Assessment of Cloud Cover in Sentinel-2 Data using Random Forest Classifier

1st Petteri Nevavuori
Mtech Digital Solutions Oy
Vantaa, Finland
petteri.nevavuori@mtech.fi

2nd Tarmo Lipping
Tampere University
Tampere, Finland
tarmo.lipping@tuni.fi

3rd Nathaniel Narra
Tampere University
Tampere, Finland
nathaniel.narra@tuni.fi

4th Petri Linna
Tampere University
Tampere, Finland
petri.linna@tuni.fi

I. Introduction

Data from the Sentinel satellites are intensively used for various applications such as land use and vegetation mapping or crop monitoring, for example. Depending on climate conditions in the region of interest, the main obstacle in using the data for practical monitoring purposes may be cloud coverage. This is especially restricting if the data should be acquired from a narrow time window corresponding, for example, to a certain growth phase of crops. The problem could be alleviated by more accurate and higher resolution cloud coverage assessment compared to that available by the product of the Sentinel data.

Currently the cloud mask of the Sentinel data is available in the form of the Level 1C product containing vector layers of dense and cirrus clouds. Also, the percentage of cloudy pixels (dense and cirrus) in the mask are provided. The Level 2A product further processes the Level 1C data to obtain the Scene Classification layer with cloud and cirrus probability values at 60 m spatial resolution. Calouzzi et.al. [1] assessed these products concluding that caution has to be taken when using the provided cloud masks and improved cloud detection algorithms are welcome. Recently, Baetens et.al. [2] compared three cloud mask calculation algorithms: MAJA (used in the Level 2A product), Sen2Cor (used by ESA) and FMask (used by USGS), using their Active Learning Cloud Detection (ALCD) method for producing reference cloud masks. Classification accuracy of about 90 % was obtained by MAJA and FMask while SenCor gave 84 % accuracy.

In this paper we train the random forest classifier to assess cloud cover in Sentinel-2 data. Our primary usage of the data is crop monitoring and yield prediction for decision support for farmers. Therefore, the classifier is trained using data acquired from crop fields by UAVs: as UAVs fly below the clouds and the data they produce is not affected by cloud cover (if properly corrected for changes in irradiance), the difference between the UAV and Sentinel data can be used as ground truth for cloud cover.

II. Data

A. Drone Images

For cloudless multispectral ground truth data, ten crop fields were selected for imaging in the vicinity of Pori, Finland (61°29'N, 21°48'E) and were imaged as a part of the MIKA DATA project [3], [4]. The total area of the selected fields was approximately 93 ha. Half of the fields had wheat (*Zebra/Mistral*), three had barley (*Harbringer/RGT Planet*) and two remaining had oats (*Ringsaker*) as the cultivated crop. The fields were imaged during the growing season for years 2018 and 2019 from the time of sowing to the time of harvest. All fields were imaged weekly. Due to varying weather conditions and the proximity of an airport, the temporal allocation of imaging flights to within a fixed daily time range was not possible. The images were thus taken during day time.

The fields were imaged with two distinct drones, using 3DR Solo for the year 2018 and Parrot Disco-Pro AG for 2019. The drones were equipped with similar Parrot Sequioa multispectral cameras. Distinct images were collated for each field to build a complete image of a field using the Pix4D software. During the process of building the image mosaics, the band data were also automatically normalized in terms of radiance utilizing the information provided by the multispectral camera's irradiance sensor. Using the red and near-infrared (NIR) channels, the normalized difference vegetation index (NDVI) was then calculated from each field's multiband mosaic. To use the drone data in conjunction with the Sentinel-2 data, the collated drone images were downsampled to match the highest resolution available in Sentinel-2 images, 10 m/px. The downsampling was done using `cubicspline` interpolation algorithm in the `gdalwarp` utility. Lastly, the images for each field were cut to proper shape with field block border data provided by Ruokavirasto (*Finnish Food Authority*) [5]. This resulted in a total of 288 distinct crop field images. The field-wise sizes, crop varieties, yearly image counts and average valid pixel counts per image are given in Table I

The use of NDVI images calculated from drone data is discussed in Sec. II-C. Next we will discuss the acquisition and processing of the Sentinel-2 satellite data.

B. Sentinel-2 Data

Sentinel-2 satellite images were selected as the source data for the study. The data provided by the dual satellite system are widely used in agriculture and is freely available. The satellite images processed to the Sentinel product Level-2A [6] were downloaded from Copernicus Open Access Hub [7].

TABLE I
SIZES, CROPS, IMAGE COUNTS AND AVERAGE PIXEL COUNTS OF FIELDS SELECTED FOR DRONE IMAGING.

Field	Size, ha	Crop	Image Counts 2018	2019	Avg. Valid Px Per Image
1	11.08	Wheat	13	16	1065.5
2	8.24	Wheat	15	14	759.1
3	11.77	Wheat	13	16	1120.9
4	11.12	Wheat	15	16	1051.9
5	7.59	Wheat	15	16	705.2
6	7.61	Oats	12	15	739.8
7	7.24	Oats	13	15	681.9
8	7.77	Barley	13	15	1016.6
9	13.05	Barley	12	16	1251.3
10	7.95	Barley	12	16	715.5

The satellite data products were downloaded for the growing seasons of 2018 and 2019.

The satellite data were selected with no limits on the estimated cloud coverage. The goal was to be able to find week-matching pairs for the drone data. The data was used as the training data for which information about the cloudless ground truth was available via drone data. The gathered data spanned initially the growing seasons of years 2018 and 2019. Part of the downloaded data was omitted during the process of week-matching Sentinel-2 data to Drone data. The satellite image data were cut to shape using field block borders already utilized with the drone data to ultimately generate image pairs of drone and satellite data aligned both temporally and geographically for distinct fields.

C. Target Data

Supervised machine learning requires the existence of *a priori* labeled data, the ground truth. With the aim of estimating cloud coverage in Sentinel-2 data in the spatial scale of crop fields, NDVI images gathered with drones at the altitudes well below clouds are considered as cloudless ground truth. This consideration is in relation to satellites flying at atmospheric altitudes. Comparing absolute values across bands for two different sensors and imaging platforms has proven to be difficult, as the data would require scaling to an unkown global maximum for Sentinel-2. However, the use of NDVI alleviates this problem by providing normalized and thus comparable data between distinct imaging systems.

Target data needs thus to be generated using the week-aligned NDVI data from both sources, the drone and the Sentinel-2 systems. Each spatially and temporally aligned satellite and drone NDVI image pair is compared pixel by pixel to determine whether the images are similar on the level of distinct pixels. A pixel correponds to an area of 10×10 meters. The similarity for a single pixel-corresponding area is determined by

$$sim_{(s,d)} = \begin{cases} 1, & |s-d| \leq threshold \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where s and d are spatially and temporally aligned pixels for a field from the satellite and drone sources respectively. The mean absolute errors (MAEs) of all week-aligned image pairs are depicted in Fig. 1. The determination of the threshold is discussed next.

To determine a proper absolute NDVI difference threshold for labeling Sentinel-2 pixels either similar or dissimilar to the drone pixels (see Eq. 1), the two data sources were compared using the Student t-test. The test was applied over the pixels in the images to compare whether the NDVI values in the images were statistically similar or not. A total of 15 statistically similar ($p = 0.01$) week-aligned image pairs were found. It is to be noted though, that the number of image pairs having MAE in close proximity to the similarity threshold was higher than just 15 (see Fig. 1).

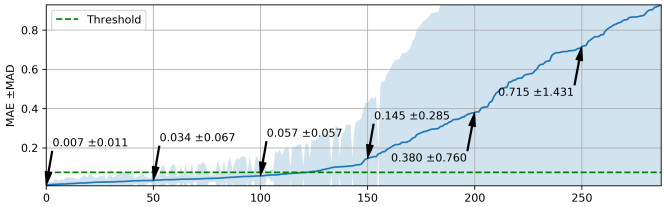

Fig. 1. The mean absolute errors (MAE) and mean absolute deviations (MAD) of week-aligned NDVI pairs in ascending order. The statistics are calculated over the pixels in the paired Sentinel-2 and drone NDVI images.

The statistically similar data (15 image pairs) were then used to empirically determine the proper threshold for classifying NDVI differences in terms of pixel-wise similarity. The tested thresholds were selected from the proximity of upper end of the MAE for the statistically similar data samples as shown in Table II. In more general terms, the task of determining the threshold for labeling is a task of balancing between (1) capturing as much similarities while (2) still excluding as many dissimilarities as possible. To elaborate, labeling every pixel in the statistically similar images as similar would require increasing the absolute NDVI threshold to levels possibly having some pixels incorrectly labeled as similar. The ratios of pixels labelled as similar for each similar image pair with different thresholds is given in Table III. In combination with visual evaluation, a threshold of 0.075 absolute NDVI difference was selected. A single image pair with the calculated similarity map is shown in Fig. 2.

TABLE II
NDVI DIFFERENCE METRICS FOR SIMILAR IMAGE PAIRS

Image pairs	15
Avg. Diff.	0.001 ±0.046
MAE	0.026 ±0.022
MSE	0.003 ±0.010
RMSE	0.046 ±0.092

TABLE III
SIMILARITY RATIOS WITH VARIOUS THRESHOLDS

Threshold	Similarity
0.025	89.13%
0.050	94.40%
0.075	96.14%
0.100	97.13%

D. Building the Modeling Data Sets

After the generation of field and week specific similarity label maps, the data required only minor preprocessing. As

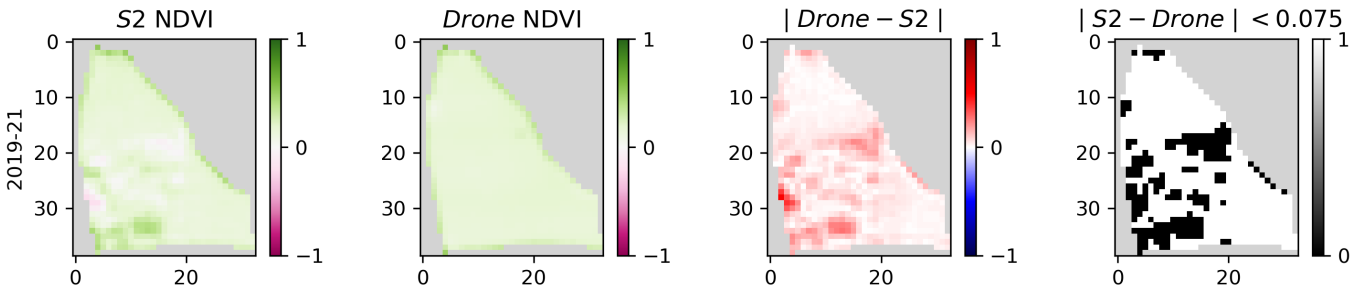

Fig. 2. A visualization of a single week-aligned Sentinel-2 and drone NDVI image pair with the absolute difference and the similarity map. The first two figures depict the NDVI maps from corresponding sources. The third figure shows the absolute difference between the aligned Sentinel-2 and drone NDVI values. The fourth figure shows the thresholded absolute difference, indicating areas where the NDVI images are similar enough.

the Sentinel-2 data products are delivered as separate files for distinct bands and layers, the satellite data were merged to construct multi-band images instead of multiple images of distinct bands. The following Sentinel-2 data were merged:

- *Sensor bands*: 1 to 8, 8A, 9, 11 and 12
- *Level-2A layers*: AOT, SCL, TCI, WVP and CLDPRB

The separately calculated NDVI data were also merged in conjunction with the alpha-channel generated during the processing of the data. As per machine learning best-pactices, the categorical values from the scene classification layer (SCL) needed to be separated to distinct binary raster layers according to the SCL classification labels, which is also known as transforming a multi-class representation to class-wise one-hot representation [8].

Thus, the final processed input data constituted 30 distinct layers of data for each pixel. The dataset was then created by extracting multi-band Sentinel-2 pixels as input samples and their spatially and temporally corresponding binary similarity label map pixels as target values. In other words, a single input sample was a $[1 \times 30]$ and its corresponding target sample a binary-valued $[1 \times 1]$ vector. A total of 381972 input-target samples (pixels) were extracted from the source data. The samples were then shuffled and split into training and test data sets with 190986 and 63661 samples, correspondingly. No scaling was applied due to the selected decision tree based model.

III. Model

Data based modeling with machine learning methods is in practice a tradeoff between model explainability and increased performance. While training an accurate model for classifying distinct Sentinel-2 pixels as similar or dissimilar to the cloud-less ground truth data from drones is the primary goal while the explainability was deemed as an important objective to pursue as well. This is why an ensemble model called Random Forest from the decision tree algorithm family was selected. The ensemble model is able to model non-linear relationships, work with unscaled data and provide easily understandable explanations of decisions' causes [9]. The model implementation was part of the Python's `scikit-learn` framework [10].

TABLE IV
The confusion matrix of similarity label predictions.

Pred/True	**0**	**1**
0	TP 23237	FP 2580
1	FN 1807	TN 36037

IV. Results

The model was allowed to train 500 sub-trees, varying the tree structure and features used for each tree, using the training data set only. The performance of the model was then evaluated with the hold out test data set. The confusion matrix of model predictions against true labels is shown in Table IV. The precision of the model is

$$PPV = \frac{TP}{TP + FP} = 0.900, \tag{2}$$

where PPV stands for positive prediction value. The model's true positive rate, i.e., recall, is then

$$TPR = \frac{TP}{TP + FN} = 0.923. \tag{3}$$

The F_1-score, a statistical test accuracy measure for binary classification analysis is then calculated using Eqs. 2 and 3 by

$$F_1 = 2 * \frac{PPV * TPR}{PPV + TPR} = 0.911. \tag{4}$$

Another interesting metric is the negative prediction value

$$NPV = \frac{TN}{TN + FN} = 0.952, \tag{5}$$

which shows the model's precision in predicting dissimilarities. In conjunction with test data set result analysis, the model was also evaluated with distinct images from the original source data.

Due to Sentinel-2 satellite data being sensitive to changes and disturbances in atmospehric conditions, the cloud estimation information from the scene classification layer (SCL) and cloud probablity mask (CLDPRB) calculated in the Level-2A processing of the Sentinel-2 data can not be taken as definitive

truth. They, however, form a proper baseline to which compare the trained model's performance against.

The model predictions are based on the similarities of Sentinel-2 and drone NDVI images, i.e. label 1 indicates predicted similarity. Taking a mean of a set of predicted values describes the mean predicted similarity for that set. The two cloudiness estimation masks in the Sentinel-2 data product are formulated differently.

As the name indicates, the CLDPRB mask contains pixel-wise probability values for the estimated degree of cloud coverage. The model-equivalent similarity measure would thus be

$$\text{CLDPRB}_{\text{SIM}} = 1 - \text{CLDPRB}, \tag{6}$$

where larger values imply increased degree of estimated similarity.

On the other hand, the SCL layer contains pixel-wise labels, with some labels indicating cloudiness (see [6]). To gain information about the SCL layer's model-equivalent similarity measure, the cloud-related label ratio

$$p_{cl} = \frac{\text{count}(\text{SCL}_{cl})}{\text{count}(\text{SCL})} \tag{7}$$

is first counted with the cl being a set of cloud-related class labels. The inverse

$$\text{SCL}_{\text{SIM}} = 1 - p_{cl} \tag{8}$$

can then be seen as the implied cloudless ratio for a set of samples. The comparison of sample-wise similarity estimations between the trained model and Sentinel-2 data products are given in Table V. The estimates are given both for when the true target value was 0 (satellite differed from drone) and when it was 1 (satellite similar to drone).

TABLE V
SIMILARITY ESTIMATES WITH HOLD OUT TEST DATA.

	$y = 0$			$y = 1$		
	Mean	Std	Median	Mean	Std	Median
Model	**0.067**	0.250	0.000	0.928	0.259	1.000
$\text{CLDPRB}_{\text{SIM}}$	0.446	0.454	0.260	**0.970**	0.138	1.000
SCL_{SIM}	0.282	0.450	0.000	0.949	0.220	1.000
Samples		38617			25044	

V. DISCUSSION AND CONCLUSIONS

Our study indicates that the Random Forest model outperforms the Sentinel-2 CLDPRB and SCL data layers in detecting cloudy areas ($y = 0$). For non-cloudy areas the detection accuracy was slightly higher for the Sentinel products (see Table V). Several issues should be considered, however, when comparing these results. Firstly, when training the Random Forest classifier, the thresholded absolute difference between the Sentinel-2 and drone data was used as the ground truth. While it can be argued that the main cause of this difference is cloudiness, there may also be other factors involved such as shadows or differences in irradiance. The satellite and drone imagery were not necessarily acquired during the same time of the day or same day of the week, although best time-matching pairs were looked for when selecting the data. In some cases a couple of days may cause significant changes in the crop development. Another limitation comes from using the NDVI data layers for ground truth assessment. While the NDVI index contains significant information for vegetation monitoring and is probably a good choice when assessing cloud cover in crop fields, its use reduces the generalizability of the results to other land cover types.

Despite the mentioned limitations, the developed method was found to improve the usability of Sentinel data in crop monitoring. By visual inspection it was observed that in many cases when the Sentinel-2 products indicated the whole crop field to be cloud-covered, there were still significant areas of almost clear skies. The proposed algorithm proved capable in detecting these areas with considerable accuracy.

REFERENCES

[1] R. Coluzzi, V. Imbrenda, L. Maria, and S. Tiziana, "A first assessment of the sentinel-2 level 1-c cloud mask product to support informed surface analyses," *Remote Sensing of Environment*, vol. 217, pp. 426–443, 09 2018.

[2] L. Baetens, C. Desjardins, and O. Hagolle, "Validation of copernicus sentinel-2 cloud masks obtained from maja, sen2cor, and fmask processors using reference cloud masks generated with a supervised active learning procedure," *Remote Sensing*, vol. 11, 02 2019.

[3] P. Nevavuori, N. Narra, and T. Lipping, "Crop yield prediction with deep convolutional neural networks," *Computers and Electronics in Agriculture*, vol. 163, no. June, p. 104859, 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0168169919306842

[4] N. Narra, P. Nevavuori, P. Linna, and T. Lipping, "A Data Driven Approach to Decision Support in Farming," in *Information Modelling and Knowledge Bases XXXI*, A. Dahanayake, J. Huiskonen, Y. Kiyoki, B. Thalheim, H. Jaakkola, and N. Yoshida, Eds. IOS Press, 2020, vol. 321, pp. 175 – 185.

[5] Ruokavirasto, "Peltolohkorekisteri." [Online]. Available: https://www.ruokavirasto.fi/tietoa-meista/avointieto/tiedonluovutukset/peltolohko-usb/

[6] ESA, "Level-2A Algorithm - Sentinel-2 MSI Technical Guide - Sentinel Online." [Online]. Available: https://earth.esa.int/web/sentinel/technical-guides/sentinel-2-msi/level-2a/algorithm

[7] ——, "Open Access Hub." [Online]. Available: https://scihub.copernicus.eu/

[8] D. Hong, N. Yokoya, N. Ge, J. Chanussot, and X. X. Zhu, "Learnable manifold alignment (LeMA): A semi-supervised cross-modality learning framework for land cover and land use classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 147, pp. 193–205, jan 2019.

[9] P. Flach, "Machine Learning: The Art and Science of Algorithms that Make Sense of Data," p. 409, 2012.

[10] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. Vanderplas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," 2013. [Online]. Available: http://arxiv.org/abs/1309.0238