DataCAP: A Satellite Datacube and Crowdsourced Street-level Images for the Monitoring of the Common Agricultural Policy*

Vasileios Sitokonstantinou[1,2][0000−0001−5506−2872], Alkiviadis Koukos[1], Thanassis Drivas[1], Charalampos Kontoes[1], and Vassilia Karathanassi[2]

[1] Institute for Space Applications and Remote Sensing, National Observatory of Athens, BEYOND Centre of EO Research and Satellite Remote Sensing, I. Metaxa and Vas. Pavlou St, Penteli, 15236 Athens, Greece

[2] Laboratory of Remote Sensing, National Technical University of Athens, 9 Heroon Polytechniou Str., Zographos, 15790 Athens, Greece

Abstract. Recently, massive amounts of satellite images are becoming available. The automated and efficient management, knowledge extraction and visualisation of these big earth data can enable the timely and comprehensive decision making in a number of operational scenarios. In this work, we demonstrate DataCAP that combines the Open Data Cube (ODC) technology on Satellite Image Time-series (SITS), with Machine Learning (ML) pipelines and crowdsourced street-level images to assist in the monitoring of the Common Agricultural Policy (CAP). DataCAP offers a suit of processing tools to simply and intuitively search, store and analyse radar and optical satellite images, along with visualisation tools that combine satellite and street-level imagery for the visual verification of algorithmic decisions.

Keywords: data cube · Sentinel-1 · Sentinel-2 · geo-tagged street-level pictures · satellite image time-series

1 Introduction

The Common Agricultural Policy (CAP) is transitioning to a new era, namely the CAP 2020+ reform, in which the current operating model is set to be simplified and improved significantly. This will be achieved by leveraging big satellite data and advanced ICT technologies that will ultimately form the so-called Area Monitoring System (AMS). The AMS is expected to fully automate and optimize the administration and control system of the CAP [2]. Until now, Paying Agencies (PAs), the implementation bodies of the CAP, have to inspect at least 5% of farmers' declarations, performing field visits or visual inspections on very high resolution satellite images. However, these methods are non-exhaustive, time-consuming, complex, and reliant on the skills of the inspector [8]. Recently,

* Supported by the ENVISION (No. 869366) and the CALLISTO (No. 101004152) projects, which have been funded by EU Horizon 2020 programs.

several solutions have emerged towards the development of a comprehensive AMS, taking advantage of Earth Observation (EO) data and Machine Learning (ML) techniques to systematically monitor the agricultural land over very large areas [8].

The backbone of all these solutions is the Copernicus program, specifically its Sentinel-1 (S1) and Sentinel-2 (S2) satellite missions. The S1 twin satellites are equipped with radar sensors, which capture useful information for the detection of mowing, grazing and harvest events. The S2 twin satellites are equipped with optical sensors, which capture multi-spectral images of the optical and near-infrared parts of the spectrum. The exploitation of S2 data enables crop type mapping, crop growth predictions and a plethora of other crop monitoring applications. The Sentinel data are freely available and characterized by high revisit frequency (6 days for S1 and 5 days for S2) and high spatial resolution (10-60 m), which are key features for an AMS.

In order to extract knowledge from Satellite Image Time-Series (SITS), and thereby make decisions on CAP compliance, ML algorithms are employed. In an AMS, SITS are combined with the parcel geometries (vector data) from the Land Parcel Identification System (LPIS) that has attached, for each parcel, the declared crop type label [8]. Two of the most common ML tasks, found in literature, are crop type classification and grassland mowing detection [6,7,3]. AMS can significantly improve the controls and reduce the administrative burden for PAs, however they are not equipped to manage big satellite data. For this reason, Analysis Ready Data (ARD) and multidimensional datacubes are key building blocks for such systems [5]. EO datacubes organise the data in such a way that anyone can intuitively exploit them. An EO datacube is an array of four dimensions; including longitude, latitude, time, and variable [1].

Having said that, EO-derived information is not panacea. Sentinels' spatial and temporal resolution limitations make it difficult to confidently decide on special scenarios, such as when we have small and narrow parcels, or broad crop type categories, or cloudy scenes. Therefore, EO data and EO-driven information needs to be accompanied by timely in-situ observations [3]. Typical in-situ data collection methods are expensive, time-consuming and therefore cannot provide continuous data streams. However, crowdsourced street-level images or images at the edge constitute an excellent alternative source [3].

In this work, we demonstrate DataCAP, an AMS module that comprises a Sentinel datacube, ML pipelines for crop classification and grassland mowing detection, and street-level images retrieved from the Mapillary API. DataCAP offers easy and efficient searching, storing, pre-processing and analyzing of big EO data, but also visualisation tools that combine satellite and street-level imagery for verifying algorithmic decisions. DataCAP comprises two components (Figure 1), the data management component (Section 2) and the visualisation component (Section 3).

2 Data management component

DataCAP's Data Management Component (DMC) is an automated module that searches, harvests (Figure 1 A) and pre-processes Sentinel data (Figure 1 B), which are then indexed as ARD in its datacube (Figure 1 C). This way, DataCAP offers easy and fast spatiotemporal data querying on SITS. The code for DataCAP DMC, including instructions on how to set up the datacube that facilitates it, can be found at https://github.com/Agri-Hub/datacap. DataCAP also enables the combination of satellite ARD with other data, including vector data for object-based image analysis, labelled data for supervised learning and crowdsourced street-level imagery for visual verification of ML outputs. The repository additionally includes two demo jupyter notebooks, showcasing the functionalities of DataCAP's DMC in the context of two CAP monitoring scenarios. Nevertheless, it should be noted that DataCAP is generic and can find applications in multiple other domains that leverage big satellite data and crowdsourced street-level images. Some indicative examples are land use/land cover change detection and burnt area mapping.

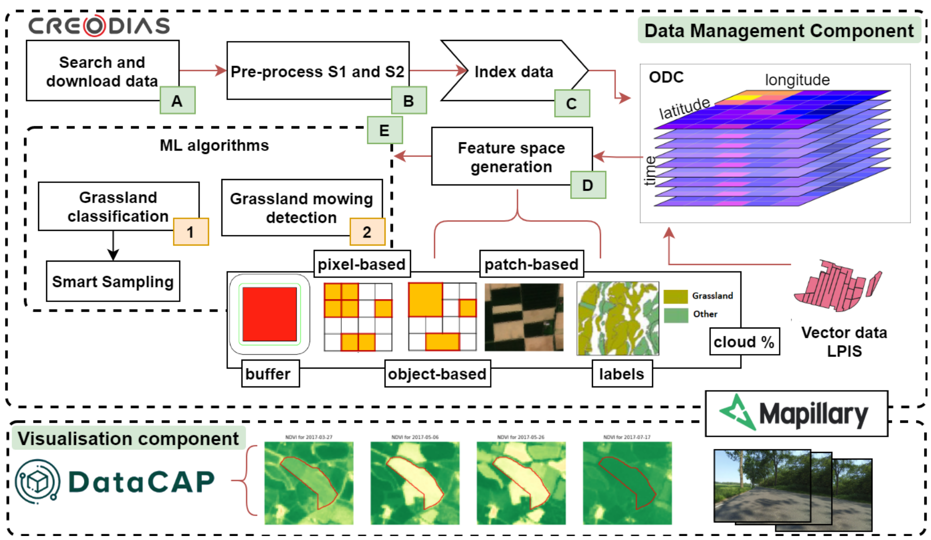

Fig. 1: DataCAP architecture: data management and visualisation components

DataCAP's datacube has been built on the Open Data Cube (ODC) software. ODC has been developed by Geoscience Australia (GA) and is supported by the Committee of Earth Observation Satellites (CEOS) [4]. ODC provides a user-friendly Python API, which loads data into Xarray data structures. This structure simplifies the management of multidimensional arrays, enabling the slicing of data on each of the dimensions of the cube. Currently, our datacube contains time-series of S1 and S2 ARD (2017) over Netherlands (6,375 km^2).

A - Search and download The S1 and S2 data products are automatically harvested and downloaded from the CreoDIAS API[3]. The user can request Sentinel products within a specified time window over the area of interest, along with other user defined parameters such as the maximum allowed cloud coverage. The metadata of each GET response is stored in a PostgreSQL/PostGIS database in order to have full control and geo-spatial querying capabilities over the products that have been downloaded.

B - Pre-processing The downloaded products are then processed to ARD. The output of this step returns time-series of i) S1 backscatter coefficient and coherence products, ii) S2 atmospherically corrected multi-spectral imagery; along with a scene classification product that includes clouds, dark pixels etc.

C - Index to ODC The ARD are automatically loaded into the datacube, triggering a batch process whenever a new image is downloaded and preprocessed. ODC provides two methods for loading data, indexing or ingesting. The indexing method stores only the metadata of each product in the database, while having the actual data products in a file system. On the other hand, the ingestion method stores the data in both. According to ODC documentation[4] Cloud Optimized Storage formats combined with GDAL or other software improve the performance of reading files, making indexing the preferred choice.

The indexing is done using YAML files, which are blueprints for storing each data type (i.e. coherence, backscatter, multi-spectral images).

D - Data analytics and feature engineering

DataCAP assists on the fast, easy and versatile generation of SITS feature spaces (Figure 1 D) to feed ML pipelines. Users can execute i) a number of complex spatio-temporal queries using the LPIS (vector data) and scene classification products (i.e. cloud mask); ii) create pixel-based, object-based or patch-based feature spaces; iii) apply inward buffers to avoid mixed pixels and more. For our two downstream tasks, we generated an *object-based (mean)* S1/S2 time-series, from *01/03/2017 - 31/10/2017* over *Utrecht, Netherlands*, with max cloud coverage *>85%* and inward buffer *10 m*.

E - 1 Crop classification and smart sampling In DataCAP we use the crop classification algorithm from Sitokonstantinou et al. (2018) that we slightly modified for mapping grasslands over Utrecht, Netherlands [8]. Then the classification results are passed through the smart sampling algorithm from [6]. The smart sampling algorithm utilizes the posterior probabilities of the classification to return to the user the most confident mismatches, between crop predictions and farmer declarations, for further visual inspection through DataCAP's Visualisation component.

E - 2 Grassland mowing event detection DataCAP leverages the Sen4CAP[5] grassland mowing detection algorithm. The algorithm calculates vegetation index time-series trends and identifies abrupt drops. If the drops are larger than a pre-defined threshold then they are characterized as a mowing event.

[3] https://finder.creodias.eu/
[4] https://datacube-core.readthedocs.io/en/latest/ops/ingest.html
[5] http://esa-sen4cap.org/

3 Visualisation component

DataCAP's visualisation component is demonstrated at http://62.217.82.91/. This demo GUI is structured based on the two downstream tasks presented earlier and enables the verification of the ML outputs via means of visual inspection. The interface offers two types of visualisations, i) time-series of S1 and S2 images and ii) crowdsourced street-level images.

Using the API of the crowdsourcing platform Mapillary, we automatically download all available street-level images over the area and time window of interest. Then, each downloaded image is matched with the corresponding LPIS object(s) it illustrates. We annotate images that are taken either towards the windshield direction (Case 1) or the window direction (Case 2).

We use Equations 1 and 2 to move the initial geo-location coordinates ($\boldsymbol{lat_1}$, $\boldsymbol{lon_1}$) to new coordinates ($\boldsymbol{lat_2}$, $\boldsymbol{lon_2}$) that are $d = 10m$ away in the direction of angle θ. For Case 1, we set $\theta = compass\_angle + 45^o$ for the right half of the image and $\theta = compass\_angle - 45^o$ for left half. For Case 2 we set $\theta = compass\_angle$.

$$lat_2 = \arcsin\left(\sin lat_1 \cdot \cos \frac{d}{R}\right) + \cos lat_1 \cdot \sin \frac{d}{R} \cdot \cos \theta) \tag{1}$$

$$lon_2 = lon_1 + \arctan\left(\sin \theta \cdot \sin \frac{d}{R} \cdot \cos lat_1, \cos \frac{R}{d} - \sin lat_1 \cdot \sin lat_2\right), \tag{2}$$

where $\boldsymbol{R}$ is the radius of the Earth. If the new coordinates of an image fall within any parcel geometry, then we match the parcel(s) label with that image. The code and the produced annotated street-level images are open.

Visual inspection of parcels to verify ML results. One can select the flagged parcels according to the smart sampling algorithm, and can verify if the farmer declaration or ML prediction was correct. This is done by inspecting parcel-focused time-series of S1 and S2 images, with an adjustable buffer around the parcel. The time-series reveal the crop growth for each parcel, thus allowing to distinguish between different crop types. In the same manner, visual inspection of the time-series of grassland parcels can reveal sudden changes in the vegetation cover, thus indicating mowing events. Finally, the street-level images offer a very high resolution complementary information to finalise the decision.

4 Conclusion

In this demo paper, we presented DataCAP, a data handling and visualisation module for the monitoring of the CAP. DataCAP consists of i) a back-end component that helps collect and prepare satellite ARD to feed pertinent ML pipelines, and ii) a front-end component that utilizes the satellite ARD and street-level images to help verify the ML outputs. The demonstrated solution is scalable, extendable and reproducible. DataCAP's code and produced annotated datasets are open, encouraging the data science community to exploit them in similar or other pertinent domains and applications.

References

1. Appel, M., Pebesma, E.: On-demand processing of data cubes from satellite image collections with the gdalcubes library. Data **4**(3) (2019). https://doi.org/10.3390/data4030092, https://www.mdpi.com/2306-5729/4/3/92
2. Devos, W., Lemoine, G., Milenov, P., Fasbender, D.: Technical guidance on the decision to go for substitution of otsc by monitoring (2018). https://doi.org/10.2760/693101, https://publications.jrc.ec.europa.eu/repository/handle/JRC112918
3. D'Andrimont, R., Lemoine, G., Van der Velde, M.: Targeted grassland monitoring at parcel level using sentinels, street-level images and field observations. Remote Sensing **10**(8) (2018). https://doi.org/10.3390/rs10081300, https://www.mdpi.com/2072-4292/10/8/1300
4. Killough, B.: Overview of the open data cube initiative. In: IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium. pp. 8629–8632 (2018). https://doi.org/10.1109/IGARSS.2018.8517694
5. Picoli, M.C., Simoes, R., Chaves, M., Santos, L.A., Sanchez, A., Soares, A., Sanches, I.D., Ferreira, K.R., Queiroz, G.R.: Cbers data cube: a powerful technology for mapping and monitoring brazilian biomes. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **3**, 533–539 (2020)
6. Rousi, M., Sitokonstantinou, V., Meditskos, G., Papoutsis, I., Gialampoukidis, I., Koukos, A., Karathanassi, V., Drivas, T., Vrochidis, S., Kontoes, C., et al.: Semantically enriched crop type classification and linked earth observation data to support the common agricultural policy monitoring. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **14**, 529–552 (2020)
7. Sitokonstantinou, V., Koutroumpas, A., Drivas, T., Koukos, A., Karathanassi, V., Kontoes, H., Papoutsis, I.: A sentinel based agriculture monitoring scheme for the control of the cap and food security. In: Eighth International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2020). vol. 11524, p. 1152407. International Society for Optics and Photonics (2020)
8. Sitokonstantinou, V., Papoutsis, I., Kontoes, C., Lafarga Arnal, A., Armesto Andrés, A.P., Garraza Zurbano, J.A.: Scalable parcel-based crop identification scheme using sentinel-2 data time-series for the monitoring of the common agricultural policy. Remote Sensing **10**(6), 911 (2018)