RAPIDA14EO: A MULTI-FORMAT DATASET FOR AUTOMATED LAND COVER CLASSIFICATION AND CHANGE DETECTION

Timothy Davis¹, Benjamin Bischke², Patrick Helber², Giovanni Marchisio³, Caglar Senaras³
Daniele Zanaga⁴, Ruben Van De Kerchove⁴, Annett Wania⁵

¹Planet Labs GmbH, Germany, ²Vision Impulse GmbH, Germany, ³Planet Labs Inc., USA,
⁴VITO NV, Belgium

ABSTRACT

RAPIDA14EO, a project funded under the European Union’s Horizon 2020 programme, aims to produce the largest corpus of remote sensing training data for land cover classification and change detection to-date. This corpus will be composed of imagery from Sentinel-2 as well as the Planet Fusion harmonization and gap-filling process, resulting in timeseries image cubes at multiple spatial resolutions and temporal cadences. Some 500,000 locations across Europe will be sampled to capture the continental distribution of CORINE land cover (CLC) classes and provide the training data for machine learning models to classify and detect changes in the next-generation product, CLC+.

Index Terms— Earth Observation Dataset, Machine Learning, Sentinel-2, PlanetScope, Change Detection

1. INTRODUCTION

Under the RapidAI4EO project¹, which receives funding from the European Union’s Horizon 2020 research and innovation programme, a consortium of partners are undertaking tasks to support the next generation of Copernicus Land Use Land Cover (LULC) products. Principal among these is the updated version of CORINE land cover (CLC) product, the so-named CLC+. To that end, the RapidAI4EO consortium is producing the largest remote sensing dataset for LULC classification and change detection to-date, which will be released as open-source in June 2022. This dataset will be multi-format, containing timeseries image cubes from Sentinel-2 as well as Planet Fusion processing for sensor-harmonized and gap-filled daily timeseries. The dataset will cover some 500,000 locations across Europe that have been selected to capture the distribution of classes in the 2018 CLC product. Each sample will be an area 640m×640m in size. This training corpus will provide the basis for a suite of machine learning approaches with the aim of automating the process of updating CLC+.

2. COPERNICUS LULC PRODUCTS

The CLC inventory [3] is a product of the Copernicus Land Monitoring Service (CLMS) that catalogs land cover across Europe with 44 distinct classes. The product has a resolution of 100 meters per pixel and a minimum mapping unit of 25 hectares. The first CLC inventory was released for the year 1990, with the most recent for 2018. Since 2000, CLC has been updated on a six-year basis and includes change mappings relative to the previous release.

A second generation of the CLC inventory, the CLC+, is under development. This product will maintain backwards compatibility with the legacy CLC, but expand the scope to include Land Use, Land-Use Change and Forestry (LULUCF) information. In order to achieve this, CLC+ has been modeled on the EIONET network’s EAGLE concept for a European land monitoring framework [1]. The EAGLE concept presents an object-oriented data model (OODM) that provides a common basis for comparing diverse LULC systems, such as national systems with transnational. This OODM likewise allows the much expanded CLC+ to be backwards-compatible with the legacy CLC.

3. RAPIDA14EO DATA SOURCES

The RapidAI4EO training corpus will be comprised of timeseries image cubes covering the year 2018 at each of the 500,000 locations. Dense information in the time domain is intended to allow for the development of novel LULC image classification and change detection models. At each location, there will be one image cube for each of the two satellite constellations from which the images are sourced: Sentinel-2 and PlanetScope. The spatial, radiometric, and temporal resolution of the cubes is dependent on the constellation, which are outlined in Table 1. The PlanetScope images will undergo an additional harmonization and gap-filling process known as Planet Fusion, which is detailed in Section 3.2. Similar to

¹This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 101004356.
²www.rapida14eo.eu
Table 1: Overview of the two satellite constellations used to derive the timeseries image cubes for the dataset.

<table>
<thead>
<tr>
<th></th>
<th>Sentinel-2</th>
<th>PlanetScope</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. satellites</td>
<td>2</td>
<td>180</td>
</tr>
<tr>
<td>No. spectral bands</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>Revisit time (days)</td>
<td>3-5</td>
<td>1</td>
</tr>
<tr>
<td>Ground resolution (m)²</td>
<td>10-60</td>
<td>3-4</td>
</tr>
<tr>
<td>Patch size (pixels)</td>
<td>64×64</td>
<td>214×214</td>
</tr>
<tr>
<td>Patch size (m)</td>
<td>640×640</td>
<td>642×642</td>
</tr>
</tbody>
</table>

BigEarthNet [9], we provide multi-label annotations for each location. The final component of the training corpus will be the associated metadata.

3.1. Sentinel-2

Sentinel-2 is a multispectral earth observation mission of the European Union’s Copernicus programme. Consisting of two satellites, this mission acquires imagery with 12 spectral bands in the visible and infrared range. The ground resolution is 10–60 meters, depending on the spectral band. Although Sentinel-2 has a repeat time between three and five days on average, many observations may contain clouds. To deal with such effects, we produce a timeseries with minimal cloud interference by fusing images from different time steps per month. We use atmospherically corrected Sentinel-2 Level 2A products and process all 12 available multispectral bands. As a result, each of the 500,000 sites is represented by a timeseries of 12 Sentinel-2 images with 10 meter ground resolution.

3.2. Planet Fusion

PlanetScope is a constellation of 180 cube satellites owned and operated by Planet Labs. These satellites capture imagery at nadir in four spectral bands from the visible and near-infrared (VNIR) range, with a ground resolution of up to three meters. PlanetScope has a daily revisit time, however it is also susceptible to interference from clouds and cloud shadows. In order to minimize this interference, the second set of data cubes included in the training corpus will be composed of images from a PlanetScope derivative, the harmonized and gap-filled Planet Fusion product.

The Planet Fusion process produces a daily timeseries of cloud-free surface reflectance images by applying the cubesat enabled spatio-temporal enhancement method (CESTEM) [5] [6]. CESTEM begins by harmonizing PlanetScope imagery against that of MODIS, Sentinel-2, and Landsat 8 to produce output with PlanetScope’s three meter ground resolution and daily cadence, and the radiometric consistency of the reference sensors.

CESTEM also includes a sophisticated approach to gap-fill pixels that are obfuscated by clouds, cloud shadows or other atmospheric phenomena, or in case no acquisition data are available. Gap-filling is informed by clear-sky observations on proximate days, as well as from the same time period in previous years. CESTEM applies a combined machine learning and regression approach trained on available observations to predict pixel values that require gap-filling. Quality assurance layers are also generated as part of the process to provide insight as to whether pixels were gap filled and if so how much information was available to inform the process. These layers allow users to make informed decisions about confidence of the gap-filling result and whether they would like to implement any masks.

![Fig. 1: NDVI analysis over an area of interest.](image-url)
The Planet Fusion data represent the highest spatial and temporal resolution data in the RapidAI4EO corpus with a ground resolution of three meters, daily cadence, and four spectral bands (VNIR) of surface reflectance. At this scale, distinct phenological patterns are easily interpretable and changes can be more accurately pinpointed in time. Consider the NDVI series presented in Figure 1. These plots represent an area of interest covering primarily agricultural land, with some forested areas and buildings present, visualized by Planet Fusion processing as acquired on 2018-07-01 in Figure 1c, and as the cloud-free Sentinel-2 sample for July 2018 in Figure 1d. Represented in each series plot in grey is the NDVI density, i.e. the proportion of image pixels that have a certain NDVI value, while the blue line shows mean NDVI. From Planet Fusion data in Figure 1a the growth patterns of the crops, as well as the revegetation of the fields following a mid-season harvest can be clearly differentiated from a background of uncultivated areas. In the case of Sentinel-2, shown in Figure 1b, the same trends are discernible but at a much coarser resolution.

4. SAMPLING OF LOCATIONS

When creating a representative earth observation dataset for machine learning, it is important to carefully consider the sampling strategy. For example, to facilitate model generalization and cope with regional differences, the sampling method should consider covering large spatial extents (e.g. to avoid spatial clusters) and include sufficient intra- and inter-class variance. Motivated by the large-scale spatial coverage in EuroSat [4], we sampled 500,000 locations over the 37 European countries present in CLC. To correctly represent each country in the dataset, we determine the number of patches to sample for each country based on the relative country surface area across all European countries.\(^3\) To avoid sampling from regional clusters, we overlay a grid over the European surface area and draw samples mainly from the checkerboard cells. We do not consider patches on the edge of the CLC map that contain a no-data class. Sea/ocean is a dominant class within the overall CLC distribution, however it is not useful for the land monitoring use case. Therefore, we intentionally undersample this class by filtering out patches that contain exclusively sea/ocean. For the remaining image patches, we perform stratified sampling with respect to the 44 classes in CORINE. We use additional statistics of each patch, such as cloud cover, slope, and snow, and prefer to sample locations with low values for these factors.

Figure 2 shows the CLC class distribution of the sampled patches compared with the overall CLC class distribution in log space. It can be seen that the sampled class distribution maintains that of the CLC inventory as a whole.

\(^3\) Overseas territories (e.g. La Reunion, France) are not considered. Only the surface area covered by the CLC inventory are considered.

5. DATASET STRUCTURE

The final training corpus will be published using emerging open source standards. The images themselves will be in a compressed GeoTIFF format based on the Cloud Optimized GeoTIFF (COG) standard [2]. COG is a specific configuration of a GeoTIFF and as such is fully compatible with any software or code that read GeoTIFFs in general. The COG specification is designed for efficiently reading data between and from remote sources, and contains optimizations for previewing large images and streaming portions of them. Given that the RapidAI4EO dataset will be composed of smaller image patches, the main advantage of the COG standard is lossless compression, which provides for more efficient storage without introducing artifacts that result from lossy compression.

![Comparison of CLC class distributions](image)

**Fig. 2:** Comparison of the CLC class distribution over the entire inventory and the sampled patches. After sampling, the original CLC class distribution is preserved.

The dataset as a whole will be discoverable using the Spatiotemporal Asset Catalog (STAC) specification [8]. STAC is a flexible standard for cataloging spatiotemporal assets, serving as an API-agnostic language for browsing and accessing the data. The categories and relationships within the catalog are wholly defined by the data provider, in this case the RapidAI4EO consortium. Given a multi-format and multi-purpose dataset as proposed here, the STAC specification provides several advantages for exposing data to the remote sens-
6. FUTURE WORK

Beyond the creation of the open-source training corpus presented in this paper, the RapidAI4EO project aims to experiment with machine learning models for LULC classification and change detection. A variety of supervised and unsupervised architectures will be explored. A key area of investigation for the machine learning experiments will be a comparison of models trained on the Sentinel-2 imagery with those trained on Planet Fusion. This will shed light on which types of class changes can be successfully captured by the models, and what spatial, temporal, and radiometric resolution of imagery is required to capture those changes. The dense temporal resolution of Planet Fusion, for example, may be necessary to identify complex phenological and harvest patterns.

Ultimately, these various change detection architectures that prove to be successful will be integrated to produce a heat map of change, identifying areas where the CLC+ is likely to require updating. The goal of this heat map is to feed into the formal process of updating the CLC+ and ultimately allow the inventory to be produced on a shorter cycle than has historically been possible. A demonstration of machine learning results under the RapidAI4EO project will be publicly released by the end of 2022 on the Copernicus DIAS ONDA [7] platform.

7. CONCLUSION

In this paper the RapidAI4EO consortium presented the training corpus for LULC classification and change detection that is currently under construction. We reviewed the multiple imagery sources that will be used, and corresponding output formats. The completed training corpus will be the largest existing training set for remote sensing classification and change detection, and will contain imagery at multiple spatial, spectral, and temporal resolutions. The ground truth for the corpus will be drawn from the 2018 CLC product. We also discuss that beyond providing these data as a resource to the broader remote sensing community, the consortium will use the training corpus to innovate machine learning models to support the next-generation CLC+ inventory.

REFERENCES


