https://doi.org/10.5194/essd-2022-298
Preprint. Discussion started: 13 September 2022

Open Access Earth System Science Data Discussions

AI4Boundaries: an open AI-ready dataset to map field boundaries with Sentinel-2 and aerial photography

Raphaël d'Andrimont[1], Martin Claverie[1], Pieter Kempeneers[1], Davide Muraro[1], Momchil Yordanov[1], Devis Peressutti[2], Matej Batič[2], and François Waldner[1]

[1]European Commission Joint Research Centre (JRC), Ispra, Italy
[2]Sinergise, Ljubljana, Slovenia

Correspondence: Raphaël d'Andrimont (raphael.dandrimont@ec.europa.eu)

Abstract. Field boundaries are at the core of many agricultural applications and are a key enabler for operational monitoring of agricultural production to support food security. Recent scientific progress in deep learning methods has highlighted the capacity to extract field boundaries from satellite and aerial images with a clear improvement from object-based image analysis (e.g., multiresolution segmentation) or conventional filters (e.g., Sobel filters). However, these methods need labels to be trained on. So far, no standard data set exists to easily and robustly benchmark models and progress the state of the art. The absence of such benchmark data further impedes proper comparison against existing methods. Besides, there is no consensus on which evaluation metrics should be reported (both at the pixel and field levels). As a result, it is currently impossible to compare and benchmark new and existing methods. To fill these gaps, we introduce AI4Boundaries, a data set of images and labels readily usable to train and compare models on the task of field boundary detection. AI4Boundaries includes two specific data sets: (i) a 10-m Sentinel-2 monthly composites for large-scale analyses in retrospect, (ii) a 1-m orthophoto data set for regional-scale analyses such as the automatic extraction of Geospatial Aid Application (GSAA). All labels have been sourced from GSAA data that have been made openly available (Austria, Catalonia, France, Luxembourg, the Netherlands, Slovenia, and Sweden) for 2019 representing 14.8 M parcels covering 376 K km^2. Data were selected following a stratified random sampling drawn based on two landscape fragmentation metrics, the perimeter/area ratio and the area covered by parcels, thus taking into account the diversity of the agricultural landscapes. The resulting "AI4Boundaries" dataset consists of 7,831 samples of 256 by 256 pixels for the 10-m Sentinel-2 dataset and of 512 by 512 pixels for the 1-m aerial orthophoto. Both datasets are provided with the corresponding vector ground-truth parcel delineation (2.5 M parcels covering 47,105 km^2) and with a raster version already pre-processed and ready to use. Besides providing this open dataset to foster computer vision developments of parcel delineation methods, we discuss perspectives and limitations of the dataset for various types of applications in the agriculture domain and consider possible further improvements.

https://doi.org/10.5194/essd-2022-298
Preprint. Discussion started: 13 September 2022

Open Access Earth System Science Data Discussions

1 Introduction

Field boundaries are at the core of many agricultural applications such as mapping crop types and yield estimation. With the development digital farming platforms, extracting and updating field boundaries automatically has gained much traction to facilitate customer onboarding. Different spatial and temporal data coverage could be needed according to desired application.

There are three broad types of methods to map field boundaries: deep learning, object-based image segmentation and conventional (edge-detection) filters. Deep learning methods can extract field boundaries from satellite/aerial images better than object-based image analysis (e.g, multiresolution segmentation) or conventional filters (Sober filters) because they can learn to emphasize relevant image edges while suppressing others. For instance, Waldner and Diakogiannis (2020) and Waldner et al. (2021) have shown that convolutional neural networks are capable of learning complex hierarchical contextual features from the image to accurately detect field boundaries and discard irrelevant boundaries, thereby outperforming conventional edge filters. More recently, similar approach has been used to unlock large-scale crop field delineation in smallholder farming systems with transfer learning and weak supervision (Wang et al., 2022). Deep learning methods need labels for training and evaluation. No benchmark data set exists to easily do so. Absence of such benchmark data impedes proper comparison with existing methods. Besides, there is no consensus on which evaluation metrics should be reported (both at the pixel and field levels). As a result, it is currently challenging to benchmark new and existing methods.

Deep learning parcel delineation based on the land parcel identification system has been evaluated in several countries such as in France (Aung et al., 2020), Netherlands (Masoud et al., 2019) and Spain (Garcia-Pedrero et al., 2019). However, as there is no European harmonised land parcel identification system, there is no dataset to properly benchmark methods over a variety of landscapes and latitude.

The Geospatial Aid Application (GSAA) refers to the annual crop declarations made by European farmers for Common Agricultural Policy (CAP) area-aid support measures. The electronic GSAA records include a spatial delineation of the parcels. A GSAA element is always a polygon of an agricultural parcel with one crop (or a single crop group with the same payment eligibility). The GSAA is operated at the region or country level in the EU-28, resulting in about 65 different designs and implementation schemes over the EU. Since these infrastructures are set up in each region, data are not interoperable at the moment, and the legends are not semantically harmonised. Furthermore, most GSAA data are not publicly available, although several countries are increasingly opening the data for public use. In this study, seven regions with publicly available GSAA are selected representing a contrasting gradient across the European Union (i.e., agricultural system depends on physical and human geography resulting in contrasted landscapes).

Despite these recent advances in parcel delineation with deep learning, there is a lack of dataset to benchmark and compare methods. Artificial Intelligence (AI) and Machine Learning (ML) algorithms have great potential to advance processing and analysis of Earth observation data. Training datasets are crucial for AI applications but they are becoming a major bottleneck in more widespread and systematic applications of AI/ML.

Creating reference AI data sets in remote sensing has been shown to accelerate method development and to help push the boundary of the state of the art. For instance, data sets such as BigEarthNet (Sumbul et al., 2019) and EuroSAT (Helber et al.,

https://doi.org/10.5194/essd-2022-298
Preprint. Discussion started: 13 September 2022


Open Access Earth System Science Data Discussions

2019) have been used for generic land cover. For agriculture, most of the previously published datasets over Europe are focusing on France (BreizhCrop and PASTIS; Rußwurm et al., 2019; Tarasiou et al., 2021) or France and Catalonia (Sen4AgriNet; Sykas et al., 2022). In addition to the fact that no harmonized dataset is currently available for multiple countries in Europe, no dataset combining remote sensing and very high resolution aerial imagery has yet been published.

To fill these gaps, we release two AI-ready data sets (pairs of images and labels) for field boundary detection to facilitate model development and comparison:

1. A multi-date dataset based on Sentinel-2 monthly composites for large-scale analyses in retrospect.
2. A single-date dataset based on orthophoto for regional-scale analyses such as the automation of GSAA.

All labels are sourced from public parcel data (GSAA) data that have been made openly available.

2 Data and study area

2.1 Sampling

The rationale behind this study is to propose ready-to-use data set of Earth observation data with corresponding parcel boundaries. Public parcel data are first obtained over several countries/regions (i.e., Austria, Catalonia, France, Luxembourg, Netherlands, Slovenia, Sweden) for the year 2019. After drawing a grid of cells of 4 by 4 km in the ETRS89-extended LAEA Europe projection (EPSG:3035), a stratified random sampling is drawn based on two variables:

1. the average parcel perimeter/area ratio (PAR) computed for each cell is then distributed over 5 percentile bins.
2. the coverage percentage of parcels within the cell distributed in 10 classes.

We designed a random stratified sampling method to extract the image chips from a variety of landscapes. First, the 4x4 km grid was overlaid over each country/region where parcel data are available. In each grid cell, the field fraction (in percent) and the perimeter/area ratio (PAR) were computed as shown for France in Figure 1. These indicators jointly describe the prevalence of agriculture (i.e. land proportion covered by agriculture) and the landscape fragmentation (i.e. perimeter area ratio) of each grid cell. Fifty strata were defined by discretising the field fraction in 10 classes (from 0 to 100% by step of 10%) and the perimeter area ratio in 5 classes defined by its 20^{th} percentile in order to obtain a representative sampling. A sample of 10,000 units was selected from the available grid cells, 170 per stratum (Figure 2). In those strata where the number of samples is larger than the number of available grid cells, the sampling units in excess were evenly distributed to the other strata. Within each stratum, grid cells were selected so as to maximise the balance between source regions.

The resulting sampling results in 7,831 samples distributed as described in Table 1 and in Figure 3.

The samples are mainly distributed from North to South (Figure 3).

https://doi.org/10.5194/essd-2022-298
Preprint. Discussion started: 13 September 2022

Open Access Earth System Science Data Discussions

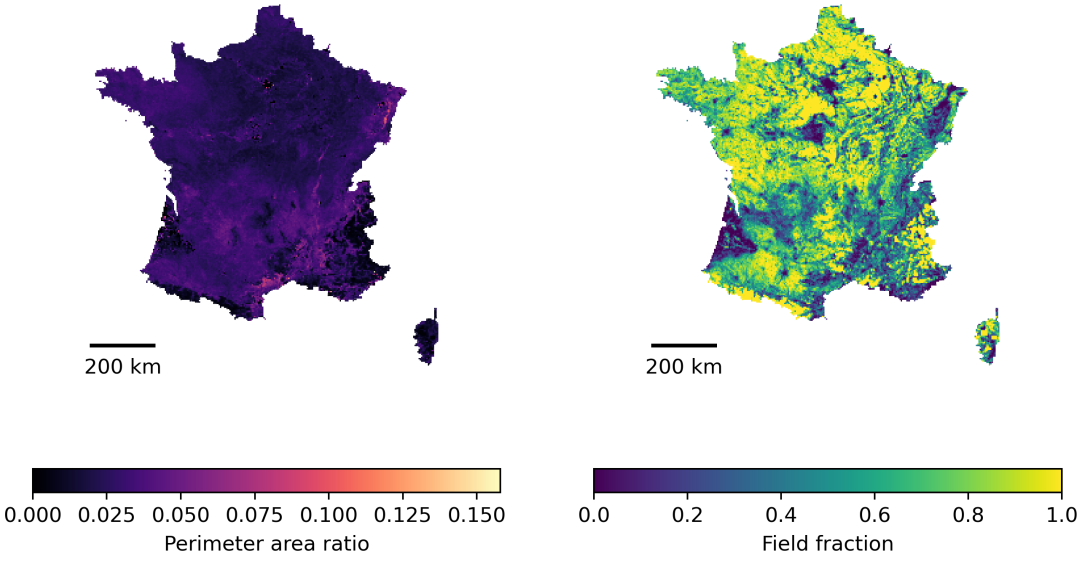

Figure 1. The stratification of the sampling is done based on Perimeter area ratio (left figure) and proportion of parcels (right figure) in 4-km × 4-km grid cells. Example across France.

Table 1. Distribution of the final stratified sampling for each region.

Country/Region	Number of sampling units
Austria	2091
Catalonia	652
France	2078
Luxembourg	132
Netherlands	1157
Slovenia	301
Sweden	1420
Total	7831

2.2 Earth Observation data

Image chips are required to feed deep learning models, image chips are required for which the EO data are extracted for the 2-specific dataset as shown in Figure 4 :

1. Sentinel-2 monthly composites (March to August 2019) 256 by 256 pixels
2. Orthophoto single-temporal imagery resampled at 1-m resolution 512 by 512 pixels

The difference of pixel extent (256 vs 512) of the two dataset is linked to the spatial resolution of Sentinel-2 (10 m) and orthophoto (1 m) respectively corresponding thus to 2560 m by 2560 m and 512 m by 512 m. The number of pixels for the orthophoto data set had to be increased to provide sufficient context.

https://doi.org/10.5194/essd-2022-298
Preprint. Discussion started: 13 September 2022

Open Access Earth System Science Data Discussions

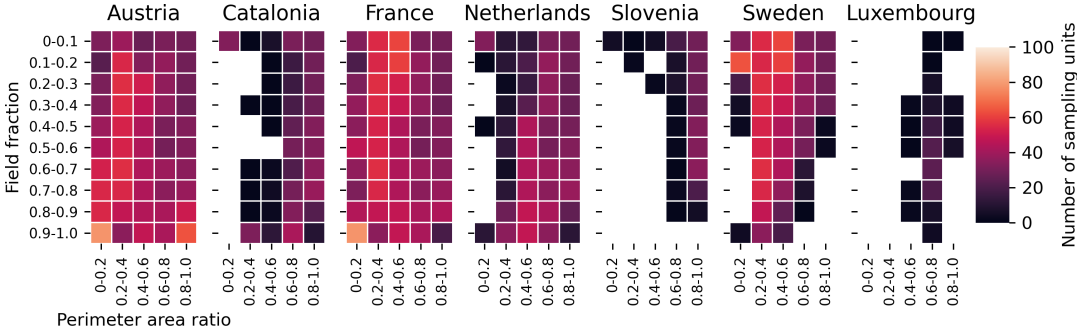

Figure 2. Distribution of sampling units among the seven regions.

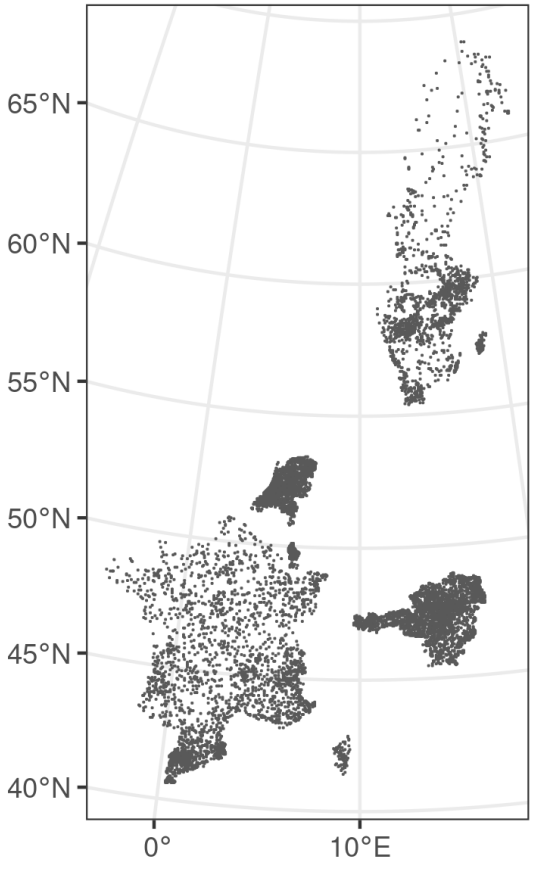

Figure 3. Spatial distribution of sampling units among the seven regions. [suggestion = overlay the country (or region) and name them with arrows]

https://doi.org/10.5194/essd-2022-298
Preprint. Discussion started: 13 September 2022

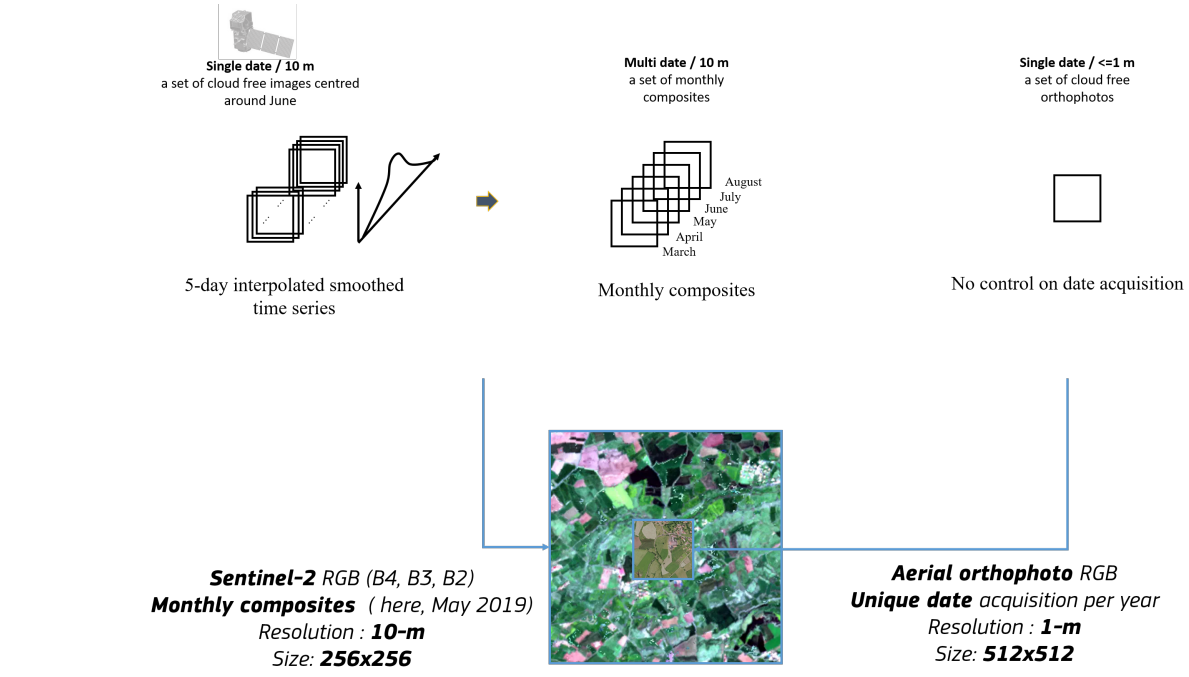

Figure 4. Earth Observation generation overview. Sentinel-2 time series are interpolated and smoothed to generate 10-m monthly composites cropped on 256 by 256 pixel. Aerial othophotos are resampled at 1-m and cropped on 512 by 512 pixels in the center of the 4-km sampled cell.

2.2.1 Sentinel-2

This section describes how the monthly cloud-free Sentinel-2 surface reflectance composites for March to August 2019 (thus 6 months of 4 bands: R, G, B; NIR) were produced.

The Sentinel-2 Level-1C top of atmosphere (TOA) reflectance data were obtained and processed using the Sen2Cor processor (Main-Knorn et al., 2017) from the ESA SNAP toolbox to generate the surface reflectance (SR). The four spectral bands that are available at a spatial resolution of 10 m were selected (B2, B3, B4, and B8). The Scene Classification Layer (SCL) obtained from Sen2Cor was added as an extra band. Sentinel-2 processing were performed on the BDAP (Soille et al., 2018) using the open source pyjeo (Kempeneers et al., 2019) Python package.

Data cubes, consisting on merging all 2019 acquisitions for all 4x4 sq.km chips, were created. The Data cubes were extended with the acquisitions of the preceding (December 2018) and successive (January 2020) months. The extra observations served to mitigate the boundary effects at the beginning and end of the time series while applying temporal operations. These months were removed after the filter was applied.

Only pixels identified as dark (SCL=2), vegetated (SCL=4), not-vegetated (SCL=5), water (SCL=6), and unclassified (SCL=7) were considered as “clear”. In addition, outliers were detected using the Hampel identifier (Hampel, 1974), based on the pixel

https://doi.org/10.5194/essd-2022-298
Preprint. Discussion started: 13 September 2022

values in the red (B4) and near infrared (B8) bands. The Hampel filter calculates the median and the standard deviation in a moving window, expressed as the median absolute deviation (MAD). For the moving window, a width of 40 days was considered. Pixels below two and above three standard deviations from the median were identified as outliers, for the NIR and the red bands, respectively. The respective values of two and three standard deviations for the lower and upper bounds were selected ad-hoc based on visual inspection of the results. The outlier in the red and NIR band domain were used to identify omitted clouds and omitted cloud-shadows, respectively. The masked pixels from the SCL and the detected outliers were removed from the time series and replaced by a linearly interpolated time series using "clear" observations; in case of outliers near the beginning and end of the time series, values were extrapolated to the nearest "clear" observation. The resulting time series was then resampled to obtain gap-filled data cubes by taking the mean of the filtered and interpolated values every 5 days.

Despite the SCL masking and the outlier detection, the resulting time series were still noisy. This is due to residual of atmospheric correction and non-accounted bidirectional reflectance distribution function (BRDF) effects. A subsequent smoothing filter was therefore applied, the recursive Savitzky-Golay filter (Chen et al., 2004). The original implementation has been developed for NDVI time-series. It was adapted in this study to smooth surface reflectance values. A total of 15 observations at 5-day temporal resolution were used for the smoothing window size: 7 leftward (past) and 7 rightward (future) observations. The order of the smoothing polynomial was set to 2.

A monthly composite image was then calculated, resulting in 12 observations for the year 2019 for each spectral band (B2, B3, B4, and B8). The composite was calculated as the median value of the smoothed values within each month.

2.2.2 Aerial Orthophoto imagery

Orthophoto extraction was done using public WMTS and WMS services for the seven regions (see details in Table A1). The orthophoto have a 1-m resolution and an extent of 512 by 512 pixels centered on the centroid of the sampled cells. After downloading a larger extent in the geographic reference system provided by the service, the samples are reprojected to EPSG 3035, cropped to the exact extent and standardised on 3 RGB by removing the NIR band when available to have consistent dataset. Finally, the histogram extraction of each sample has been used to filter out 233 of the 7,831 samples (almost exclusively in Sweden for which filtering removed images with no data). Random examples for each country are shown in the Figure 7 along with the vector label data.

2.3 Label data

The labels are obtained from vector parcels of the GSAA for each specific regions. The GSAA refers to the annual crop declarations made by EU farmers for CAP area-aid support measures. The electronic GSAA records include a spatial delineation of the parcels. A GSAA element is always a polygon of an agricultural parcel with one crop (or a single crop group with the same payment eligibility). The GSAA is operated at the region or country level in the EU-28, resulting in about 65 different designs and implementation schemes over the EU. Since these infrastructures are set up in each region, at the moment data is not interoperable, nor are legends semantically harmonised. Furthermore, most GSAA data is not publicly available, although several countries are increasingly opening the data for public use. In this study, seven regions with publicly available GSAA

https://doi.org/10.5194/essd-2022-298
Preprint. Discussion started: 13 September 2022

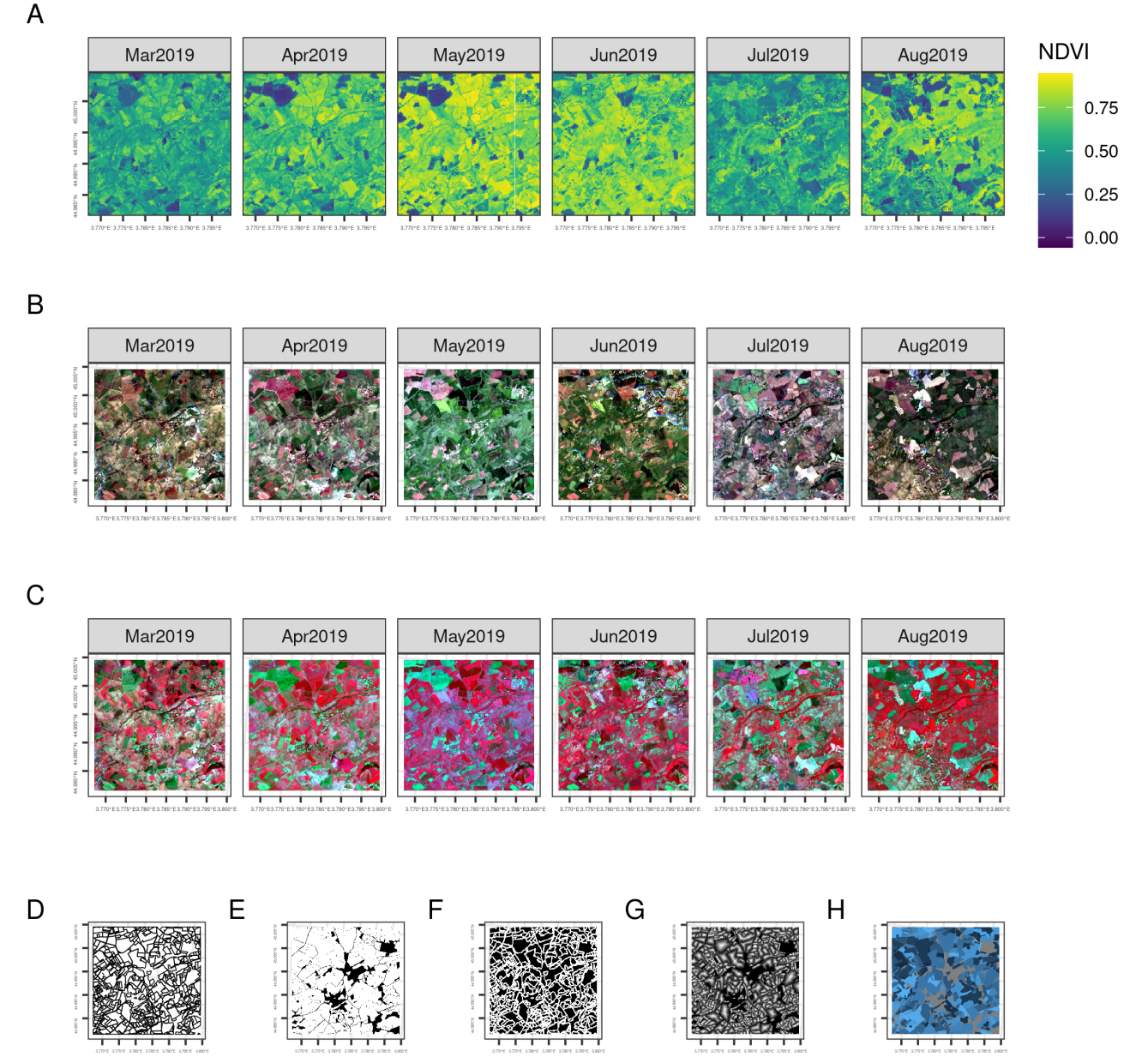

Figure 5. Examples of the Sentinel-2 10-m dataset consisting of an extent of 256 pixels of 10-m (thus 2560 m by 2560 m). (A) shows NDVI monthly composite, (B) RGB monthly composites and (C) NIR false colour monthly composites. The label at the same resolution and extent consist in four layers : (E) an extent mask, (F) boundary mask , (G) a distance mask and (H) a field enumeration.

https://doi.org/10.5194/essd-2022-298
Preprint. Discussion started: 13 September 2022

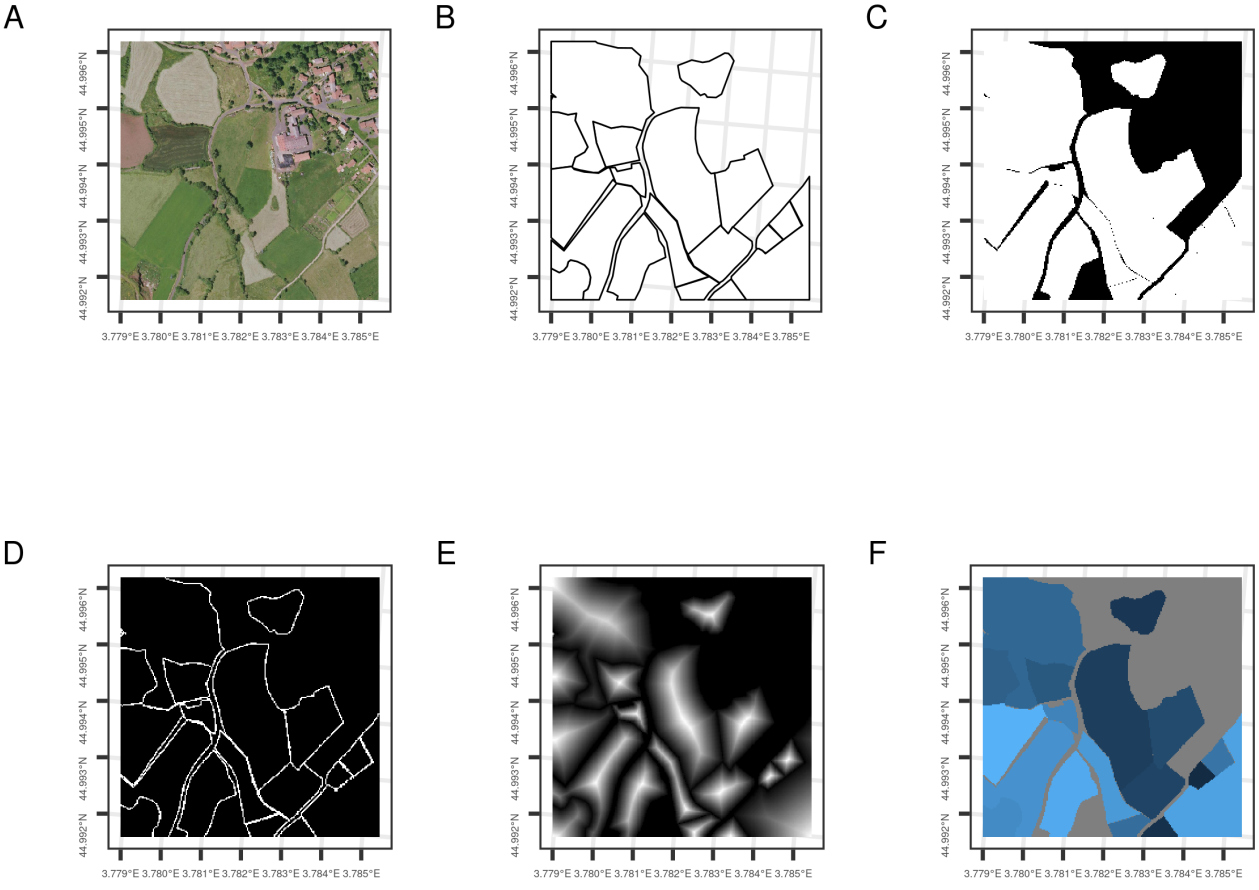

Figure 6. Examples of the aerial ortophoto 10-m dataset consisting of an extent of 512 pixels of 1 m (thus 512 m by 512 m). (A) aerial orthophoto RGB. The vector label (B) and the raster label at the same 1-m resolution and extent consist in four layers : (C) an extent mask, (D) boundary mask , (E) a distance mask and (F) a field enumeration.

are selected representing a contrasting gradient across the EU. After downloading the dataset, a reprojection to EPSG 3035 was done. From the original set of 14.8 M parcels covering 376 K km^2 (Table 2), the 7831 4-km samples contains 2.5 M parcels covering 47,105 km^2 were selected. Finally, for both Sentinel-2 and ortophoto datasets, vector data was rasterised. The label is composed of four bands (example in Figure 6 B, C, D, E): extent mask, boundary mask, distance mask and field enumeration.

https://doi.org/10.5194/essd-2022-298
Preprint. Discussion started: 13 September 2022

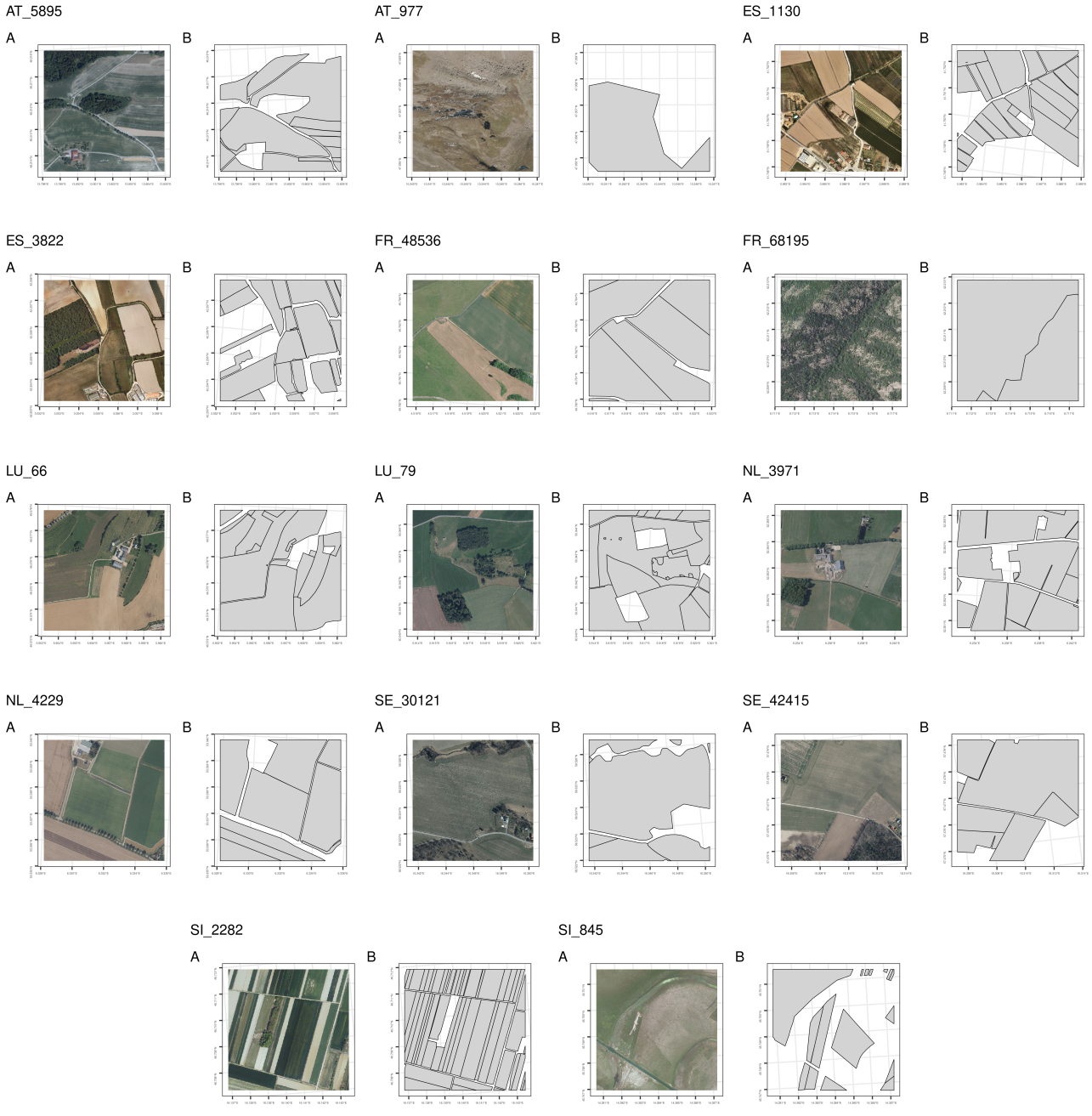

Figure 7. Two random aerial orthophotos (A) examples for each country along with the corresponding parcel vector labels (B). File id including NUTS0 (e.g. AT_5895) is above each example.

https://doi.org/10.5194/essd-2022-298
Preprint. Discussion started: 13 September 2022

Table 2. The original dataset contains 14.8 M parcels covering 376 K km^2. The stratified sampling resulting in 7,831 4-km samples contains 2.5 M parcels covering 47,105 km^2.

Region	Full set		Sampling	
	Count	*Area (km^2)*	*Count*	*Area (km^2)*
Austria	1,631,360	31,920.77	609,849	12,138.88
Catalonia	644,376	7,267.31	351,403	4,589.57
France	9,604,463	279,750.40	562,568	12,613.27
Luxembourg	92,397	1,280.16	76,657	1,044.46
Netherlands	772,565	18,686.18	399,849	8,277.49
Slovenia	820,151	4,684.92	249,271	1,522.60
Sweden	1,282,363	32,492.69	222,513	6,919.22
Total	**14,847,675**	**376,082.50**	**2,472,110**	**47,105.48**

2.4 Train, Validation and Test

In order to be used as a benchmark dataset, the stratification between the samples is also provided respecting typical distribution: training 70%, validation 15% and test 15%. As 233 sampling (almost exclusively in Sweden have no orthophoto available as described previously, the split was done on 7598 files (corresponding thus to 7,831 minus 233). The resulting random split provides 5,319 files for training, 1,140 for validation, 1,139 for testing and 233 as NA as they don't have aerial orthophoto. This information is stored in the column 'split' of the CSV tables with the URLs of the files.

3 Limitations and Perspectives

In this section, we point some limitations and potential improvements of the approach and the proposed dataset.

The atmospheric corrections and cloud screening remains a challenge for Sentinel-2. We implemented a pragmatic approach to improve the bottom-of-atmosphere reflectance obtained from sen2cor (Main-Knorn et al., 2017). The Hampel outlier detection approach followed by a Savitsky-Golay smoothing allows to produce a 5-day interpolated smoothed data. However, residual cloud, cloud-shadow or haze thus jeopardize the development of applications. From the interpolated data, we obtain median monthly composite to reduce the size of the data. This approach has also limitations and we could question the usefulness of interpolating the data if the ultimate goal is to produce median monthly composite as it could represent an extra computing burden wit a limited added value.

The access to the orthophoto services was done either via WMTS or via WMS. A specific server access has to be used for each countries with different projection (most in EPSG:3857 but some in local projections as shown in Table A1). While for most of the country a specific capability layer allows to select the specific year of the service, the specific data of acquisition is most of time unavailable. Additionally, the data quality is heterogeneous and depend of the specific acquisition.

https://doi.org/10.5194/essd-2022-298
Preprint. Discussion started: 13 September 2022

The labels are obtained from GSAA containing inherent caveats. First of all the geometry accuracy, referred as 1/5000 i.e. better than 1 m. Sometimes, parcels do not correspond to the agricultural field. Limitations of the label dataset could be the geometries, the timeliness and also the semantics. As agricultural fields might be missing (e.g. due to not being present in original GSAA data), the datasets are really only suitable for the masked approach in training - the models trained on AI4boundaries should only learn about the borders, extent and distance of the included fields.

Several potential improvements have to be considered in the future.

First, in addition to the field boundaries, the crop type could be added to enable semantic segmentation similarly to Sykas et al. (2022). To do it properly over a large scale, this would require to harmonise the legend of the GSAA from the different countries. A recent work (Schneider et al., 2021) has proposed a semantic harmonisation framework for this type data and could thus serve as basis.

Another potential improvement would be to add other data sources such as radar data (e.g. Sentinel-1 coherence) or high resolution satellite time series such as Planet data.

It is also very important to align such type of dataset with new emerging standard such as the one proposed by Radiant Earth ML HUB (https://mlhub.earth/, Alemohammad (2019)). Their Spatio Temporal Asset Catalog (STAC) helps to make geospatial assets openly searchable and indexable.

A limitation of the proposed AI4boundaries is its availability on a FTP server only, not directly callable as python packaged dataset Having the dataset accessible similarly to the Crop Harvest (Tseng et al., 2021) or to Calisto (https://github.com/Agri-Hub/Callisto-Dataset-Collection) would be more user friendly.

Finally, the AI4boundaries Sentinel-2 dataset has been used to train a model that is available on Euro Data Cube (EDC) as an algorithm for on-demand automatic delineation of agricultural field boundaries over user-defined Area Of Interest (AOI). The algorithm (https://collections.eurodatacube.com/field-delineation/) uses Sentinel-Hub services for accessing Sentinel-2 data, and can be employed to produce a baseline results for benchmarking purposes. In future work, the AI4boundaries could be made available directly on EDC along with a tutorial on how to use it to increase the outreach to the community of potential users.

4 Conclusions

The AI4boundaries dataset provides a statistical sampling of agricultural parcel boundaries over key regions of Europe along with 10-m Sentinel-2 satellite time series and 1-m aerial orthophoto imagery. This unique dataset allows to benchmark and compare parcel delineation methodologies in a transparent and reproducible way.

Data availability. This section describes each data-set provided along with this document and downloadable here https://jeodpp.jrc.ec.europa.eu/ftp/jrc-opendata/DRLL/AI4BOUNDARIES (d'Andrimont et al., 2022):

./sampling :

https://doi.org/10.5194/essd-2022-298
Preprint. Discussion started: 13 September 2022

ai4boundaries_sampling.gpkg is a geopackage vector file containing the 7831 4-by-4 km polygons of the sampling along with the stratification values as attributes;

ai4boundaries_ftp_urls_all.csv is a table that contains the path on the JRC FTP server of each Sentinel 2 tiles, orthophotos, and the respective labels of each. This also contains the split (i.e. train, test, val).

ai4boundaries_parcels_vector.gpkg consisting in a vector file (geopackage) with the original parcel boundaries on the 4-km grid cell of the sampling

./sentinel2 :

./images : the folder contains 7 folders - one for each NUTS0 region, amounting to a total of 7831 files named *NUTS0_sampleID_-S2_10m_256.nc*. The files are NetCDF of Sentinel 2 tiles at 10m ground resolution, of 256 by 256 pixels, and containing 5 bands (R, G, B, NIR and NDVI) from March to August 2019.

./masks : the folder contains 7 folders - one for each NUTS0 region, amounting to a total of 7831 files named *NUTS0_sampleID_-S2label_10m_256.tif*. The files are Geotiff at 10m ground resolution, of 256 by 256 pixels and containing 4 bands.

ai4boundaries_ftp_urls_sentinel2_split.csv contains the URLs of the sentinel2 image and corresponding mask files along with the split (i.e. train, test, val).

./orthophoto :

./images : the folder contains 7 folders - one for each NUTS0 region, amounting to a total of 7598 files named *NUTS0_sampleID_-ortho_1m_512.tif*. the files are Geotiff at 1m ground resolution, of 512 by 512 pixels, and containing 3 bands (R, G, B) acquired in 2019.

./masks : the folder contains 7 folders - one for each NUTS0 region, amounting to a total of 7598 files named *NUTS0_sampleID_-ortholabel_1m_512.tif*. the files are Geotiff at 1m ground resolution, of 512 by 512 pixels, and containing 4 bands.

ai4boundaries_ftp_urls_orthophoto_split.csv contains the URLs of the orthophoto image and corresponding mask files along with the split (i.e. train, test, val).

Appendix A

Table A1. Aerial orhtophoto WMTS and WMS services and projections.

country	epsg	capablity layer	type
Netherlands	28992	"https://service.pdok.nl/hwh/luchtfotorgb/wmts/v1_0?REQUEST=GetCapabilities,layer=2019_ortho25"	wmts
Luxembourg	3857	"http://wmts1.geoportail.lu/opendata/wmts/1.0.0/WMTSCapabilities.xml,layer=ortho_2019"	wmts
Austria	3857	"https://maps.wien.gv.at/basemap/1.0.0/WMTSCapabilities.xml,layer=bmaporthofoto30cm	wmts
Catalonia	3857	"https://geoserveis.icgc.cat/icc_mapesmultibase/noutm/wmts/topo/1.0.0/WMTSCapabilities.xml,layer=orto"	wmts
Slovenia	3794	"https://prostor4.gov.si/ows2-gwc-pub/service/wmts?request=GetCapabilities,layer=SI.GURS.ZPDZ:DOF050"	wmts
France	3857	"https://wxs.ign.fr/8ir1y6t0lrcpvtp6up6vc3h7/geoportail/wmts?SERVICE=WMTS&REQUEST=GetCapabilities, layer=ORTHOIMAGERY.ORTHOPHOTOS"	wmts
Sweden	3857	"https://minkarta.lantmateriet.se/map/ortofoto/?SERVICE=WMS&VERSION=1.1.1&LAYERS=Ortofoto_0.25"	wms

https://doi.org/10.5194/essd-2022-298
Preprint. Discussion started: 13 September 2022

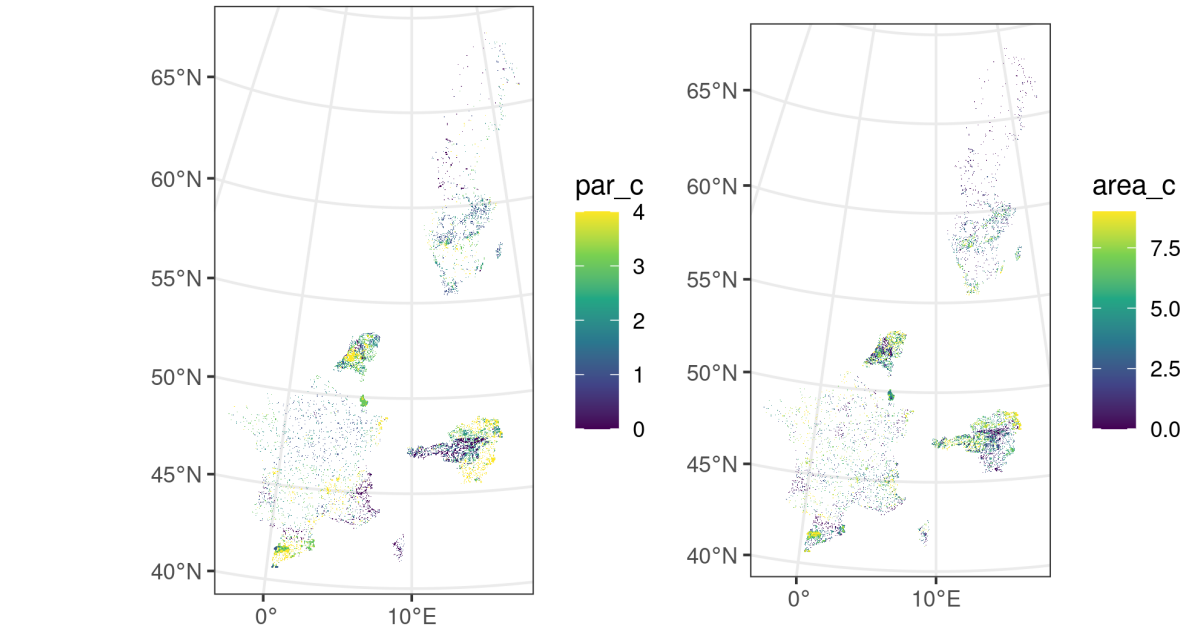

Figure A1. Distribution of sampling units among the seven regions with the two variables used for the stratification.

Author contributions. R.D. and F.W. conceptualized the study and designed the methodology. R. D., M. C., P. K., M. Y., D. P. and F.W. processed the data. R. D., M. C., P. K., D. M., M.Y., D. P. and F.W. analyzed the data and wrote the paper.

Competing interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

Acknowledgements. The authors thank Ferdinando Urbano, Guido Lemoine, Marijn van der velde and Wim Devos for their support.

https://doi.org/10.5194/essd-2022-298
Preprint. Discussion started: 13 September 2022

References

Alemohammad, H.: Radiant ML Hub, https://www.radiant.earth/mlhub/, 2019.

Aung, H. L., Uzkent, B., Burke, M., Lobell, D., and Ermon, S.: Farm parcel delineation using spatio-temporal convolutional networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 76–77, 2020.

Chen, J., Jönsson, P., Tamura, M., Gu, Z., Matsushita, B., and Eklundh, L.: A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky–Golay filter, Remote Sensing of Environment, 91, 332–344, https://doi.org/https://doi.org/10.1016/j.rse.2004.03.014, 2004.

d'Andrimont, R., Claverie, M., Kempeneers, P., Muraro, D., Martinez Sanchez, L., and Waldner, F.: AI4boundaries, 2022.

Garcia-Pedrero, A., Lillo-Saavedra, M., Rodriguez-Esparragon, D., and Gonzalo-Martin, C.: Deep learning for automatic outlining agricultural parcels: Exploiting the land parcel identification system, IEEE Access, 7, 158 223–158 236, 2019.

Hampel, F. R.: The influence curve and its role in robust estimation, Journal of the american statistical association, 69, 383–393, 1974.

Helber, P., Bischke, B., Dengel, A., and Borth, D.: EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 12, 2217–2226, 2019.

Kempeneers, P., Pesek, O., De Marchi, D., and Soille, P.: pyjeo: A Python Package for the Analysis of Geospatial Data, ISPRS International Journal of Geo-Information, 8, 461, https://doi.org/10.3390/ijgi8100461, 2019.

Main-Knorn, M., Pflug, B., Louis, J., Debaecker, V., Müller-Wilm, U., and Gascon, F.: Sen2Cor for sentinel-2, in: Image and Signal Processing for Remote Sensing XXIII, vol. 10427, pp. 37–48, SPIE, 2017.

Masoud, K. M., Persello, C., and Tolpekin, V. A.: Delineation of agricultural field boundaries from Sentinel-2 images using a novel super-resolution contour detector based on fully convolutional networks, Remote sensing, 12, 59, 2019.

Rußwurm, M., Pelletier, C., Zollner, M., Lefèvre, S., and Körner, M.: Breizhcrops: A time series dataset for crop type mapping, arXiv preprint arXiv:1905.11893, 2019.

Schneider, M., Broszeit, A., and Körner, M.: Eurocrops: A pan-european dataset for time series crop type classification, arXiv preprint arXiv:2106.08151, 2021.

Soille, P., Burger, A., De Marchi, D., Kempeneers, P., Rodriguez, D., Syrris, V., and Vasilev, V.: A versatile data-intensive computing platform for information retrieval from big geospatial data, Future Generation Computer Systems, 81, 30–40, 2018.

Sumbul, G., Charfuelan, M., Demir, B., and Markl, V.: Bigearthnet: A large-scale benchmark archive for remote sensing image understanding, in: IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium, pp. 5901–5904, IEEE, 2019.

Sykas, D., Sdraka, M., Zografakis, D., and Papoutsis, I.: A Sentinel-2 multi-year, multi-country benchmark dataset for crop classification and segmentation with deep learning, https://doi.org/10.48550/ARXIV.2204.00951, 2022.

Tarasiou, M., Guler, R. A., and Zafeiriou, S.: Context-self contrastive pretraining for crop type semantic segmentation, arXiv preprint arXiv:2104.04310, 2021.

Tseng, G., Zvonkov, I., Nakalembe, C. L., and Kerner, H.: CropHarvest: A global dataset for crop-type classification, in: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.

Waldner, F. and Diakogiannis, F. I.: Deep learning on edge: Extracting field boundaries from satellite images with a convolutional neural network, Remote Sensing of Environment, 245, 111 741, 2020.

Waldner, F., Diakogiannis, F. I., Batchelor, K., Ciccotosto-Camp, M., Cooper-Williams, E., Herrmann, C., Mata, G., and Toovey, A.: Detect, consolidate, delineate: Scalable mapping of field boundaries using satellite images, Remote Sensing, 13, 2197, 2021.

https://doi.org/10.5194/essd-2022-298
Preprint. Discussion started: 13 September 2022

Open Access Earth System Science Data Discussions

Wang, S., Waldner, F., and Lobell, D. B.: Unlocking large-scale crop field delineation in smallholder farming systems with transfer learning and weak supervision, arXiv preprint arXiv:2201.04771, 2022.