MARRYING DEEP LEARNING AND DATA FUSION FOR ACCURATE SEMANTIC LABELING OF SENTINEL-2 IMAGES

*G. Fonteix[1], M. Swaine[2], M. Leras[1], Y. Tarabalka[1], S. Tripodi[1],
F. Trastour[1], A. Giraud[1], L. Laurore[1], J. Hyland[2]

[1]LuxCarta Technology, Mouans Sartoux, France – gfonteix@luxcarta.com
[2]LuxCarta South Africa, Cape Town, South Africa – mswaine@luxcarta.com

KEY WORDS: Deep learning, time-series, optical satellite images, semantic segmentation, U-net, confidence maps

ABSTRACT:

The understanding of the Earth through global land monitoring from satellite images paves the way towards many applications including flight simulations, urban management and telecommunications. The twin satellites from the Sentinel-2 mission developed by the European Space Agency (ESA) provide 13 spectral bands with a high observation frequency worldwide. In this paper, we present a novel multi-temporal approach for land-cover classification of Sentinel-2 images whereby a time-series of images is classified using fully convolutional network U-Net models and then coupled by a developed probabilistic algorithm. The proposed pipeline further includes an automatic quality control and correction step whereby an external source can be introduced in order to validate and correct the deep learning classification. The final step consists of adjusting the combined predictions to the cloud-free mosaic built from Sentinel-2 L2A images in order for the classification to more closely match the reference mosaic image.

1. INTRODUCTION

The ever-increasing amount of available optical satellite images has enabled the development of new techniques for global land-cover classification. Manual annotations are time-consuming and sometimes less accurate than automatic methods. Moreover, traditional semantic segmentation approaches, as well as methods based on machine/deep learning, usually present misclassified parts in the case where prediction is limited to a single image.

Single-date satellite images are traditionally used for land-use segmentation (Desclée *et al.*, 2006). Access to time-series images can improve recognition integrity and accuracy by using new algorithms and strategies as shown in several studies (Yan and Roy, 2014, Kamdem de Teyou *et al.*, 2020).

With a revisit time of 5 days, the freely-available images from high-temporal resolution satellites Sentinel-2A and Sentinel-2B, launched in 2015 and 2017 respectively, enable the production of a 10-meter cloud-free Earth mosaic and multi-temporal classification. In this work we address the problem of a pixel-wise classification, where each image pixel is assigned to a thematic class.

To produce a high-quality mosaic and accurate land-cover map in an automatized way, solutions to three major issues were focused on:

- A cloud-free mosaic from Sentinel-2 images must be automatically generated, with natural and aesthetically pleasing colors, while simultaneously preserving details of the landscape.
- High-quality classification maps need to be extracted from time-series images, which must closely fit the cloud-free mosaic. The goal is to limit manual correction (adding omissions and deleting over-detected pixels), which has been essential in most operational scenarios to provide compliant data.
- Finally, it is desirable to have an estimated measure of the extraction accuracy by zone, with the purpose to speed up quality control procedures.

To overcome the above mentioned issues, we propose an automated chain that combines multi-date predictions in a probabilistic way, and further refines the obtained classification map to fit the mosaic images. The major contributions of this work lie in combining several images and sources in order to assess and improve semantic labeling results on the enhanced mosaic. The combination of predictions from different dates increases the completeness. The resulting classification map is then automatically assessed and corrected using an external data source as well as an unsupervised classification of the mosaic image. The aim is to obtain an up-to-date land use map together with its confidence score.

2. PROPOSED METHOD

Single-date classifications can yield imperfect results, thus this paper proposes a multi-temporal approach. A limitation however is that the resultant classification must correspond to a single reference image. Thus it is essential to automatically readjust the prediction according to this reference image. To reach this goal, the proposed pipeline consists of four main steps:

1) Generation of a cloud-free mosaic from Sentinel-2 L2A images.
2) Multi-date deep learning-based classification from Sentinel-2 L1C images.
3) Combining deep-learning based classification with an external data source (e.g., outdated prediction, OpenStreetMap, etc.).
4) Automatic correction of natural environments to fit the cloud-free mosaic.

* Corresponding author

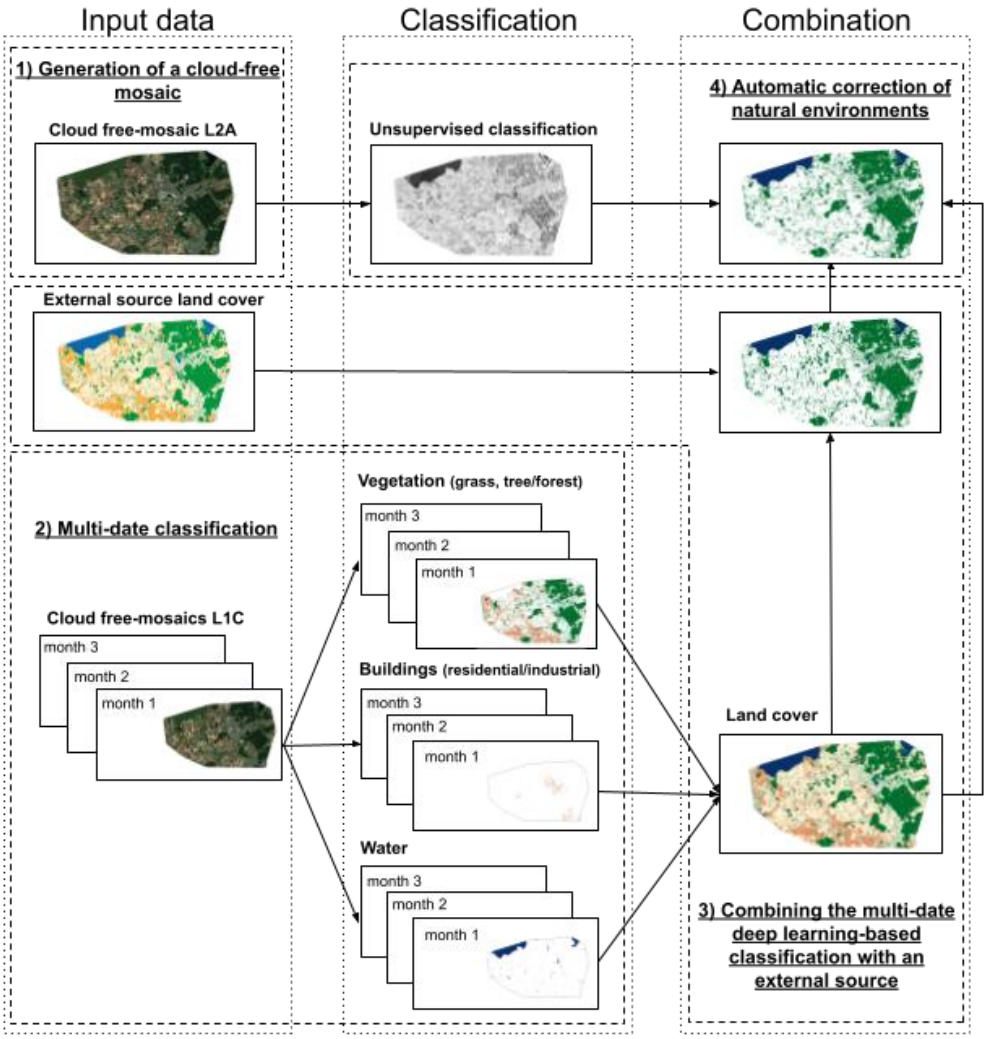


Figure 1: The proposed workflow to improve land-cover segmentation.

Each step will be explained in detail and illustrated with examples from different parts of the world. The proposed pipeline for reaching an optimal land-cover extraction in a reduced time is illustrated in Figure 1.

2.1 Generation of a cloud-free mosaic from Sentinel-2 L2A images

Recently, LuxCarta has completed a 2019/2020 vintage global 10-meter mosaic, named BrightEarth[1]. This followed on from the work (Swaine *et al.*, 2020) where an operational pipeline for a global 10-m resolution mosaic was produced. A number of enhancements have been added to this chain, namely; (a) the used source imagery processing level and (b) the image normalization methodology.

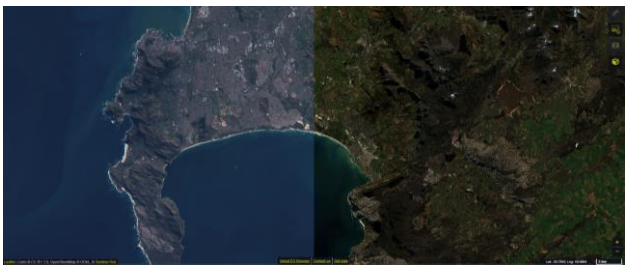
2.1.1 Source Imagery Processing Level: A major improvement on the existing pipeline has been the integration of Sentinel-2 L2A processed data as the source input data as opposed to Sentinel-2 L1C processed data. Sentinel-2 data has five processing levels, namely; level-0, level-1A, level-1B, level-1C and level-2A, with only the last two processing levels available to users. Table 1 provides the description for each processing level as set out by Sentinel's technical guides.

Processing level	Description
L0	Compressed raw data.
L1A	Uncompressed raw data with spectral bands coarsely co-registered and ancillary data appended.
L1B	Radiometrically corrected radiance data. The physical geometric model is refined using available ground control points and appended to the product, but not applied.
L1C	Orthorectified Top-Of-Atmosphere (TOA) reflectance, with sub-pixel multispectral registration.
L2A	Orthorectified Bottom-Of-Atmosphere (BOA) reflectance, with sub-pixel multispectral registration.

Table 1: Sentinel-2 Processing levels description.

Since December 2018, ESA has provided L2A processed data globally (ESA, 2018). We have taken advantage of this product as it saves thousands of processing hours to manually process L1C data to L2A using Sen2Cor processor (see Figure 2 for comparison between L1C and L2A processed data)

*[1] https://www.luxcarta.com/products/brightearth/



(1) (2)

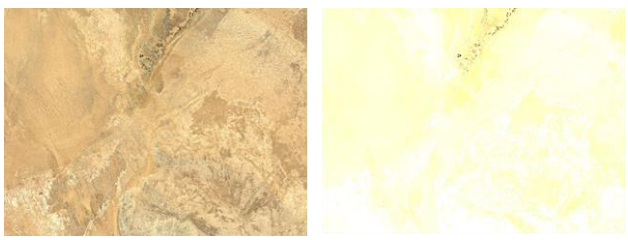
Figure 2: (1) L1C and (2) L2A processed data as viewed in EO Browser.

2.1.2 Image normalization: The 8-bit True Color Image (TCI) provided by ESA is not adequate to create a global mosaic. From previous experience, the TCI lacks a balance in detail retention and brightness. Thus, LuxCarta has produced a scaling process which retains detail in bright areas where detail would traditionally have been blown out, yet still maintains an adequate brightness throughout the image. Due to different reflectance values across the Earth, Sentinel grid tiles were split into three scaling zones representing differing reflectance categories, namely bright desert, glacial and mixed. Bright desert areas mainly consists of North Africa where very high reflectance values are observed. Similarly, glacial areas include grids where permanent ice or snow occurs.

Each scaling zone utilizes a different set of scaling parameters which is optimized for each zone in order to retain detail in the downscaling to 8bit depth process. Figures 3 and 4 illustrate the difference in the LuxCarta-based and ESA-based downscaling for glacial and desert zones.

(1) (2)

Figure 3: (1) LuxCarta downscaled 8bit, (2) ESA downscaled 8bit.



(1) (2)

Figure 4: (1) LuxCarta downscaled 8bit, (2) ESA downscaled 8bit.

2.2 Multi-date classification from Sentinel-2 L1C images

With a revisit time of five days and free online data access, the satellites from the Sentinel-2 constellation provide multi-spectral images, making multi-date predictions possible.

Obtaining a convincing result from a single Sentinel-2 image is a challenge. For example, due to cloud shadows or difference of reflectance between images, results might vary from date to date. To minimize omissions and wrong detections, we have opted for the use of multiple images acquired on different dates.

As mentioned in Section 2.1, two products can be used as input for classification. The Level-1C (L1C) introduced in 2015 and the Level-2A (L2A) introduced at the end of 2018. As the L1C images have a larger catalog and better retain the original physical properties of the image, this processing level has been chosen for the multi-date classification.

A cloud-free image can be obtained by combining several parts of Sentinel-2 L1C images (Swaine *et al.*, 2020). Thus, three recent cloud-free mosaics are produced for semantic labeling by U-Net convolutional neural network models. The U-Net architecture (Ronneberger *et al*., 2015) has shown its high performances for pixel-wise classification in various remote sensing studies (Huang *et al*., 2018, Le Saux *et al.*, 2019). The applied network architecture is adopted from (Tasar *et al.*, 2018) and illustrated in Figure. 5.

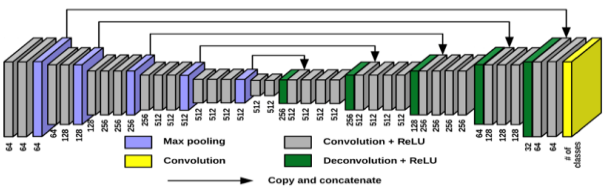


Figure 5: U-net model used for landcover classification.

Three distinct models have been trained to separately classify vegetation classes (grass, tree/forest), buildings (residential and industrial/commercial) and water.

The near-infrared, red and green bands were used as an input for the vegetation and water models while red, green and blue spectral channels were exploited for building classes. A large amount of accurately annotated ground-truth data (property of the LuxCarta company[2]) has yielded efficient models that succeed in generalizing all around the world.
The proposed method consists of applying all of the trained models on the three mosaics (on different periods of time) and coupling the results using a probabilistic algorithm. Following this approach, a resulting 6-class land-cover in a raster format is generated by inferring a thematic label for every pixel: water, grass, forest, barren, industrial building or residential building.

Even though the combined results are more accurate than a single-date prediction, at the end of this automatic classification step no information is available about reliability of each classified pixel truly belonging to the assigned class. Our

*[2] https://www.luxcarta.com

probabilistic model purposely tends to over classify forest and water classes (denoted as natural environment classes in this work). This choice allows us to boost up the number of well-detected pixels for these classes knowing that the next step will refine the contours to gain accuracy.

In the following section, we propose a procedure for automatic assessment and correction of classification maps. For the purpose of illustration in this work, we are going to focus on the automatic correction of the forest class. However, the method is applicable to other classes as well.

2.3 Combining deep learning-based classification with an external data source

The main idea of the proposed method consists in using another available data source, even if this data is not accurate (outdated, lower resolution or contains artefacts). The assumption is that if a multi-date classification on the recent imagery agrees with an independent external data source, the reliability of such prediction is very high.

Providing geodata for thirty years, LuxCarta (www.luxcarta.com) has collected a very large quantity of data all around the world. From coarse to very-high resolution satellite imagery, such as Sentinel-2 or Pléiades, land-cover maps have been created (we call them OTS - On The Shelf).

To generate high-quality geodata for our customers, manual quality control and correction allow us to assert that our OTS maps contain few false positive predictions, even though the data is *not up to date*. These data can be used as an external source.

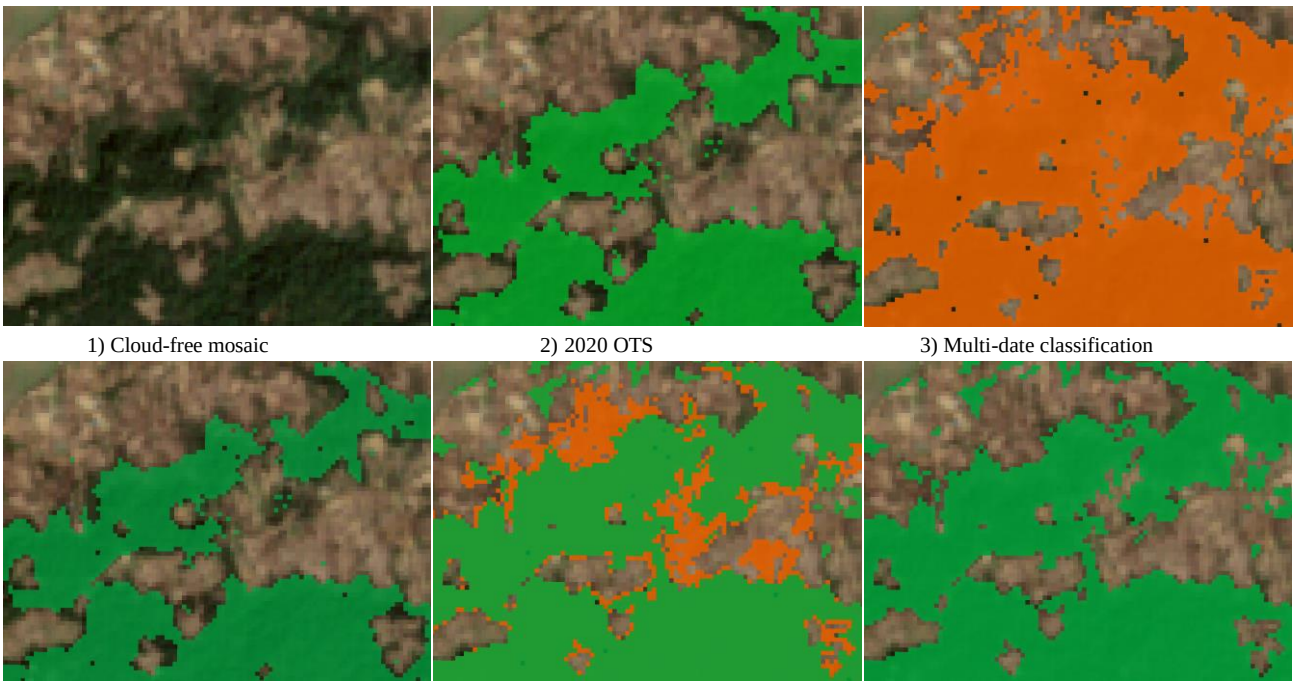
There are many other possibilities of free external data sources. For example, open databases, such as Corine Land cover (Bossard *et al*., 2000) or OpenStreetMap (OSM) (Johnson and Iizuka, 2016) can be exploited in the same way as an input.

The method for an automatic enhancement and assessment of the classification maps uses three input data:

- Cloud-free mosaiced image (ref. Section 2.1).
- Deep learning-based multi-temporal classification (ref. Section 2.2).
- External classification map (e.g. OSM, OTS, etc.).

An intersection between a multi-date classification map and an external classification map yields an under detected, but rather confident area (see Figure 6(4)). In our automatic pipeline, we assign a reliability code of 1 to each pixel in the intersection area, which denotes a very reliable classification.

To proceed to the next step, the multiband mosaic image must be converted to a single raster band using an unsupervised classification algorithm (e.g., k-means was chosen in our study (Jain., 2008), with 255 clusters). With this step, a unique value is associated to each image pixel.



1) Cloud-free mosaic | 2) 2020 OTS | 3) Multi-date classification

4) Intersection | 5) Multi-date (orange) cover by the final result | 6) Final result fitting the image

Figure 6: Results of the colour propagation (Guatamala).

2.4 Automatic correction of the classification map to fit the cloud-free mosaic

The last step consists of an automatic enhancement of the classification map. Our method is inspired by the hysteresis thresholding algorithm (Sornam *et al.*, 2016). The novelty consists in combining deep learning-based classification results approved by an additional external data source, with unsupervised classification results, using the principles similar to hysteresis thresholding. The output of this step is a more precise classification map, together with the associated reliability codes for every pixel, as well as the estimated confidence score for each image.

The following methodology can be applied to each class independently.

We seek to determine if the pixels of the multi-date prediction not contained in the high confidence areas (ref. Section 2.3), are part of the class and with which confidence score.

The main idea to solve this problem is to collect the pixel values of the unsupervised classification result (i.e. indices of clusters) which are significant for the class. For this purpose, we first polygonise high confidence areas. Then, for each polygon *Pi* we collect the cluster indices it contains, we call this set of pixel values *SPi*. Only cluster indices with an acceptable number of occurrences are retained to represent the class of interest, avoiding spreading false information. For the same reason, we do not include the polygons composed of less than 10 pixels.

We defined two cases:

1. For the pixels of the multi-date prediction, which are connected to a polygon *Pi*: if its cluster index in the unsupervised classification is contained in *SPi*, we add this pixel to the final classification map with the confidence score = 1; otherwise this pixel is rejected from the classification result.
2. For the pixels of the multi-date prediction, which are not connected to any *Pi*: if its cluster index in the unsupervised classification is contained at least in one *SPi* for the polygons *Pi* present within a radius of 500 pixels, we add this pixel to the confidence map with a confident score = 2 (medium confidence); otherwise with a confident score = 3 (low confidence).

Finally, a confidence map is obtained where every pixel is assigned a low to high reliability index, an example of this is shown in Figure 7. For the final evaluation, a confidence score is calculated by dividing the number of pixels considered as well-classified (code 1) by the set of pixels of the final layer (code 1 + code 2 + code 3) on the work area.

(1) (2)

Figure 7: (1) Mosaic image over the city of Ballarat, Australia, and (2) the corresponding confidence map: the green corresponds to code 1, the orange to code 2 and the red to code 3.

This classification accuracy assessment is an efficient tool to quickly validate the quality of the land-cover map and highlight complicated areas where the predictions are not accurate.

3. EXPERIMENTS

Two evaluation strategies are used to better understand the efficiency of the proposed pipeline. The first strategy consists of observing the time required for manual correction on the same zone based on the single image prediction, the multi-date extraction and the result from this methodology. The second approach use very accurate ground truth to compare pixel-wise classification results.

Throughout the paper, the automatic chain has been illustrated for the forest layer on Sentinel-2 images over samples in Guatemala (confidence score of 98%), Denmark (confidence score of 89%) and Australia (confidence score of 92%) at the spatial resolution of 10m/pixel. As the presented methodology is currently operational to produce land-cover for our customers, it has already been tested on five other zones. We have chosen data over countries (Austria, France, Vietnam, USA and Brazil) with different types of forest (coniferous and broadleaved), the obtained confidence scores are between 79 and 98%. In all study areas and based on the feedbacks of the quality control team, the land-cover map was significantly improved by applying the proposed methodology. The confidence score allows quick validation of the accuracy. It provides an unambiguous indication of the prediction reliability.

Further validation has been carried out over Austria, which has a confidence score of 91%. An accuracy assessment was made, in which a validation team has done manual correction on the prediction from a single date, the multi-date prediction and the final result obtained following the proposed chain. The results are indisputable, to obtain an equivalent result, it was necessary to spend twice as long as manual corrections on the prediction carried out in a single date compared to the extraction from three dates. Better yet, the final result required half the effort for the validation team when compared to multi-date prediction. Moreover, the quality control team has determined that some

manual corrections were less accurate than the predictions at the end of the proposed processing chain.

Furthermore, an automatic evaluation strategy was applied to the single-date prediction, the multi-date extraction and the result after applying the developed framework. Corine Landcover is the classification map used as an external source in this case. A very precise, manually labeled, ground truth dataset (at a much higher resolution, 50 cm/pixel) was used to outline forest of 20 km^2 over the cities of Mailly and Nozay in France allowed the comparison of the results obtained pixel by pixel.

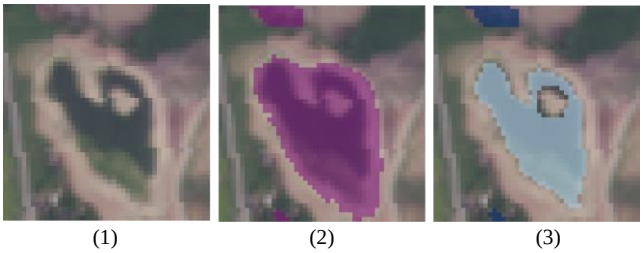
For the evaluation, the most common metrics for semantic segmentation were used. The Intersection over Union (IoU), also known as Jaccard index (Rahman and Wang, 2016) and used during the deep learning training as a loss function, compares the overlap between two shapes. It measures the number of pixels in common divided by the total number of pixels present in both predictions. If the prediction corresponds totally to the ground truth, the IoU = 1 and it decreases with an increasing difference.

The second metric is the F1 score also called F-score (Tatbul *et al.*, 2018), it combines the complementary effects of recall and precision. The recall indicates the proportion of successfully detected objects in the image, it is a measure of completeness and the precision describes the accuracy by calculating the fraction of all the detected pixels that are real forest. As it takes false positives and false negatives into account, it is an efficient measure for evaluating data.

The metrics obtained in Table 2, in line with the visual results illustrated in Figures 8 and 9, show that the multi-date prediction greatly improves the quality of the forest detection. In addition, an even higher score is attributed to the method applied to improve the contours of the prediction and remove false positives through the unsupervised classification (ref. Section 2.4). This is mainly due to an increase in precision as the over-detection is broadly reduced.

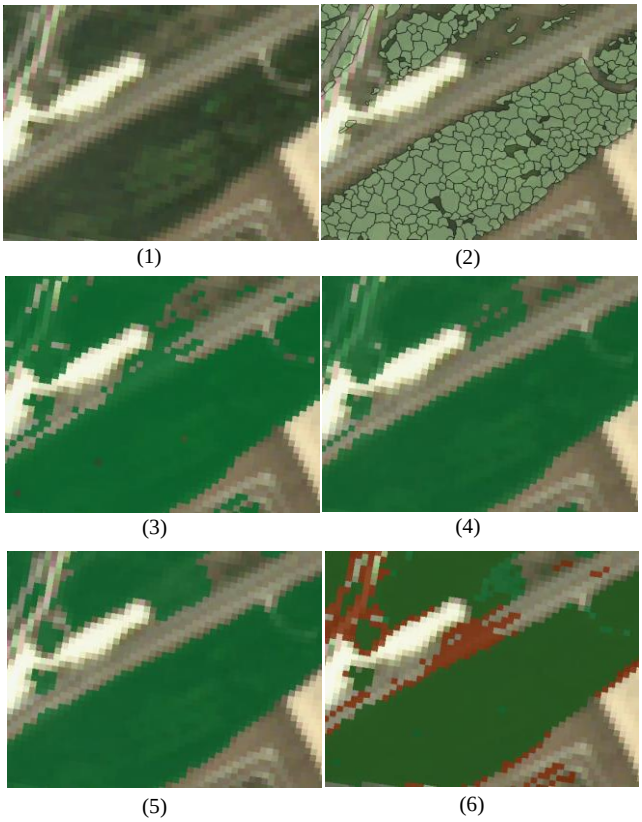
	Single-date	**Multi-date**	**Final layer**
	Mailly – France		
IoU	0.72	0.74	**0.76**
F1 score	0.84	0.85	**0.86**
	Nozay – France		
IoU	0.65	0.70	**0.72**
F1 score	0.79	0.82	**0.83**

Table 2: Experimental assessment on the cities of Mailly and Nozay



(1) (2) (3)

Figure 8: (1) Mosaic image over Denmark with the corresponding (2) multi-date water body extraction in purple and (3) the final result fitting the image in light blue.



(1) (2) (3) (4) (5) (6)

Figure 9: (1) Sentinel-2 L2A cloud-free mosaic over Mailly, France, and (2) corresponding ground truth, (3) prediction from a single image, (4) multi-date extraction, (5) final result and (6) comparison between (3: orange) and (5: green).

The latest experiments concern the water bodies, for this class the same pipeline can be applied. We have made only a few minor changes to adapt the algorithm to the water class, notably to keep the rivers continuity. The obtained results on several tests sets are very promising, significantly outperforming the state-of-the-art deep learning-based semantic labeling.

4. CONCLUSION

In this paper, we have proposed an operational pipeline to enhance land-cover maps using a multi-date and multi-source approach. This has been made possible through the use of time-series images provided by the Sentinel-2 constellation that continuously acquires data.

As manual annotations of multiple images over large areas are time-consuming and inefficient, new strategies were required to reach high-quality products. The emergence of new techniques based on deep learning opens doors for automation of Earth observation analysis.

By exploiting these methods, LuxCarta is now more than ever able to provide a high-quality mosaic derived from Sentinel-2 L2A images and the corresponding 6-class land use map with a high level of detail, an example of which can be seen in Figure 10.

(1)

(2)

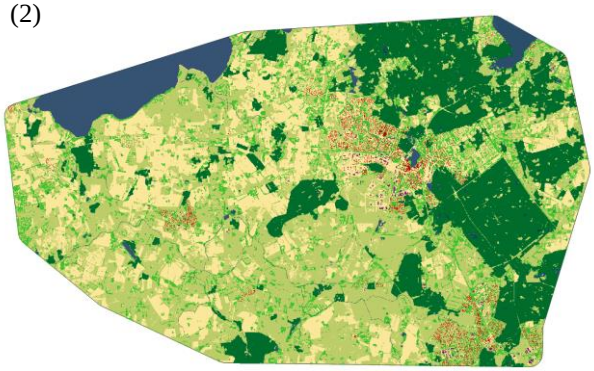


Figure 10: (1) Sentinel-2 L2A cloud-free mosaic over Hillerod, Denmark, and (2) corresponding 6-class land-cover map: water (blue), grass (light green), tree/forest (dark green), residential building (red), industrial/commercial building (orange) and barren (cream).

The multi-model deep learning approach trained on a large set of data coupled with the use of multiple well-chosen temporal images have proved to be an efficient way to improve the classification results. Additionally, the confidence index as a measure of agreement between several sources is now available for the quality control team to decrease the time required for the manual corrections.

REFERENCES

Bossard, M., Feranec, J., Otahel, J., 2000. CORINE land cover technical guide: Addendum 2000. ESA, 2018. Level-2A Algorithm Overview. https://sentinel.esa.int/web/sentinel/technical-guides/sentinel-2-msi/level-2a/algorithm

Desclée, B., Bogaert, P., Defourny, P., 2006. Forest change detection by statistical object-based method. *Remote sensing of environment*, *102*(1-2), 1-11.

Huang, B., Lu, K., Audeberr, N., Khalel, A., Tarabalka, Y., Malof, J. El-Saban, M., 2018. Large-scale semantic classification: outcome of the first year of inria aerial image labeling benchmark. *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium* (pp. 6947-6950). IEEE.

Jain, A. K., 2008. Data clustering: 50 years beyond k-means. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 3-4). Springer, Berlin, Heidelberg.

Johnson, B. A., Iizuka, K., 2016. Integrating OpenStreetMap crowdsourced data and Landsat time-series imagery for rapid land use/land cover (LULC) mapping: Case study of the Laguna de Bay area of the Philippines. *Applied Geography*, *67*, 140-149.

Le Saux, B., Yokoya, N., Hänsch, R., Brown, M., 2019. 2019 IEEE GRSS data fusion contest: Large-scale semantic 3d reconstruction. *IEEE Geoscience and Remote Sensing Magazine (GRSM)*, *7*(4), 33-36.

Rahman, M. A., Wang, Y., 2016: Optimizing intersection-over-union in deep neural networks for image segmentation. *International symposium on visual computing* (pp. 234-244). Springer, Cham.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham.

Sornam, M., Kavitha, M. S., Nivetha, M., 2016. Hysteresis thresholding based edge detectors for inscriptional image enhancement. In *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)* (pp. 1-4). IEEE.

Swaine, M., Smit, C., Tripodi, S., Fonteix, G., Tarabalka, Y., Laurore, L., Hyland, J., 2020. Operational Pipeline for a Global Cloud-Free Mosaic and Classification of SENTINEL-2 Images. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *43*, 195-200.

Tasar, O., Tarabalka, Y., Alliez, P., 2019. Incremental learning for semantic segmentation of large-scale remote sensing data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *12*(9), 3524-3537.

Tatbul, N., Lee, T. J., Zdonik, S., Alam, M., Gottschlich, J., 2018. Precision and recall for time series. *arXiv preprint arXiv:1803.03639*.

Kamdem de Teyou, G., Tarabalka, Y., Manighetti, I., Almar, R., Tripod, S., 2020. Deep Neural Networks for automatic extraction of features in time series satellite images. *arXiv preprint arXiv:2008.08432*.

Yan, L., Roy, D. P., 2014. Automated crop field extraction from multi-temporal Web Enabled Landsat Data. *Remote Sensing of Environment*, *144*, 42-64.