Unsupervised Pretrain, Supervised Fine-Tune and Knowledge Distillation in the domain of Artificial Intelligence for Earth Observation applied to Developing Countries

Gianfausto Bottini[1]

[1]Geospatial Data Scientist at Food and Agriculture Organization of United Nations (FAO)

Abstract: The use of Deep Neural Networks to extract patterns and features from Satellite imagery is nowadays widely used. Neural Networks exploit their multitude of parameters and a ground truth to classify images with one or more labels or to segment objects within the image itself. The Ground Truth is a stack of hundreds or thousands of labeled images and it's a key object to training a neural network; a labeled Dataset is paramount to train a Supervised model but it is not always available. Here is presented a state-of-the-art approach able to dribble the scarcity of in-situ data in difficult areas, such as Yemen, Syria, and Lesotho to eventually compute a coarse estimation of a land cover map at zero cost and at any administrative level (i.e., regions, sub-regions, districts, sub- districts, etc.). The procedure implemented in this paper is composed of two cascading Neural Networks. In the first step an Unsupervised model (UM), from the family of Contrastive Learning techniques, is trained to learn the general features of an unlabeled dataset by teaching the model which data points are similar or not. The hereby dataset is composed

of 50x50 pixels tiles of Sentinel2 acquisitions (High-Resolution images released for free by ESA). This action is indispensable to replace the massive training over thousands of labeled images. In the second step, weights trained in (UM) are passed to a second model leveraging the Transfer Learning technique. The second model is a Supervised model (SM) trained over a balanced dataset of a few hundred hand-labeled tiles divided into 8 categories: Tree cover, Shrubland, Grassland, Cropland, Built-up, Bare/sparse Vegetation, Permanent water bodies. Once the model (SM) is able to generalize it will be possible to harvest a good approximation of the Land Cover Map with no need for expensive in-situ data. The study, eventually, provides a benchmark with the most known pixel-wise models for land cover, ESA WorldCover (*1*) and Google Dynamic World (*2*).

keywords: Artificial Intelligence, Machine Learning, Deep Learning, Land Cover, Earth Observation, Remote Sensing

1 Introduction

In recent years, tremendous improvements have been made to the field of remote sensing, creating opportunities for new modeling capabilities to be applied to solve image classification tasks. Modeling techniques have been extended from traditional Machine Learning to, more recently, Deep Learning techniques. While conventional Machine Learning approaches require extensive knowledge of the input data and are often limited in their capability to distinguish essential features from each other, Deep Learning models use multiple levels of abstractions that capture more salient features from the data (*3*). As a result, these latter models have shown significant performance improvements in image classification tasks as laid out in several review papers (Dargan et al. (*4*), Hoeser and Kuenzer (*5*), Ball et al. (*6*), Ma et al. (*7*)). In particular, when

applied to land cover classification tasks, Deep Learning models have shown a strong ability to extract high-level spatial information rather than the less informative low-level features identified by traditional Machine Learning techniques (Ma et al. (*7*)). However, a main disadvantage of these models is that they require a considerable amount of labeled input data, which is often unavailable or costly to obtain.

This paper presents a new method to compute the land cover of a given country, which, by analyzing snapshots from a Satellite Acquisition from a single period in time, solves the need for large quantities of in-situ data. Depending on the size of the country and on the infrastructure used, the model itself can harvest the results in a few hours, meaning that it is possible to have a monthly trend for each class. With respect to other more famous algorithms for land cover, this method is very cheap, fast, flexible, and less precise but comparable in terms of accuracy.

This paper is organized as follows:

- the Section 2 is about the novelty of the model itself and the novelty of the proposed methodology applied to this exercise.
- in Section 3 it is presented an overview of the use of Deep Learning applied to land cover estimations and the recent advances that have been made in these fields.
- the Section 4 introduces the methodology employed in this project and the data used. Eventually, it presents the results of the analysis comparing the proposed method with those developed by the FAO-OCS group, the European Space Agency (ESA), and Google Dynamic World.
- in Section 5 it is discussed the potential and the limitations of using the designed model, where and when it best performs or underperforms with respect to other land cover methods.

2 Novelty of the proposed methodology

The SimCLR approach stands out for its simplicity, effectiveness, and the ability to learn powerful representations from unlabeled data, which has implications for a wide range of computer vision and machine learning applications.

The novelty brought by this project lies in applying this state-of-the-art approach in the filed of semi-supervised learning in an ambitious field as the Remote Sensing one with composite data such as the High-Resolution acquisitions.

The operation of this exercise must be seen as a pure Research and Development project with no will of replacing any existing land-cover method benchmarked in this article, but rather to show the potentiality of Semi-supervised Learning applied to Land-Cover classification starting with the encouraging results shown in the next paragraphs.

3 Literature Review

3.1 Land-Cover classification

Land Cover is a fundamental variable for both environmental and human purposes. This visual data is cardinal to grasp an area's agricultural potential, manage its natural resources, and estimate its inhabitants' living conditions (Garcia-Mora, Mas, and Hinklet (*8*)). Therefore, assessing these data is imperative to formulate proper policy interventions in areas with little information.

The most widely implemented approach to estimate land cover is remote sensing analysis. Remote sensing techniques analyze a physical area or phenomenon by collecting data obtained by measuring emitted and reflected radiation from a distance, usually from an aircraft or a satellite (Hay (*9*)). Thus, estimating land cover using remote sensing consists of collecting remotely sensed imagery as features and associating them with land cover classes, such as grass or shrub-

land, thus rendering a map of predicted land cover classes over a given area (Aplin (*10*)).

Remote sensing technologies have made remarkable improvements over the last decade following the launch of optical and synthetic aperture radars (SAR) remote sensing satellites (Kussul et al. (*11*)). The European Space Agency sent the Sentinel-1A/B and Sentinel-2A satellites within the Copernicus Program (Drusch et al. (*12*); Torres et al. (*13*)). Prior to this the United States of America launched Landsat-8 as part of the Landsat Project (Roy et al. (*14*)) and now Landsat-9 is available. These new technologies have generated enormous amounts of high-resolution images made freely available and open to the public, creating opportunities for a broader range of computer vision applications.

In addition to the progress made in remote sensing data, recent years have seen a wave of new and improved Artificial Intelligence models, which have in turn led to solid performance results in land cover classification and forecasting (Wulder et al. (*15*)).

3.2 Deep Learning for Land-Cover classification

With the surge in remote sensing data, new and more advanced Machine Learning models have been implemented to solve land cover classification tasks. Traditional supervised classification models such as Random Forest, Support Vector Machine, Naïve Bayes, Spectral Angle Mapper, Radial Basis Function, and Multilayer Perception have attracted much attention for solving land cover classification tasks (Talukdar et al. (*16*)). More recently, developments in Deep Learning techniques have shown outstanding performance results compared to traditional learning models. Their flexibility advantages and incurring lower modeling costs have made them widely used in fields where unstructured data, such as images and texts, are being analyzed.

Deep Learning is a sub-field of Machine Learning covering a family of algorithms whose structure mimics the neurons in the human brain (LeCun et al. (*17*)). While initially a single layer of neurons was employed, research over time demonstrated that the performance of these models

improved significantly by adding more layers, hence the term "Deep" Learning. These models take advantage of a backpropagation process whereby the model automatically adjusts its weights at each layer according to a performance measure. While traditional computer vision techniques required knowing which important annotated input features were associated with each image, Deep Learning techniques have overcome this by letting neural networks understand underlying patterns that are salient to each object (O'Mahony et al. (*18*)). Since the recent developments in Deep Learning theory, the most widely used models for computer vision tasks have based themselves on Convolutional Neural Networks (CNNs) and, even more recently, on Generative Adversarial Networks (GANs) (Cheng et al. (*19*)).

However, a significant challenge encountered with these techniques applied to remote sensing data is that a large amount of input labeled information is needed, which can be either difficult to obtain or costly to collect and annotate. The development of new techniques that leverage the available data has helped solve this problem.

3.3 Contrastive Learning, a Self-supervised technique

One such prominent technique is a subset of Unsupervised Learning called Self-supervised Learning. It is a powerful technique with which models are trained to learn about the data and the setting of the dataset without any annotations or labels, hence the term *Self-supervised Learning*. This approach leverages the underlying structure of the data to detect supervisory signals, which are then transferred to other downstream tasks where very few annotated labels are available. Specifically, using a "pretext task", the model first learns about the visual features of the data by training it on automatically generated labeled data and then transfers this knowledge to the actual computer vision task that needs to be solved (Jing and Tian (*20*)). Pretext tasks include, for example, image colorization (teaching the model to predict the actual colors of an image after converting it to grayscale), image super-resolution (the model learns to output

high-resolution images from pictures for which the resolution was decreased), image inpainting (cropping part of an image and teaching the model to identify the missing piece). Consequently, these methods serve as a backbone to the training process as they can significantly improve the performance of a model by feeding it only a small quantity of labeled data.

Self-supervised methods can be separated into Contrastive and Generative Learning models. These groups differ in the input data samples, the model's objective, and the underlying architecture (Liu et al. (*21*)). While Generative Learning models solely focus on identifying similar pairs of images, Contrastive Learning models additionally distinguish dissimilar pairs of images, often making them the preferred approach for image classification (Jaiswal et al. (*22*)).

Contrastive Learning models are characterized by their discriminative approach, whereby they find similar samples and group them together, and identify different samples and place them far from each other. Similar augmented images are generated and paired with the original images during the pretext task to form positive samples, while the rest are considered negative samples (Jaiswal et al. (*22*)). This step then allows the model to be trained to learn positive from negative samples and therefore understand the underlying visual representations that matter for the downstream task.

There exist many contrastive learning models depending on the task at hand. One of them is SimCLR, a state-of-the-art model presented by Chen et al (*19*) that has recently received much attention for its application to computer vision tasks (Siddiqui et al (*23*)). By using SimCLR as a pretext as a solution to the challenge of too little labeled data available, this paper presents a novel approach to land cover classification.

4 Methodology, Data and Results

4.1 Methodology

The procedure implemented in this paper is composed of three cascading actions. In the first place, an Unsupervised Model is trained; this action is needed to handle the scarcity of ground truth data. The model is trained to recognize similar pairs of images. Once the first model is trained, the refined weights are passed to the second model using Transfer Learning. The second model is a classifier that classifies tiles into the previously mentioned classes. Once the classifier can generalize, the model is applied to the country, district by district. The final output is a digital report containing the numerosity of each class for each district for the period of interest.

4.1.1 Unsupervised training as a backbone

A prominent self-supervised model for image representation tasks is called SimCLR.
SimCLR stands for "simple framework for Contrastive Learning" and is one of the most accessible models from this family. The idea of the technique is to force the model to pair the most similar couples of images into vectors, where each couple of images is composed of the original image and an augmented version from the same original acquisition. This action replaces the need for considerable amounts of labeled data, which are not easy to retrieve in several domains and often expensive (i.e., in-situ data in developing countries, medical annotations, etc., as in most real-world scenarios). Hence, this modeling technique provides an intelligent backbone at the basis of a classification model.
Specifically, the unsupervised training process can be defined in 3 steps:

1. **Data Augmentation**: A way to overcome the lack of labeled data is to automatically generate transformed pairs of images starting from a sample of images or tiles (see Figure

1). Two transformations are performed for each original tile in the dataset to have two similar pairs starting from the original tile. Therefore, the model is forced to learn that these two augmented tiles are similar in several images and come from the same original tile.

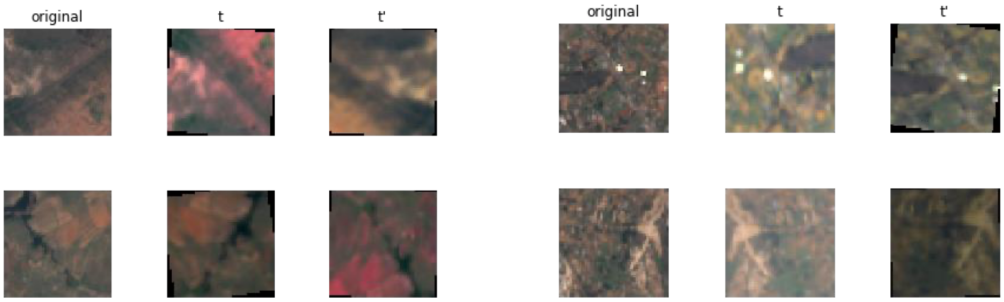

Figure 1: Data Augmentation applied to tiles.

In this project the original tiles are transformed by applying, in order, Random Cropping, Random Color Distortions, and Random Gaussian Blur.

2. **Vector Representation, Encoding** The augmented pairs, represented as vectors, are then passed through a Big-Convolutional Neural Network (Big-CNN). The goal is to train the model to have significantly correlated vectors for similar images and very uncorrelated vectors for dissimilar images.

3. **Maximize Agreement** In this last step, a measure of accuracy and a function to minimize the loss is needed to understand how the model works correctly and how to recalibrate the model's weights to improve its performance.

Once the pre-trained model has gained a general understanding of the data, it can be fine-tuned for a specific task (*downstream task*), such as image classification, where labels are scarce to improve label efficiency and potentially surpass supervised methods significantly.
Once the SimCLR model is trained on the contrastive learning task, it can be used for transfer

learning. For this, the representations from the encoder are used instead of representations obtained from the projection head. These representations can be used for downstream tasks like classifying each tile of a Sentinel2 acquisition into one of the types of land previously mentioned.

4.1.2 Supervised Fine-tuned model

Following the original simCLR paper, Resnet is the architecture chosen and used for training the head of the workflow. Resnet, which stands for "Residual Neural Network", is a very stable Deep Neural Network architecture widely used in computer vision tasks. It was released by Microsoft in 2015 following the idea that "the deeper, the better" when it comes to Convolutional Neural Networks. Deeper networks, having more parameters, are more flexible and capable of learning more up to some moments in which, after some depth, the performance decreases. This was the problem of the VGG architecture that going deeper with parameters starts losing generalization on validation and test set. One problem brought about by the depth of a network is the "vanishing gradient effect," which is tackled and solved by Resnet. The various type of Resnet that have been coded and trained are Resnet18, Resnet34, Resnet50, and Resnet152, each number symbolizing the number of layers used.

4.1.3 Knowledge Distillation, Inference

Once trained, the model used for classification is applied to the country pilot. Then, the model is ready to generalize and classify each tile of a country in each of the previously considered labels. Specifically, the model takes an aggregate acquisition as input and gives the numerosity in square kilometers of each class for each administrative level of the country.

4.1.4 Steps leading to the Digital Report

The final product is a digital report containing information about the numerosity of each class for each administrative level of a country to have a deep understanding of the physiognomy of the area. To have this, the complete workflow foresees the following steps:

1. Retrieving the coordinates in WKT format of the boundaries of each district of a country.

2. Downloading all the latest Sentinel2 acquisitions over the country and selecting the most recent and cloudless ones.

3. Merging all the different acquisitions in a single image without overlapping.

4. Cutting the single images into smaller tiles of 50x50 pixels, where a sample of about 10,000 unlabeled tiles is used for the Unsupervised pre-train model, and a sample of about 1,000 are hand-labeled and utilized for the Supervised fine-tune model. Minimum domain expertise is needed for both datasets to choose which tiles to use. To allow the model to better generalize, the tiles must be selected from different areas, and it is essential to have a balanced dataset.

5. Feeding the Unsupervised pre-train model. As an output of the semi-supervised model, it is expected to have refined weights capable of extracting high-level dataset features. As previously mentioned, this action is needed to replace the need for a considerable amount of labeled data.

6. Feeding the Supervised model. Refined weights trained by the unsupervised model are transferred to the first layers of the Supervised neural network. The last layers of the model are trained over the hand-labeled dataset and left free to learn the features of each of the eleven labels selected.

7. Running the inference. Each District is cut accordingly and divided into sub-tiles of 50 square pixels each. It is possible now to run the inference and let the model classify each tile into one of the eleven classes. The output is a report containing the numerosity of each class for each administrative level of a country and which answers to the question, “How many square meters of a class do we have for a given administrative level and in what percentage concerning the other classes?”

4.2 Data

4.2.1 Dataset

The images used for this task are 10 meters in resolution, and they are owned and distributed by ESA (European Space Agency), acquired by the Sentinel-2A and Sentinel-2B satellites launched respectively in 2015 and 2017 for the Sentinel2 earth observation mission from the Copernicus Program. These acquisitions are commonly called “Sentinels” and are widely used in the Earth Observation community because of their excellent resolution and because they are distributed for free.

4.2.2 Labels, types of land cover considered

The legend has been chosen according to the ESA World Cover product. It includes seven generic classes that appropriately describe the land surface at 10 meters: "Tree cover", "Shrubland", "Grassland", "Cropland", "Built-up", "Bare/sparse vegetation”, “Permanent water bodies”. These seven classes have been selected because they are the most expressive of the land cover of a country and because they are commonly reported by major Land Cover methods such as those released by Google, ESA, and FAO.

4.2.3 Case Study

The proposed methodology is applied to the kingdom of Lesotho. The kingdom of Lesotho is administratively divided into ten districts: Berea, Butha-Buthe, Leribe, Mafeteng, Maseru, Mohale's Hoek, Mokhotlong, Qacha's Nek, Quthing, and Thaba-Tseka. Each district is further divided into 80 Constituencies (or second administrative levels), the unit considered for this case study. As a result, the model's output consists of land cover estimations for the Second administrative level and above. The acquisitions considered for the project are those that are cloudless and the most recent possible. Specifically, these acquisitions come from a period ranging from late December 2021 to February 2022, when the model was ideated and trained.

4.3 Results

4.3.1 Benchmarking various Supervised fine-tune models

As previously mentioned, we select ResNet as the architecture for the training process. There are many variants of the ResNet architecture with different numbers of layers. The name ResNet followed by two or more-digit numbers indicates the number of layers of the given architecture. To choose the optimal architecture, we begin by training and evaluating ResNet18, ResNet34 with Mixup, ResNet50, and ResNet152. The table 1 below shows the reported scores:

	ResNet18	ResNet34+MU	ResNet50	ResNet152
Embedded Validation	0.93	0.83	0.95	0.84
Validation 1	0.56	0.18	0.957	0.45
Validation 2	0.36	0.22	0.96	0.6

Table 1: Validation results across selected ResNet architectures.

Given the results presented in table 1, ResNet50 is selected – that is, with 50 layers – as this solution perfectly generalizes on three different validation sets. In fact, the outstanding results obtained on all three sets, all above the 95 percent of accuracy, give the perception that the model perfectly extracts patterns of each label.

4.3.2 Coherence Analysis

We define a coherence index as the correlation between the same district Land Cover computed with different methods.

It is computed using the cosine similarity which calculates the similarity between two vectors of an inner product space. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction.

The coherence ranges between -1 and 1. The closer the measure is to 1 the more similar the two vectors are. The coherences between the methods are computed for each class and for each administrative level. The formula is the following:

$$\cos\varphi = \frac{CA'CB \cdot CA'CD}{|CA'CB| \cdot |CA'CD|}.$$

Tables 2 and 3 below show the results in terms of districts and classes.

District	wrt ESA WorldCover	wrt FAO-OCS	wrt Google DynamicWorld
Berea	0.99457	0.92289	0.56838
Butha-Buthe	0.99976	0.95944	0.74237
Leribe	0.99785	0.93260	0.63217
Mafeteng	0.91189	0.98585	0.29727
Maseru	0.99519	0.96304	0.71376
Mohale's Hoek	0.99229	0.98198	0.54063
Mokhotlong	0.99992	0.97339	0.94837
Quacha's Nek	0.99864	0.99389	0.97009
Quthing	0.99647	0.95886	0.64854
Thaba-Tseka	0.99956	0.96667	0.90062

Table 2: Coherence Indexes obtained by district relative to ESA, OCS, and DW.

Class	wrt ESA WorldCover	wrt FAO-OCS	wrt Google DynamicWorld
Built-up	0.97518	0.90572	0.94782
Cropland	0.97707	0.94429	0.47858
Tree	0.79038	0.81077	0.64517
Shrubland	0.50909	0.40835	0.76543
Grassland	0.99620	0.99799	0.95165
Water	0.80179	0.80817	0.48538

Table 3: Coherence Indexes obtained by class relative to ESA, OCS, and DW.

From the tables above, we find that with respect to the methodologies developed by ESA and OCS, coherences in terms of districts are always above 0.9. Analyzing the results by class, the highest coherences are obtained for permanent classes such as Built-Up areas. The lowest coherences, on the other hand, are linked to classes that strongly depend on seasonality such as Shrubland, Water and Tree areas.

5 Discussion: advantages, limitations and further ideas

This paper analyzed a new methodology primarily used in the field of Computer Vision exploiting a model from the family of Contrastive Learning techniques. This project perfectly meets the capability of our Lab of using non-conventional sources and state-of-the-art techniques in the field of Deep Learning with the need to have a quantitative estimation of Land Cover in poor countries without spending money on expensive High-Resolution images and/or in-situ data.
There are many advantages with applying this methodology to land cover classification. Compared to other methods implemented, the costs are incredibly low since it does not require large amounts of in-situ data which are expensive to obtain. The only costs incurred therefore arise from data storage and computing time. There is also great flexibility in applying this method, as it is very easy to add or remove new types of land cover to the dataset used in the training process. Finally, this method allows to harvest land cover results at any administrative level for a country in a few hours, depending on the extension of the land and the GPU used. Hence, this

methodology is perfectly tailored for situations where a country's setting changes rapidly, i.e., in the case of recent natural disasters such as major floods, earthquakes, and eruptions.

There also exist some limitations to this method. Specifically, the accuracy with respect to pixel-wise methodologies such as the ESA World Cover or Google Dynamic World is lower for two reasons. First, the methodology implemented in this paper is based on Contrastive Learning, which aims to classify tiles of pixels rather than a single pixel. In addition, replacing a huge amount of labeled training data with our approach implies a trade-off between lower cost but also lower accuracy. The limitations relative to the accuracy of the methods are clearly showed in Figure 2 and Figure 3 below:

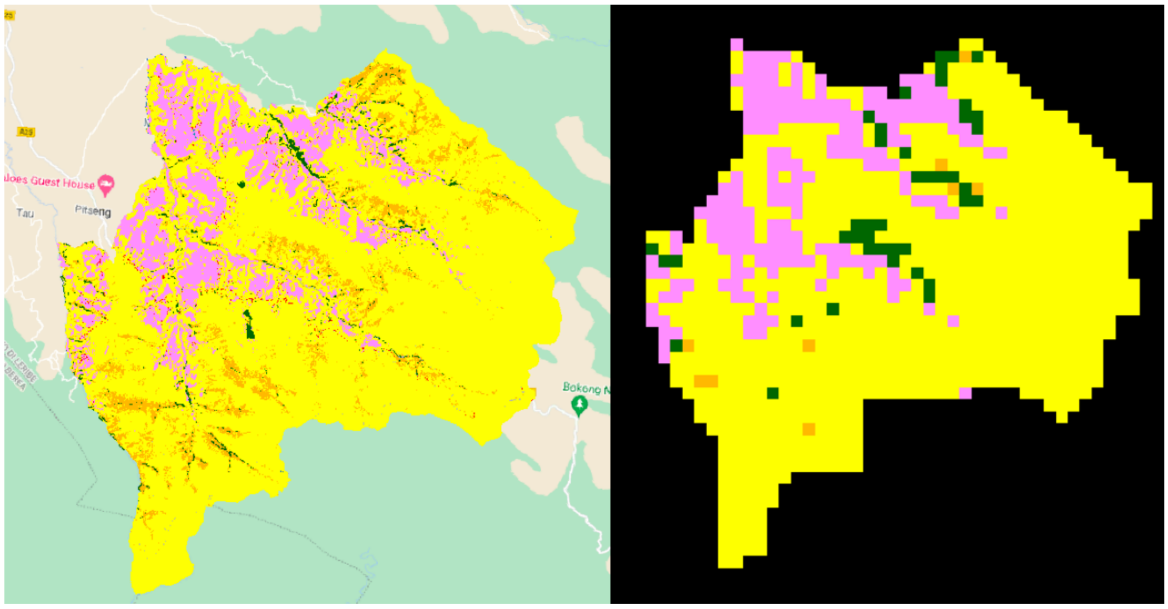

Figure 2: Bolahla, Lesotho: ESA World Cover appraoch versus FAO Data Lab approach.

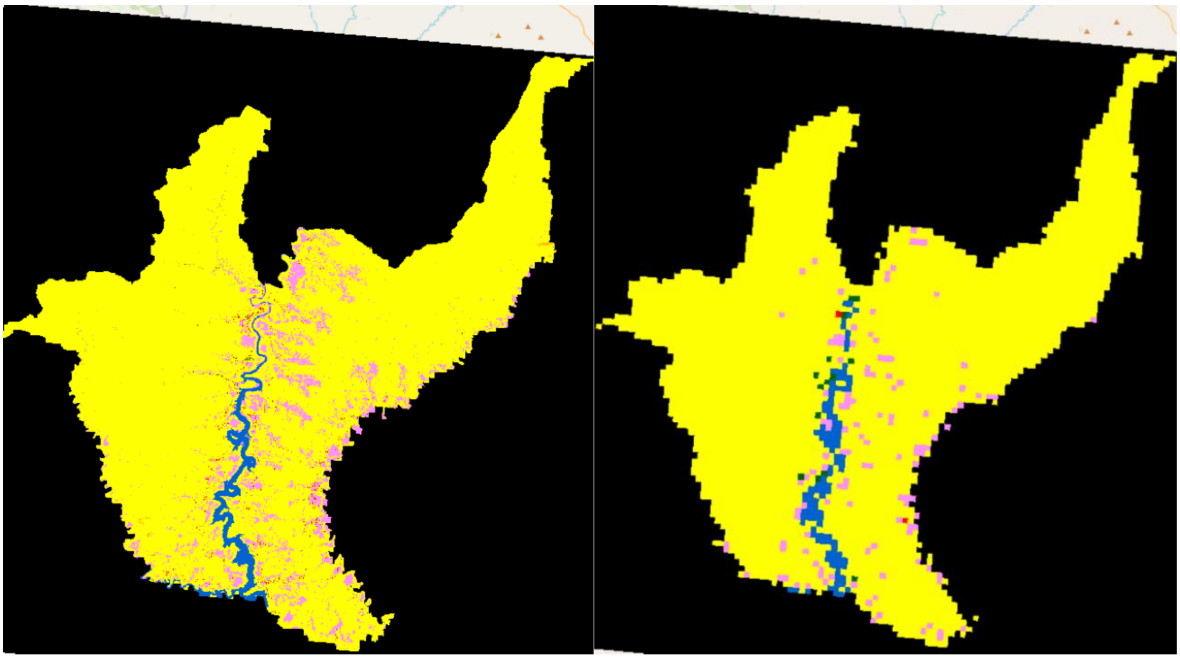

Figure 3: Matsoku, Lesotho: ESA World Cover appraoch versus FAO Data Lab approach.

While the ESA World Cover visualizations is a pixel-wise approach and therefore more precise, the FAO Data Lab results, which were obtain a low cost and with larger tiles, remain comparable as previously shown. Limitations are also relative to the extension of the country. The software runs perfectly for small regions (such as Lesotho, the country pilot of the exercise) while it could need massive resources for very extended regions.

There's room for so many improvements for this exercise and most of them are relative to the Data Fusion of several different source not taken into account such as the usage of Crop Calendars, the exploitation of refined indexes instead of pure RGB, the merge of Very-High-Resolution images when possible and the merge of socio-economic and ancillary data in general.

References and Notes

1. V. D. K. R. D. D. D. K. W. B. C. K. G. W. J. C. O. S. M. F. S. L. M. H. M. T. N. X. P. R. F. A. O. Zanaga, D. **ESA WorldCover 10 m 2021** (2022).

2. B. S. G.-W. B. e. a. Brown, C.F., *Sci Data* **Dynamic World, Near real-time global 10 m land use land cover mapping.** (2022).

3. S. K. Chauhan NK, *IEEE International conference on computing, power and communication technologies (GUCON)* **A review on conventional machine learning vs deep learning**, 347 (2018).

4. A. M. K. G. Dargan S, Kumar M, *Archives of Computational Methods in Engineering.* **A survey of deep learning and its applications: a new paradigm to machine learning.**, 1071 (2020).

5. K. C. Hoeser T, *Remote Sensing* **Object detection and image segmentation with deep learning on earth observation data: A review-part i: Evolution and recent trends.**, 1667 (2020).

6. C. C. Ball JE, Anderson DT, *Journal of applied remotesensing* **Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community.** (2017).

7. Z. X. Y. Y. Y. G. J. B. Ma L, Liu Y, *ISPRS journal of photogrammetry and remote sensing.* **Deep learning in remote sensing applications: A meta- analysis and review.**, 166 (2019).

8. H. E. García-Mora TJ, Mas JF, *International Journal of Digital Earth* **Land cover mapping applications with MODIS: a literature review.**, 63 (2012).

9. H. SI., *Advances in parasitology* **An overview of remote sensing and geodesy for epidemiology and public health application.**, 1 (2000).

10. A. P., *Progress in Physical Geography* **Remote sensing: land cover**, 283 (2004).

11. S. S. S. A. Kussul N, Lavreniuk M, *IEEE Geoscience and Remote Sensing Letters.* **Deep learning classification of land cover and crop types using remote sensing data.**, 778 (2017).

12. C. S. C. O. F. V. G. F. H. B. I. C. L. P. .-M. P. M. A. Drusch M, Del Bello U **Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote sensing of Environment.** (2012).

13. G. D. B. D. D. M. A. E. P. P. R. B. F. N. B. M.-T. I. Torres R, Snoeij P **GMES Sentinel-1 mission. Remote sensing of environment.**, 9 (2012).

14. L. T. W. C. A. R. A. M. H. D. I. J. J. D. K. R.-S. T. Roy DP, Wulder MA, *Remote sensing of Environment* **Landsat-8: Science and product vision for terrestrial global change research.**, 154 (2014).

15. R. D. W. J. H. T. Wulder MA, Coops NC, *International Journal of Remote Sensing* **Land cover 2.0**, 4254 (2018).

16. M. S. P. S. L. Y. R. A. Talukdar S, Singha P, *Remote Sensing* **Land-use land-cover classification by machine learning classifiers for satellite observations—A review.**, 1135 (2020).

17. H. G. LeCun Y, Bengio Y, *Nature* **Deep learning**, 436 (2015).

18. C. A. H. S. H. G. K. L. R. D. W. J. O'Mahony N, Campbell S, *Springer International Publishing* **Deep learning vs. traditional computer vision. In Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC)**, 128 (2020).

19. H. J. G. L. X. G. Cheng G, Xie X, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.* **Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities.**, 3735 (2020).

20. T. Y. Jing L, *transactions on pattern analysis and machine intelligence* **Self-supervised visual feature learning with deep neural networks: A survey.**, 4037 (2020).

21. H. Z. M. L. W. Z. Z. J. T. J. Liu X, Zhang F, *IEEE Transactions on Knowledge and Data Engineering* **Self-supervised learning: Generative or contrastive.**, 857 (2021).

22. Z. M. B. D. M. F. Jaiswal A, Babu AR, *Technologies* **A survey on contrastive self- supervised learning.**, 2 (2020).

23. A. S. Siddiqui SA, Dengel A, *IEEE* **Self-supervised representation learning for document image classification.**, 164358 (2021).