DivByAll, the probability scores are summed up separately for each land cover class found in the investigated pixel from all images of a time series. The sum is then divided by the number of all occurrences in a time series regardless of the classes label. In the second method, named DivByClass, the probability scores are summed up in the same way as in the DivByAll method, however, for each land cover class the sum is divided by the number of occurrences of that particular class (Figure 3). In both methods, the occurrences of pixels covered by clouds are excluded from calculation.
The third method is a plurality voting, for which only the classification maps are used in the aggregation process without including information on the posterior probability. In the plurality voting method the winning class is the one with the highest number of occurrences within the time series. As opposite to the proper majority voting the applied plurality voting

does not require to reach more than 50% of the occurrences [21]. In this way the winning class is simply the one most frequently occurring.

Table 1. The land cover/land use classes analyzed in the predefined test areas.

Class name	**Description**
Artificial surfaces and constructions	All surfaces where landscape has been changed by or is under influence of human construction activities by replacing natural surfaces with artificial abiotic constructions or artificial materials
Consolidated areas	Consolidated surfaces, which are in most parts impervious to water, formed by natural material and with a solid surface. It also may have been modified through man-made processes like on extraction sites
Un-Consolidated areas	Any surface with loose mineral particles of any size range, either as an outcome of natural physical sedimentation processes or human activity. E.g. mountain slope debris, glacier moraines, river pebble banks, beaches, sand dunes and quarries
Evergreen coniferous tree cover	Land covered with needle-leaved tree canopy that do not seasonally lose needles
Deciduous broadleaf tree cover	Land covered with broadleaved tree canopy that seasonally lose leaves
Bush, shrub	Any type of vegetation with woody character (ligneous stem) and with a growth form and height between herbaceous and trees. This class also includes dwarf shrubs
Herbaceous vegetation	Lands covered by natural herbaceous vegetation
Cultivated and managed areas	Land surface managed by human (including temporary bare soil) - arable land, permanent and temporary crops
Water bodies	Water, either in liquid or in frozen solid state
Permanent snow-covered surfaces	Snow cover that persists throughout the year, above the climatic snow line. Persistent ice cover formed by accumulation of snow
Areas permanently covered by clouds	Areas covered by clouds in all images from a time series

Figure 3 presents an example of the aggregation rules applied to a single pixel and illustrate the calculation process for all three methods used. An important aspect of the presented methods is that the individual classifications used are not weighted neither according to acquisition date, cloud coverage or individual thematic accuracy. The only weighting factor is the posterior probability applied to each pixel individually as explained above.

The three aggregation methods presented above were applied to the land cover maps derived from time series of single-dates Sentinel-2 images from four analyzed areas (Table 2a-d). For the 32UME study site (Germany), additional two scenarios were followed, in which all three aggregation approaches were tested on two sets of land cover maps derived from initial classifications with RF. In the first scenario (S1), the set was composed of five land cover maps derived from five the least cloudy Sentinel-2 images from the available set, while in the second scenario (S2) five land cover maps derived from images with the greatest cloud cover were used. The idea of this analysis was to check whether the aggregation methods provide stable results while performed on land cover maps with different cloud coverage.

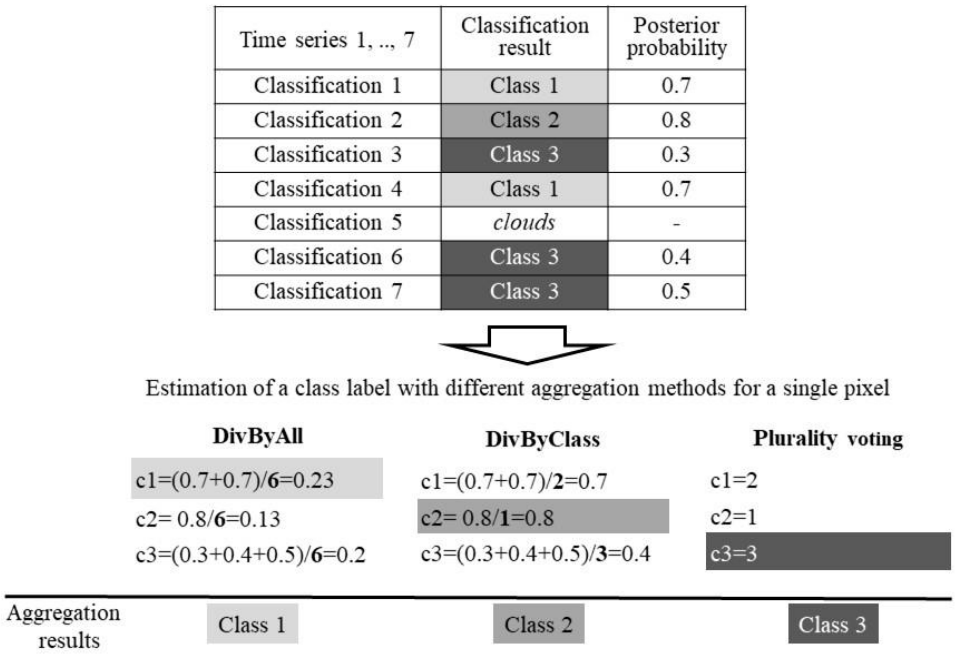

Figure 3. Example of estimation of a class label for a single pixel with three aggregation methods, DivByAll, DivByClass and Plurality voting.

4. RESULTS

In order to address the goal of the study we calculated accuracy scores for all individual classifications and compared them with accuracy measures derived for the results of the three aggregation methods proposed. The accuracy score used here is the overall accuracy (OA) which indicate the overall performance of the classification by presenting the ratio of the correctly recognized samples for all classes and the total number of used samples. The second accuracy measure is the Kappa coefficient which represents the measure of agreement between model predictions and the reference data [22].

Figure 4a – d presents accuracies for the analyzed study areas including German 32UME and 32ULB, Namibian 33KYU and Colombian 18NVM sites, respectively. The presented data allows direct comparison between the performed aggregations and individual classifications. Additional accuracy statistics are presented in Table 2a – d. An example of a land cover map derived with DivByAll aggregation method is presented in Figure 5 for a fragment of 32UME test site.

Due to the apparent correlation between OA and Kappa coefficient we focus our analysis on the former. Considering the OA there was no single best aggregation method and the performance of the proposed aggregations are in most cases at the level of the highest accuracy results of the individual classifications or above them and reach circa 90% for German sites and nearly 71% and 60% for Namibian and Colombian sites, respectively. This is the case for DivByAll and Plurality approaches, which in two analyzed areas (i.e. 32ULB and 18NVM) outperformed the single-date classifications and at two other are at similar level. The DivByAll method reaches the highest scores in both Germans areas while Plurality method is slightly more efficient in Namibian and Colombian sites. The OA differences between DivByAll and Plurality methods varies from 0.1–4.0%. The DivByClass method tend to perform the worst in all analyzed areas and its scores are lower by 4.0–8.0% from the best performing aggregation. Only in one case of the Colombian test site the DivByClass method outperforms the result of a single-date classification.

The results of additional aggregations performed for the test site 1 – 32UME are presented in Table 3 that provides OA statistics for the three aggregations and two tested scenarios. These results confirm the previous statistics received for this test site. The DivByAll and plurality voting got higher OA scores then the DivByClass method. The DivByAll method achieved the highest scores in both analyzed scenarios. The difference, however, between the two more accurate methods and the DivByClass approach was in this case lower than the one received while analyzing all available images (max. 2.6% of OA).

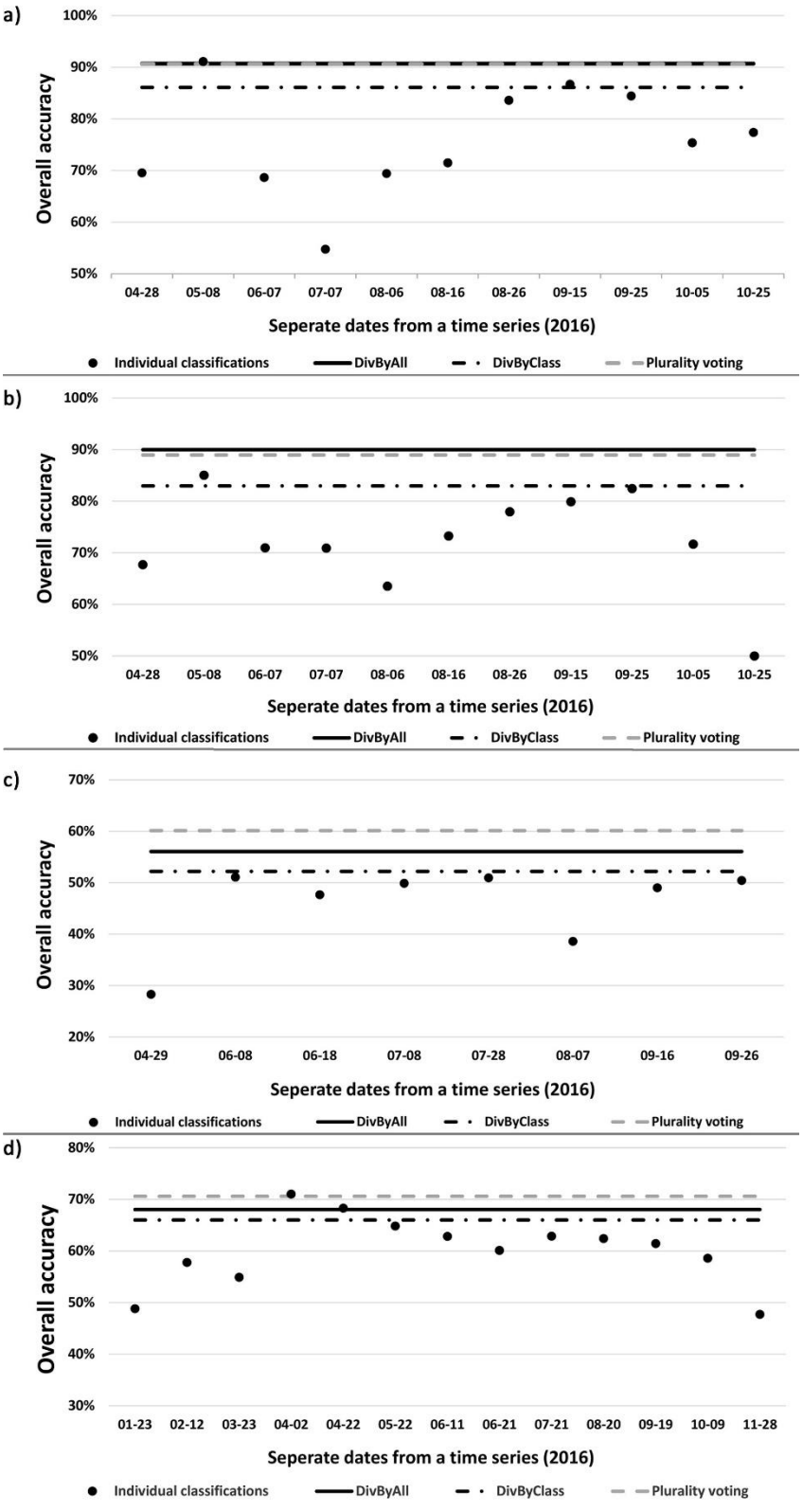

Figure 4. Graphs presenting overall accuracies of the three aggregation methods and classifications of single-date Sentinel-2 images for 32UME (a), 32ULB (b), 18NVM (c) and 33KYU (d) test sites.

Table 2. Details on Sentinel-2 images used in the study with information on cloud coverage and thematic accuracy of individual classifications and aggregation methods applied.

a) 32UME test site

Images of a time series	Acquisition date	Cloud cover [%]	Overall accuracy [%]	Kappa coefficient
1	2016-04-28	48.2	69.5	0.48
2	2016-05-08	0.8	91.1	0.88
3	2016-06-07	21.2	68.7	0.60
4	2016-07-07	45.5	54.7	0.43
5	2016-08-06	54.5	69.4	0.60
6	2016-08-16	45.0	71.5	0.62
7	2016-08-26	3.2	83.6	0.78
8	2016-09-15	1.2	86.7	0.83
9	2016-09-25	29.5	84.4	0.80
10	2016-10-05	39.8	75.4	0.68
11	2016-10-25	0.6	77.4	0.69
Aggregation	**DivByAll**	**0.2**	**90.7**	**0.88**
	DivByClass		**86.1**	**0.82**
	Plurality voting		**90.6**	**0.88**

b) 32ULB test site

Images of a time series	Acquisition date	Cloud cover [%]	Overall accuracy [%]	Kappa coefficient
1	2016-04-28	64.7	67.7	0.59
2	2016-05-08	2.9	85.0	0.81
3	2016-06-07	33.2	71.0	0.63
4	2016-07-07	45.2	70.9	0.64
5	2016-08-06	62.5	63.5	0.54
6	2016-08-16	5.2	73.3	0.66
7	2016-08-26	8.6	78.0	0.72
8	2016-09-15	21.3	79.9	0.75
9	2016-09-25	2.7	82.5	0.78
10	2016-10-05	17.4	71.7	0.64
11	2016-10-25	99.7	50.0	0.0
Aggregation	**DivByAll**	**0.9**	**90.0**	**0.87**
	DivByClass		**83.0**	**0.79**
	Plurality voting		**89.0**	**0.86**

c) 18NVM test site

Images of a time series	Acquisition date	Cloud cover [%]	Overall accuracy [%]	Kappa coefficient
1	2016-01-23	21.7	48.8	0.34
2	2016-02-12	41.4	57.8	0.36
3	2016-03-23	7.0	54.9	0.32
4	2016-04-02	1.9	71.0	0.53
5	2016-04-22	3.7	68.3	0.49
6	2016-05-22	0.1	64.8	0.45
7	2016-06-11	0.1	62.9	0.42
8	2016-06-21	0.1	60.1	0.38
9	2016-07-21	0.1	62.9	0.41
10	2016-08-20	0.1	62.4	0.41
11	2016-09-19	0.3	61.5	0.41
12	2016-10-09	0.1	58.6	0.39
13	2016-11-28	20.9	47.7	0.26
Aggregation	**DivByAll**	**0.01**	**68.0**	**0.49**
	DivByClass		**66.0**	**0.44**
	Plurality voting		**70.6**	**0.51**

d) 33KYU test site

Images of a time series	Acquisition date	Cloud cover [%]	Overall accuracy [%]	Kappa coefficient
1	2016-04-28	52.9	28.3	0.13
2	2016-05-08	35.8	51.1	0.35
3	2016-06-07	46.3	47.7	0.28
4	2016-07-07	44.4	49.9	0.33
5	2016-08-06	18.1	50.9	0.36
6	2016-08-16	38.6	38.6	0.26
7	2016-08-26	25.8	49.0	0.34
8	2016-09-15	40.8	50.4	0.27
Aggregation	**DivByAll**	**0.47**	**56.1**	**0.4**
	DivByClass		**52.2**	**0.34**
	Plurality voting		**60.1**	**0.44**

Table 2a – d presents also information about the cloud cover on the analyzed Sentinel-2 images and the amount of cloud found on the final land cover map resulting from the aggregations. It can be seen that for each analyzed region the map resulting from aggregation (all aggregation methods result in the same cloud coverage in the final map) provides considerably reduced cloud cover. The difference in the cloud coverage between the classification of the least cloudy Sentinel-2 image and the aggregation map was 0.38% for 32UME, 1.83% for 32ULB, 0.01% for 33KYU and 17.6% for 18NVM sites.

Table 3. Overall accuracy of separate aggregations performed on sets of five land cover classifications derived from five the least and the worst cloudy Sentinel-2 images for the test site 32UME.

Aggregation scenarios	Overall accuracy [%]		
	DivByAll	**DivByClass**	**Plurality voting**
S1 - 5 least cloudy land cover maps	91.3	88.6	90.5
S2 - 5 most cloudy land cover maps	88.4	86.3	88.1

5. DISCUSSION

From the point of view of thematic accuracy, the application of the proposed aggregation methods (i.e. DivByAll and Plurality voting) provided satisfactory and improved results while compared to accuracies of classifications of single-date images. Therefore, the stated objective of the study was accomplished as the proposed method increased or at least maintained the thematic accuracy of a land cover map while minimizing the cloudiness in the map. Considerably lower overall scores were received (for all three methods) for the sites in Columbia and Namibia then in Germany (from 20% to 30%) but this is due to much lower overall performance of individual classifications in those areas. The analysis of a problem of lower individual and aggregation accuracies in Columbia and Namibia is not, however, within the scope of this paper.

The aggregation provides the greatest benefits especially in cases with frequent cloud cover in analyzed area. In such case only part of the image scene could be classified with a single-date image. Aggregation, however, enable to combine data represented on images from multiple dates and select only those which provide real information (Figure 5). If any pixel is classified as a cloud during the pre-processing step of atmospheric correction it will be assigned with lower weight. This will result in reclassification of such pixel to any other land cover class if such occur for the analyzed pixel in the imagers from a time series. Therefore, the only constraint on applying the aggregation technique is availability of a reliable cloud mask. This issue, however, remains questionable as it was shown that the automatically derived cloud mask is not always dependable [23]. This was the case for all areas analyzed in this study. A good example of the advantage of using aggregation is provided with the study site from Colombia. Even though the least cloudy image was covered with clouds by over 18%, the aggregation provided a map with only 0.47% of cloud cover eliminating cloudy pixels significantly.

Plurality voting was found as one of the two most accurate aggregation methods. Similarly as in other studies [4,15] this method provided stable but not necessarily the highest accuracies. However, in most cases it still managed to outperformed individual initial classifications. Besides its good performance it has the advantage that can be used in situation when an initial classification of single-date images does not provide information on posterior probability or when it is difficult to interpret such information.

The lowest performance was observed for the DivByClass method which was the case in all study sites. From the computational point of view this method tends to favor candidates with lower number of occurrences. This is because the division in the computing formula is made for each separate class by the number of occurrences only of that particular class as opposite to DivByAll method, in which the division is always made by the number of all valid classifications regardless of the class label. Therefore, the less occurrences of some class the less its final probability is averaged. In this way even the class with low probability score and possibly incorrect class recognition but singular occurrence may often be selected as a winner. This is expected to be the reason for lower accuracy scores achieved by this method.

The additional analysis conducted with sets of images with different level of cloudiness also provides important founding. It was shown that the decrease in OA was only in the range of less than 3% while the cloudiness in the two sets analyzed was considerably different. The first set was composed of data with 0.65%, 0.8%, 1.2%, 3.2% and 21.2% of cloud cover. The second included data with 39.8%, 45.0%, 45.5%, 48.2% and 54.5% of cloudiness. Therefore, the cloud cover of the worst image from the set with low cloudiness was twice smaller than the least cloudy image of the set with high cloudiness. This big difference was sufficiently overcome by using the aggregation approach.

It is worth to mention that the performance of aggregation (all methods) is highly dependent on the quality of images from a time series used for the initial classification. It will result in greater disagreement between land cover classes originating from individual classifications composing the aggregated map and will probably reduce its final accuracy. The issue of inaccurate image co-registration, however, would most probably concern any method used for analysis of multitemporal data and data from any sources. A problem of misregistration of Sentinel-2 images mentioned in Kukawska et al.(2017) [23] may serve here as an example for this problem. However, no analysis was performed to assess to what degree this issue influenced results presented in this paper.

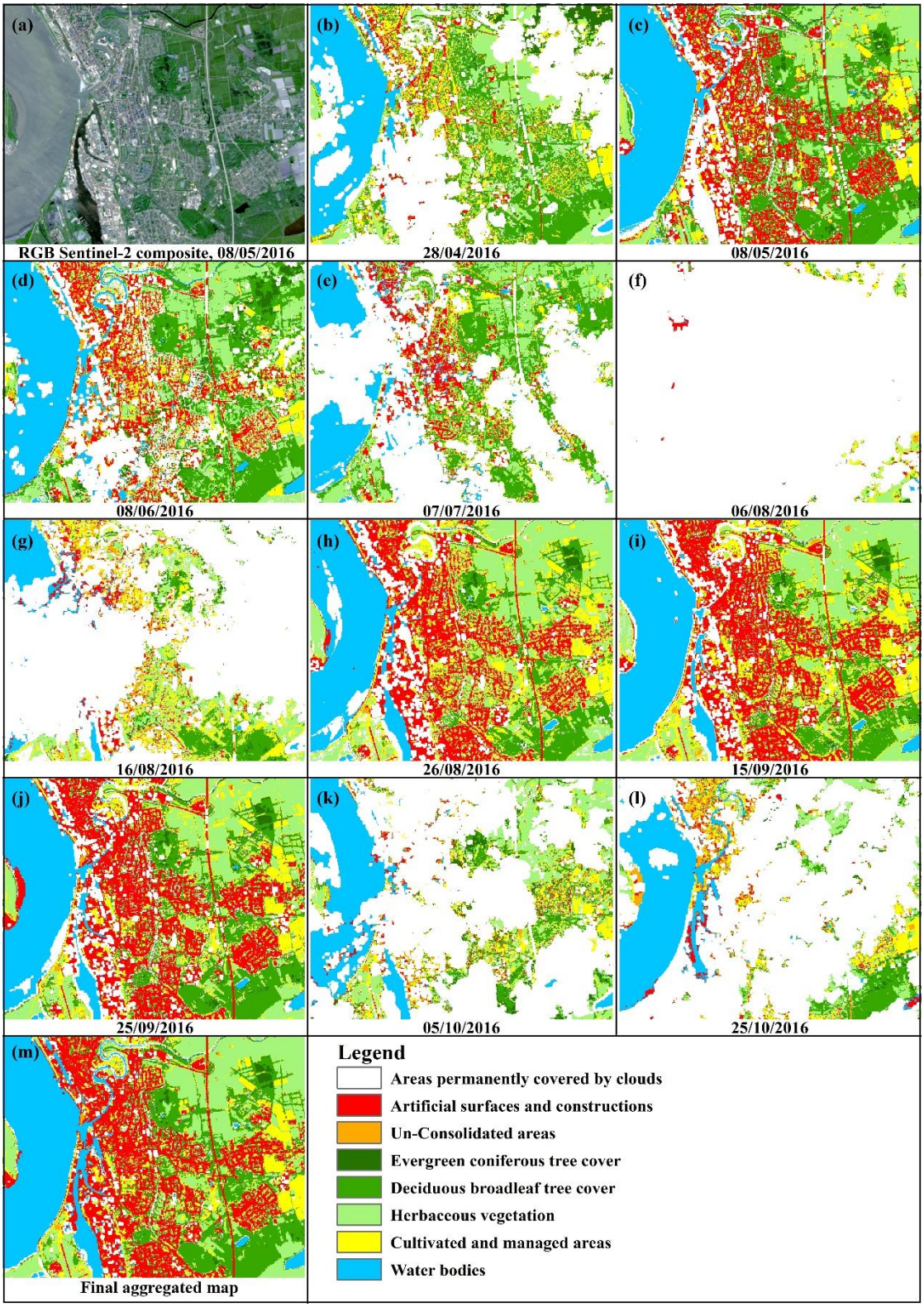

Figure 5. Land cover classification maps from single-date Sentinel-2 images ((a) – (l)) and an aggregated land cover map resulting from the DivByAll aggregation method (m) for a extract of the test site 1 – 32UME. The map presents area of city Bremerhaven, Germany. The presented legend includes only those classes that were recognized on that test site area.

6. CONCLUSIONS

Our research presented above demonstrates application of aggregation techniques, which allow to combine information from multiple land cover maps derived from separate classifications of satellite images from a time series. Two methods, namely the DivByAll and Plurality voting, show superior performance over the third analyzed approach, the DivByClass. The indicated methods, depending on the settings, outperform classifications of single-date images or reach the same level of accuracy. In addition to this, the proposed methods maximize the use of cloud free pixels from the multiple input data allowing for considerable reduction of cloudiness in the resulting land cover map. This property of the aggregation technique is a very important advantage when areas characterized by frequent cloud cover are being analyzed. It allows to utilize the most up-do-date image data, acquiring during the year of interest, without the need for extending analysis to data from previous years that not necessary represent the most current situation. Similarly, it facilitates land cover analysis in the climatic zones where the cloudy season is the most vital and dynamic period from the perspective of analysis of growing conditions.

The next important step in further development of the aggregation technique will be an analysis of changes in land cover taking place during the analyzed period (e.g. one year, one growing season, etc.) and should consider how those changes are addressed in the aggregation algorithm. The performance of aggregation should be also compared, in the future study, with other land cover classification methods that utilize multitemporal image data.

ACKNOWLEDGEMENT

This research study has been conducted as part of the European Space Agency founded Sentinel-2 Global Land Cover project (S2GLC) within the framework of Scientific Exploitation of Operational Missions (SEOM) S2-4Sci Land and Water - Study 3: Classification program.

REFERENCES

[1] Inglada, J. *et al.* Operational High Resolution Land Cover Map Production at the Country Scale Using Satellite Image Time Series. *Remote Sens.* **9,** (2017).

[2] Matton, N. *et al.* An automated method for annual cropland mapping along the season for various globally-distributed agrosystems using high spatial and temporal resolution time series. *Remote Sens.* **7,** 13208–13232 (2015).

[3] Valero, S. *et al.* Production of a dynamic cropland mask by processing remote sensing image series at high temporal and spatial resolutions. *Remote Sens.* **8,** (2016).

[4] Löw, F., Conrad, C. & Michel, U. Decision fusion and non-parametric classifiers for land use mapping using multi-temporal RapidEye data. *ISPRS J. Photogramm. Remote Sens.* **108,** 191–204 (2015).

[5] Vieira, M. A. *et al.* Object Based Image Analysis and Data Mining applied to a remotely sensed Landsat time-series to map sugarcane over large areas. *Remote Sens. Environ.* **123,** 553–562 (2012).

[6] Waldner, F., Canto, G. S. & Defourny, P. Automated annual cropland mapping using knowledge-based temporal features. *ISPRS J. Photogramm. Remote Sens.* **110,** 1–13 (2015).

[7] S2GLC. Sentinel-2 Global Land Cover Project. **2017,** (2016).

[8] Gascon, F. *et al.* Copernicus Sentinel-2A calibration and products validation status. *Remote Sens.* **9,** (2017).

[9] Copernicus. Copernicus - The European Earth Observation Programme. **2017,** (2017).

[10] Breiman, L. Random Forests. *Mach. Learn.* **45,** 5–32 (2001).

[11] Belgiu, M. & Drăguţ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **114,** 24–31 (2016).

[12] Loosvelt, L., Peters, J., Skriver, H., De Baets, B. & Verhoest, N. E. C. Impact of Reducing Polarimetric SAR Input on the Uncertainty of Crop Classifications Based on the Random Forests Algorithm. *IEEE Trans. Geosci. Remote Sens.* **50,** 4185–4200 (2012).

[13] Polikar, R. Ensemble based systems in decision making. *IEEE Circuits Syst. Mag.* **6,** (2006).

[14] Benediktsson, J. A., Sveinsson, J. R. & Swain, P. H. Hybrid consensus theoretic classification. *IEEE Trans. Geosci. Remote Sens.* **35,** 833–843 (1997).

[15] Clinton, N., Yu, L. & Gong, P. Geographic stacking: Decision fusion to increase global land cover map accuracy. *ISPRS J. Photogramm. Remote Sens.* **103,** 57–65 (2015).

[16] Du, P. *et al.* Multiple classifier system for remote sensing image classification: A review. *Sensors* **12,** 4764–4792 (2012).

[17] Wolpert, D. H. Stacked generalization. *Neural Networks* **5,** 241–259 (1992).

[18] Benediktsson, J. A. & Sveinsson, J. R. Multisource remote sensing data classification based on consensus and pruning. *IEEE Trans. Geosci. Remote Sens.* **41,** 932–936 (2003).

[19] Joshi, N. *et al.* A review of the application of optical and radar remote sensing data fusion to land use mapping and monitoring. *Remote Sens.* **8,** (2016).

[20] Solberg, A. H. S. Data Fusion for Remote-Sensing Applications. in *Signal and Image Processing for Remote Sensing* (ed. Chen, C. H.) 249–271 (CRC Press, 2007).

[21] Kuncheva, L. I. & Rodríguez, J. J. A weighted voting framework for classifiers ensembles. *Knowl. Inf. Syst.* **38,** 259–275 (2014).

[22] Congalton, R. G. & Green, K. *Assessing the accuracy of remotely sensed data: principles and practices*. *Mapping Science Series* (CRC Press, 2009).

[23] Kukawska, E. *et al.* Multitemporal Sentinel-2 data – remarks and observations. *2017 9th International Workshop on the Analysis of Multitemporal Remote Sensing Images (Multi-Temp)* (2017).