Open Data Science

WORKSHOP

13. - 16. June 2022

Prague, Czech Republic

Co-financed by the Connecting Europe Facility of the European Union

CTU CZECH TECHNICAL UNIVERSITY IN PRAGUE

mundialis

MultiOne

OpenGeoHUB

TERRASIGNA

Introduction to ODSE datasets in Python

Jun 13, 2021: 15:30 - 17:00

https://bit.ly/3MHRRLX

Tomáš Bouček

tomas.boucek@fsv.cvut.cz

https://www.fsv.cvut.cz/

Leandro Parente

leandro.parente@opengeohub.org

https://opengeohub.org

Introduction to ODSE datasets in Python - Outline

- Open Data Science Europe
- Available datasets
- Cloud-free Landsat ARD (2000–2020)
- Annual land use and land cover (2000–2020)
- Multi depth soil organic carbon, pH, nutrients, clay and sand content (2000–2020)
- Distributed processing

Open Data Science Europe

The Open Data Science Europe data portal / viewer aims at serving decision-ready layers such land cover, air quality and pollution, potential natural vegetation and similar.

INEA ceased operations on 31 March 2021. The **European Health and Digital Executive Agency (HaDEA)** *was established on 1 April 2021 to take over the CEF Telecom legacy portfolio as well as additional EU funding programmes.*

Geo-harmonizer: EU-wide automated mapping system for harmonization of Open Data based on FOSS4G and Machine Learning

Programme:
CEF Telecom

2018-EU-IA-0095

Call year:
2018

Location of the Action:
Croatia, Czech Republic, Germany, Netherlands, Romania

Implementation schedule:
September 2019 to June 2022

https://ec.europa.eu/inea/en/connecting-europe-facility/cef-telecom/2018-eu-ia-0095

Open Data Science Europe

The Open Data Science Europe data portal / viewer aims at serving decision-ready layers such land cover, air quality and pollution, potential natural vegetation and similar.

Open Source software

Publicly available data

Open Data Science Europe

The Open Data Science Europe data portal / viewer aims at serving decision-ready layers such land cover, air quality and pollution, potential natural vegetation and similar.

Decision-ready layers

Spatiotemporal Machine Learning Algorithms

Data Harmonization

Open Source software

Publicly available data

Open Data Science Europe

The Open Data Science Europe data portal / viewer aims at serving decision-ready layers such land cover, air quality and pollution, potential natural vegetation and similar.

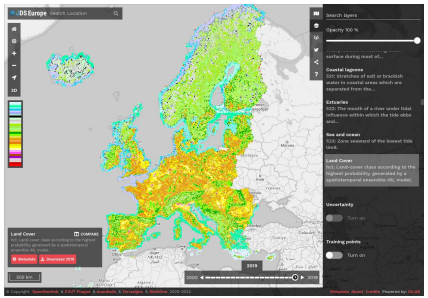
Decision-ready layers

Spatiotemporal Machine Learning Algorithms

Data Harmonization

Open Source software

Publicly available data



https://maps.opendatascience.eu

GeoNetwork opensource

https://data.opendatascience.eu/geonetwork

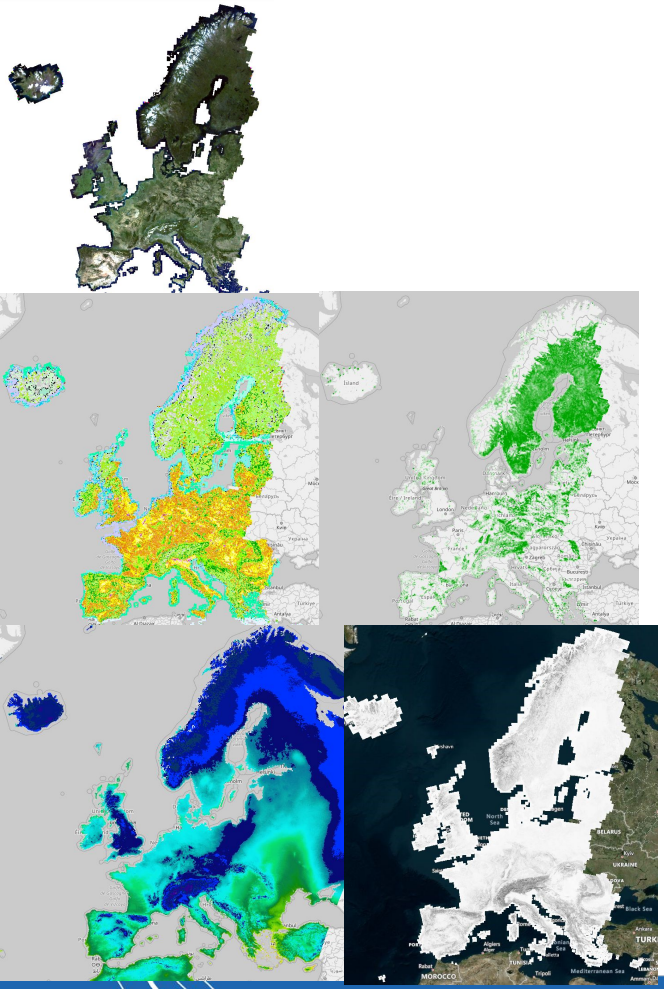
STAC SpatioTemporal Asset Catalog

http://stac.opendatascience.eu

GitLab

https://gitlab.com/geoharmonizer_inea

Open Data Science



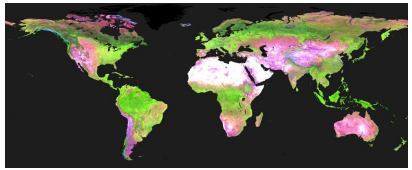
Available datasets

- Cloud-free GLAD **Landsat** ARD (2000–2020)
- Cloud-free **Sentinel** ARD (2018–2020)
- Annual **land cover** with dominant class, probabilities and uncertainties (2000–2020)
- Digital **terrain** model
- Daily **precipitation** and **air / land** temperature (2000–2020)
- Potential and realized **forest tree species** distribution (2000–2020)
- Actual and potential natural vegetation (2000–2020),
- Multi depth **soil organic carbon**, **pH**, nutrients, **clay** and **sand content** (2000–2020),
- **Aerosol** Optical Depth-AOD at 550 nm (2018-2020),
- Harmonized **OpenStreetMap** (OSM).

Cloud-free Landsat ARD (2000–2020)

GLAD Landsat ARD

- Globally consistent analysis ready data (ARD) for multi-decadal LCLU monitoring
- 16-day time-series composites from Landsat 5, 7 and 8 (TM, ETM+ and OLI)
- Per-pixel observation quality flag
- MODIS (MOD44C) surface reflectance calibrated
- Product organized by 1 × 1 degree tiles
- Automatically download through HTTP API
- Product under Creative Commons Attribution License



FORCE

- An all-in-one remote sensing processing framework for Sentinel-2 A/B MSI and Landsat 5, 7 and 8 (TM, ETM+ and OLI)
- Advanced cloud and cloud shadow detection
- Integrated atmospheric, topographic and BRDF correction
- Reprojection and grinding capabilities
- Different strategies to generate composites (e.g. best available pixel, spectral temporal metrics)
- Free software under GNU License v.3

FORCE
Framework for Operational Radiometric Correction for Environmental Monitoring

Cloud-free Landsat ARD (2000–2020)

FORCE
(All levels)

GLAD Landsat ARD
(Level 3 product)

esa USGS science for a changing world
(Level 1 and on-demand level 2 products)

Level 4
(Model output)

Level 3
(Temporal composites and gridded data)

Level 2
(Atmospheric correction)

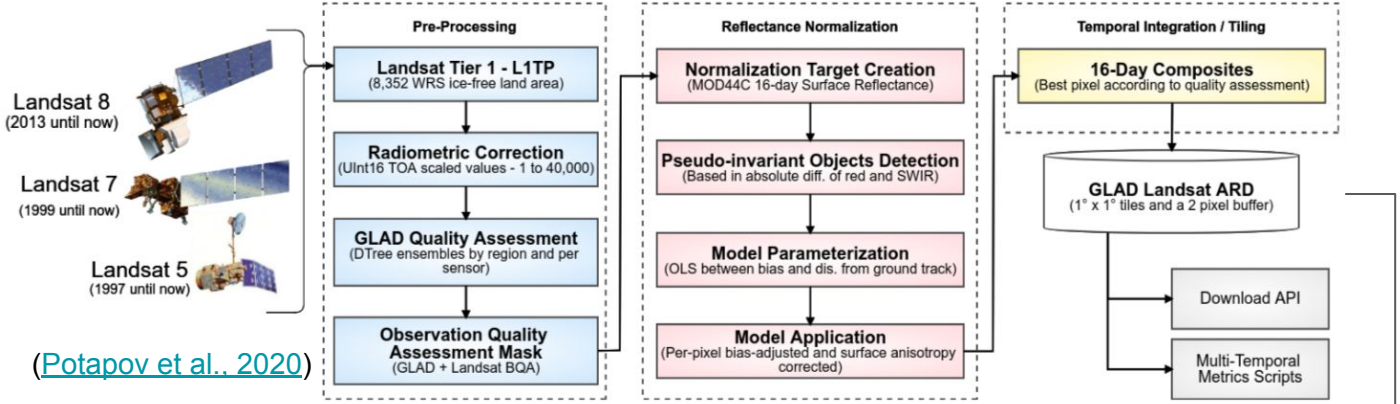
Level 1
(Radiometrically calibrated and georectified data)

GeoHarmonizer

- 4 mosaics per year (Level 3)
- LULC Maps (Level 4)
- Environmental Quality Maps (Level 4)

Asrar, G.; Greenstone, R. (Eds.) MTPE EOS Reference Handbook; NASA/Goddard Space Flight Center: Greenbelt, MD, USA, 1995; p. 281.

Cloud-free Landsat ARD (2000–2020)



(Potapov et al., 2020)

Aggregate 16-days composites in 4 time periods per year

Interval ID	Start	End	Composite
1	1-Jan	16-Jan	
2	17-Jan	1-Feb	
3	2-Feb	17-Feb	1
4	18-Feb	4-Mar	
5	5-Mar	20-Mar	
6	21-Mar	5-Apr	
7	6-Apr	21-Apr	
8	22-Apr	7-May	2
9	8-May	23-May	
10	24-May	8-Jun	
11	9-Jun	24-Jun	
12	25-Jun	10-Jul	
13	11-Jul	26-Jul	
14	27-Jul	11-Aug	
15	12-Aug	27-Aug	3
16	28-Aug	12-Sep	
17	13-Sep	28-Sep	
18	29-Sep	14-Oct	
19	15-Oct	30-Oct	
20	31-Oct	15-Nov	
21	16-Nov	1-Dec	4
22	2-Dec	17-Dec	
23	18-Dec	31-Dec	

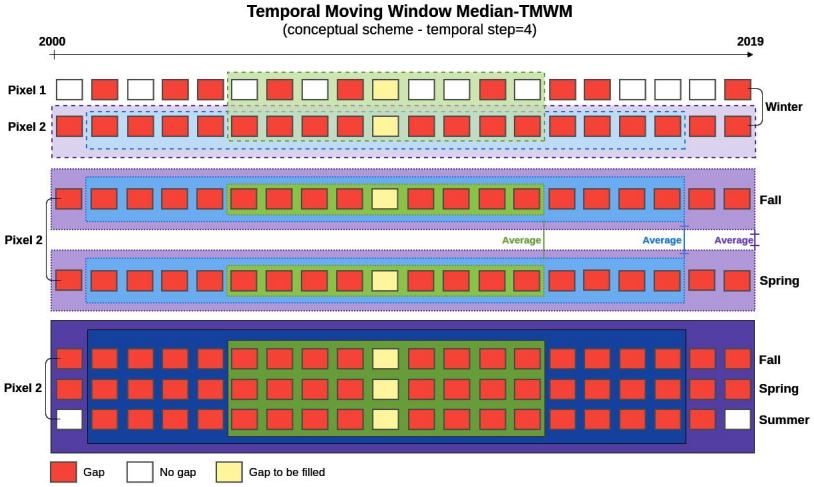
Percentile 25th

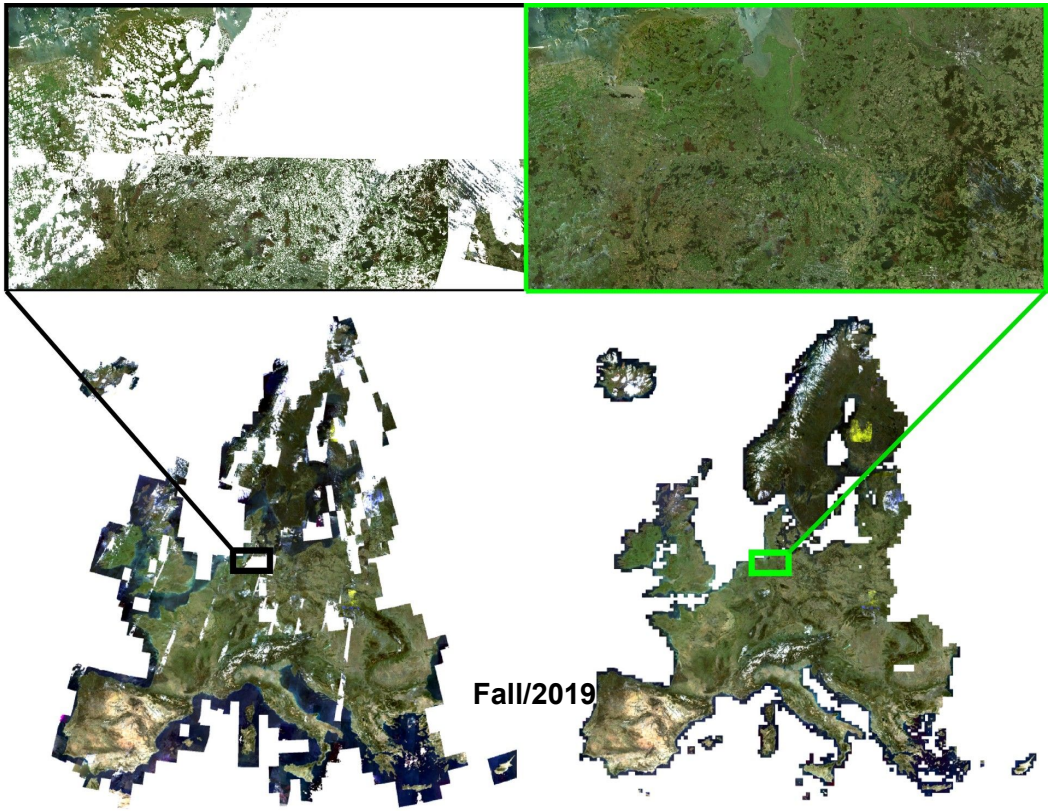
Percentile 50th

Percentile 75th

Gap filling
(temporal moving window median)

Cloud-free Landsat ARD (2000–2020)





Total amount of yearly images: **84**

(7 bands x 3 percentiles x 4 trimesters)

Cloud-free Landsat ARD (2000–2020)

04_introduction_odse_datasets.ipynb

File Edit View Insert Runtime Tools Help Saving...

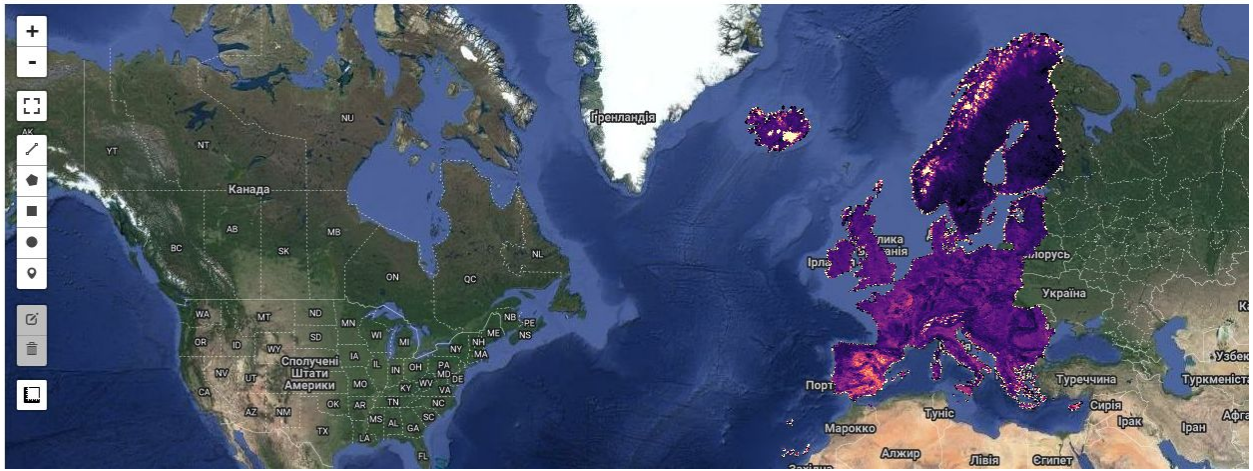
+ Code + Text

Cloud-free Landsat ARD (2000–2020)

The first dataset that you will access is the quarterly green band of GLAD Landsat ARD (2000–2020). Use this link to choose a specific period and band (e.g. `green_p50`) in ODSE STAC Browser, copying the download link to the variable `url`.

The `display_cog` function accepts a palette and basemap arguments.

```
url = 'https://s3.eu-central-1.wasabisys.com/eumap/lcv/lcv_green_landsat.glad.ard_p50_30m_0..0cm_2019.06.25..2019.09.12_eumap_epsg3035_v1.1.tif'
display_cog(url)
```

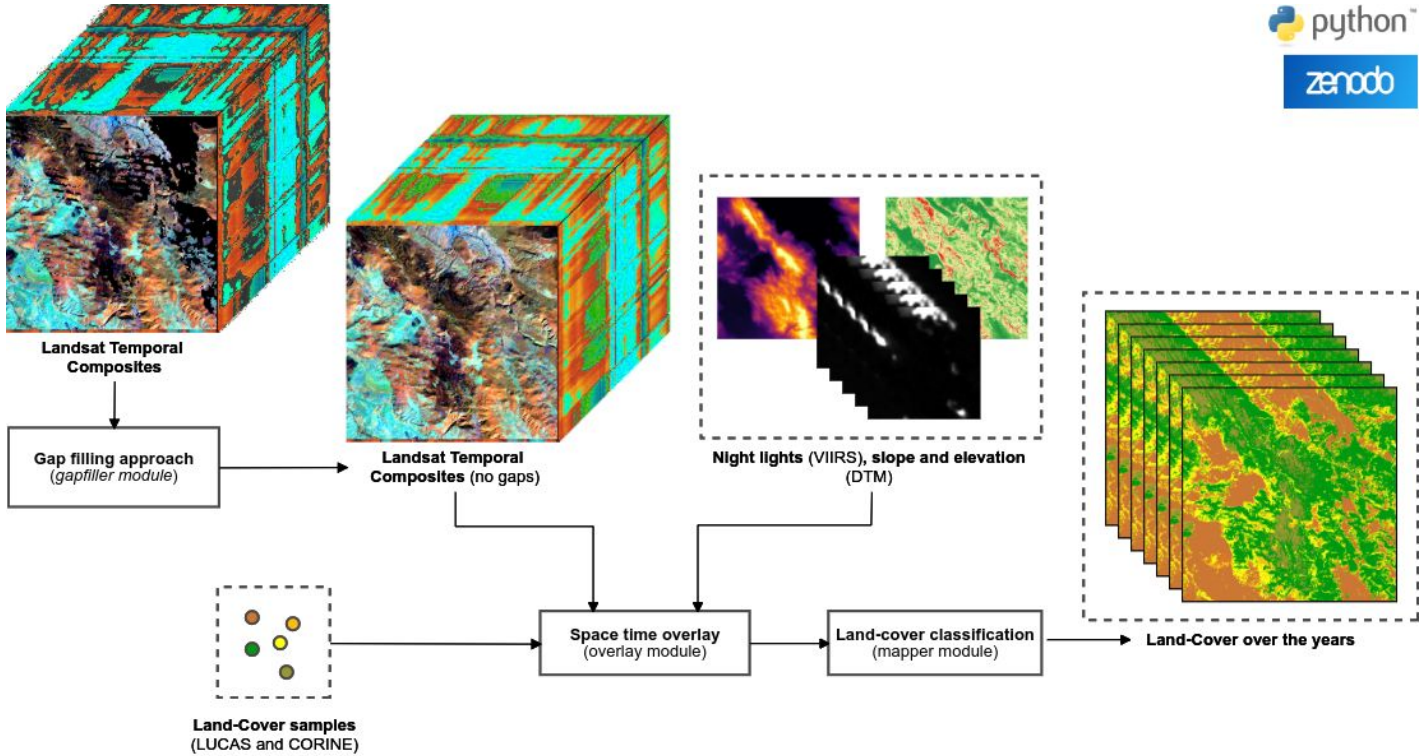


Google colab

Colab link

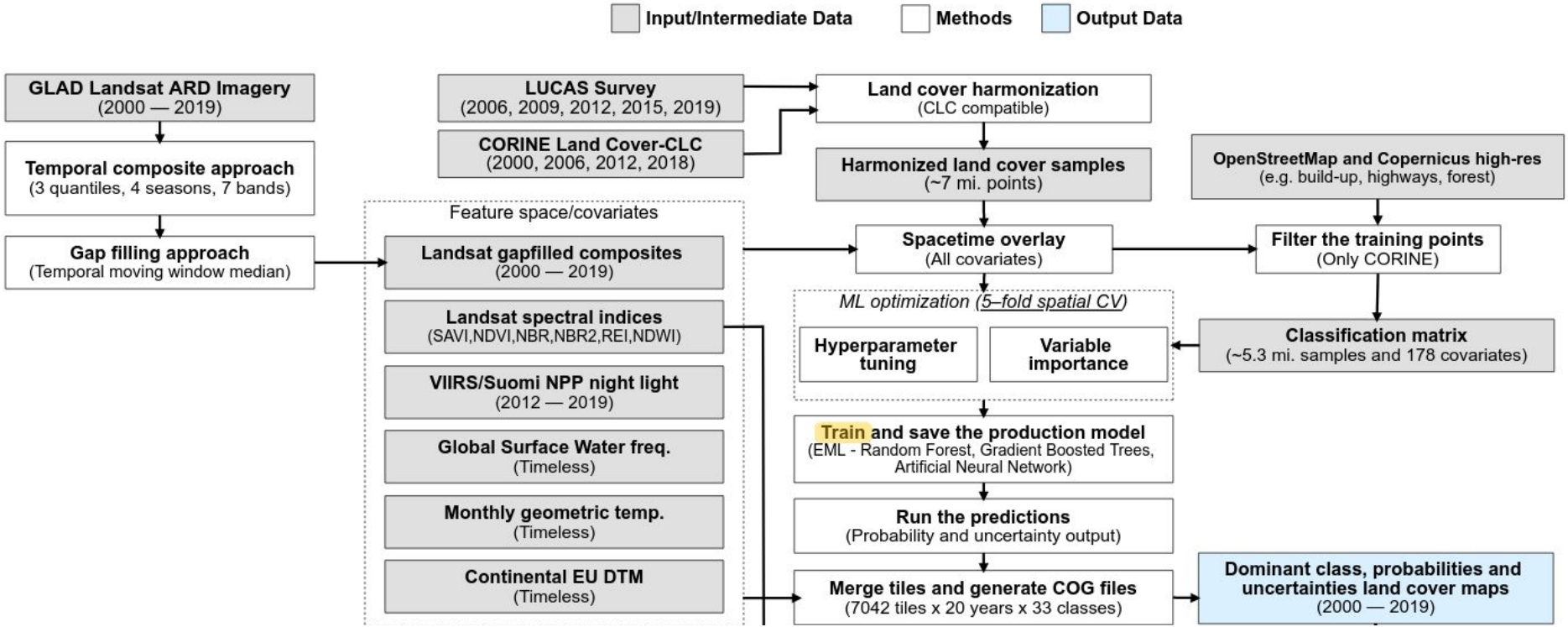
Open Data Science

Annual land cover with dominant class, probabilities and uncertainties (2000–2020)



https://doi.org/10.21203/rs.3.rs-561383/v4

Annual land cover with dominant class, probabilities and uncertainties (2000–2020)

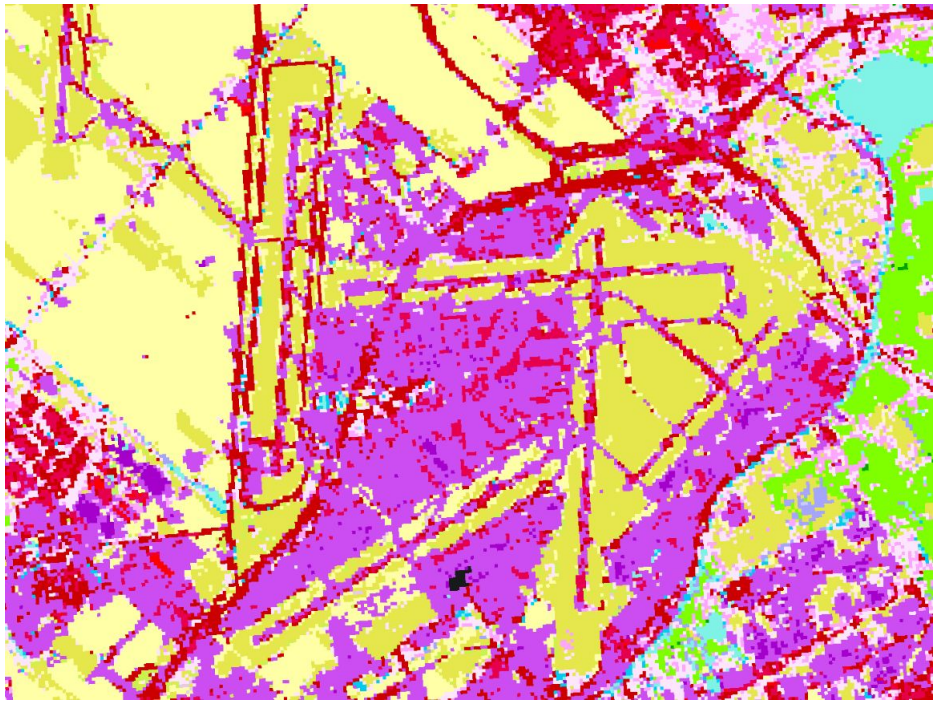


Annual land cover with dominant class, probabilities and uncertainties

Schiphol Airport, NL

GLAD Landsat ARD - RGB

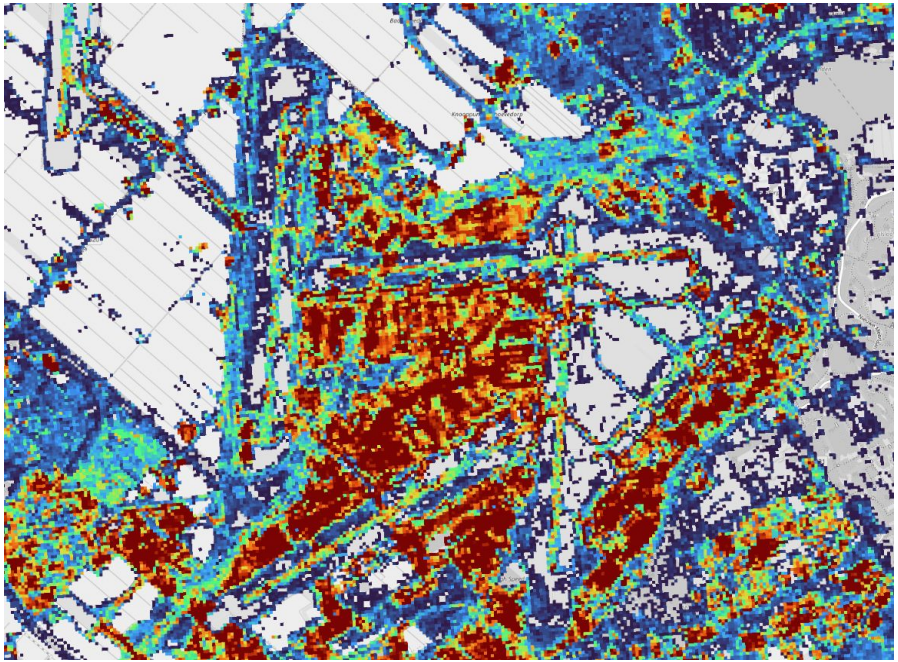
Predicted most likely class



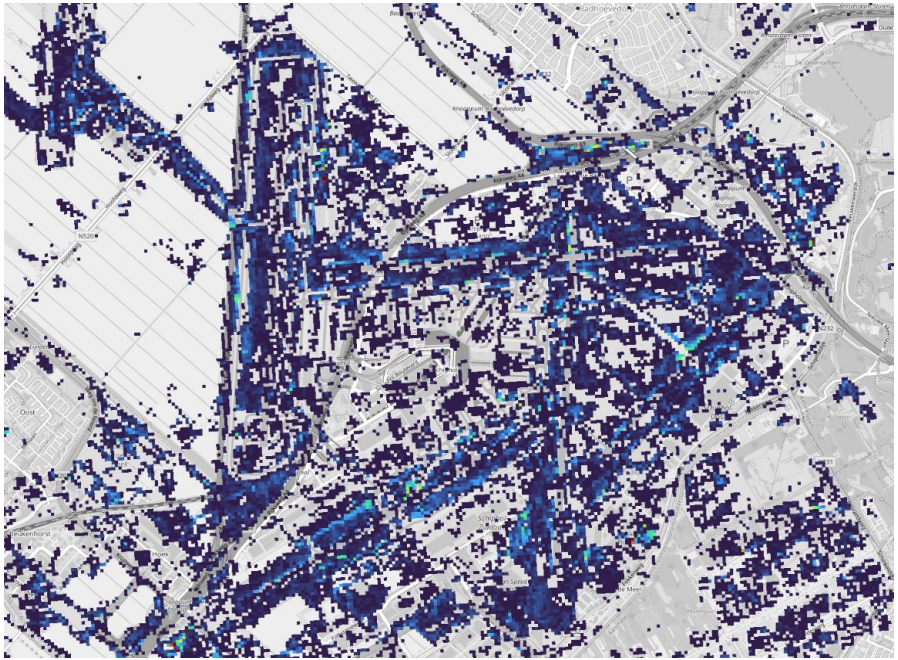
Annual land cover with dominant class, probabilities and uncertainties

Schiphol Airport, NL

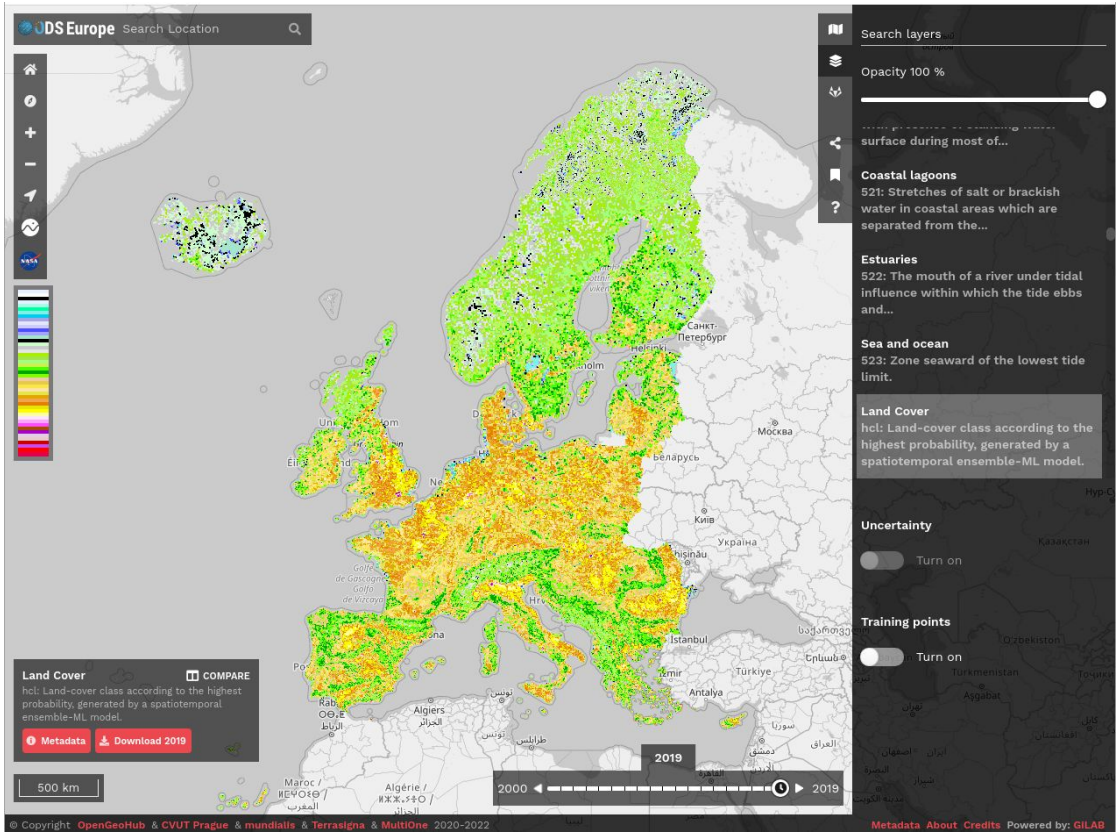
Industrial & Commercial buildings



….Airports

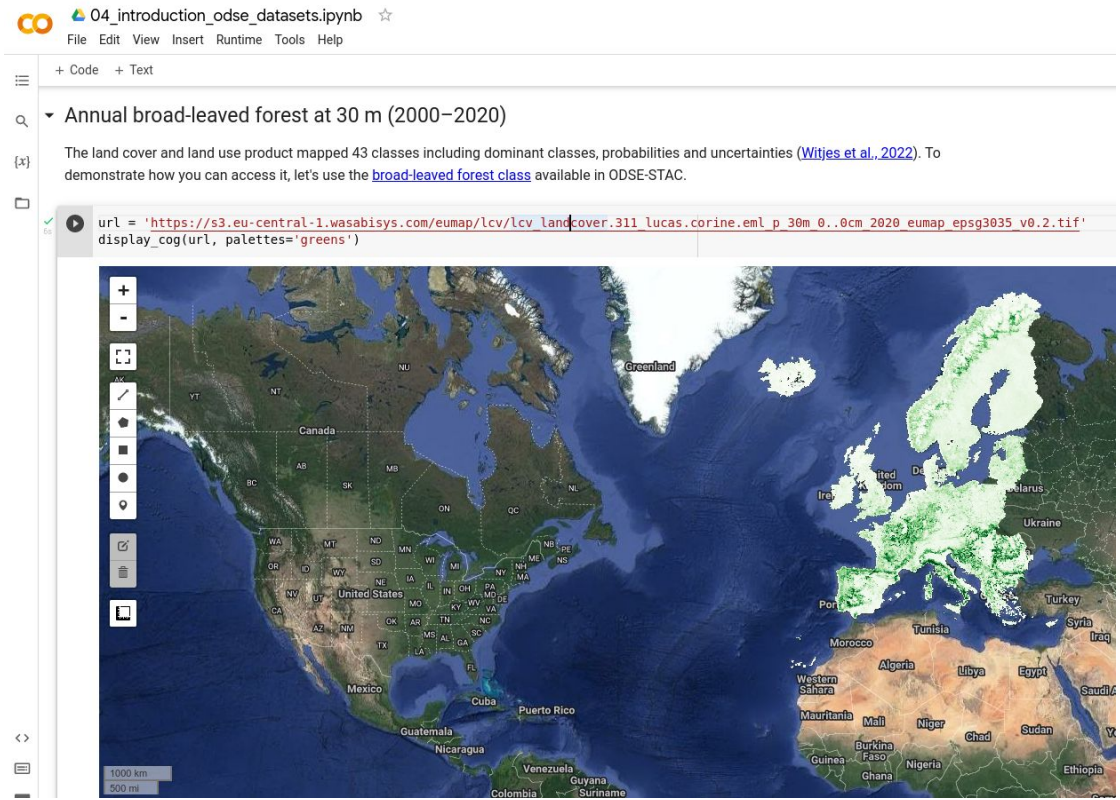


Annual land cover with dominant class, probabilities and uncertainties



ODSE Viewer Link
https://maps.opendatascience.eu

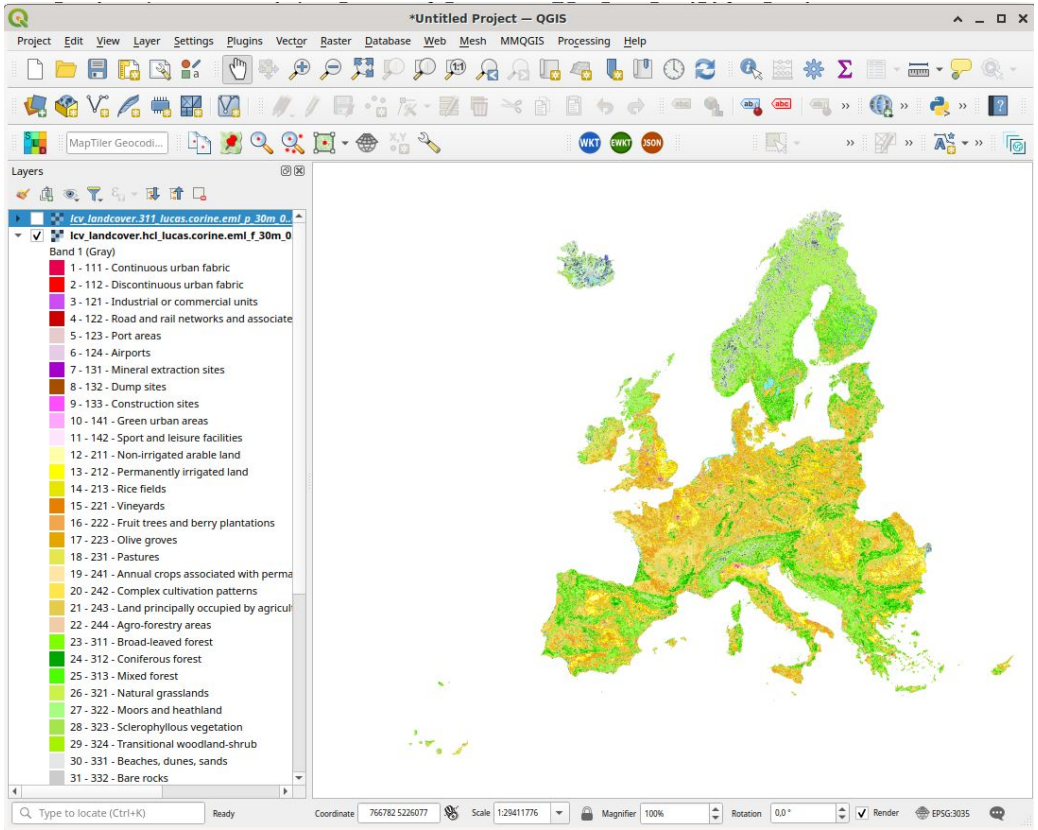
Annual land cover with dominant class, probabilities and uncertainties



Google colab

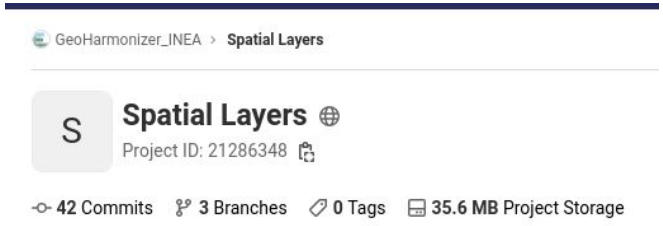
Colab link

Annual land cover with dominant class, probabilities and uncertainties



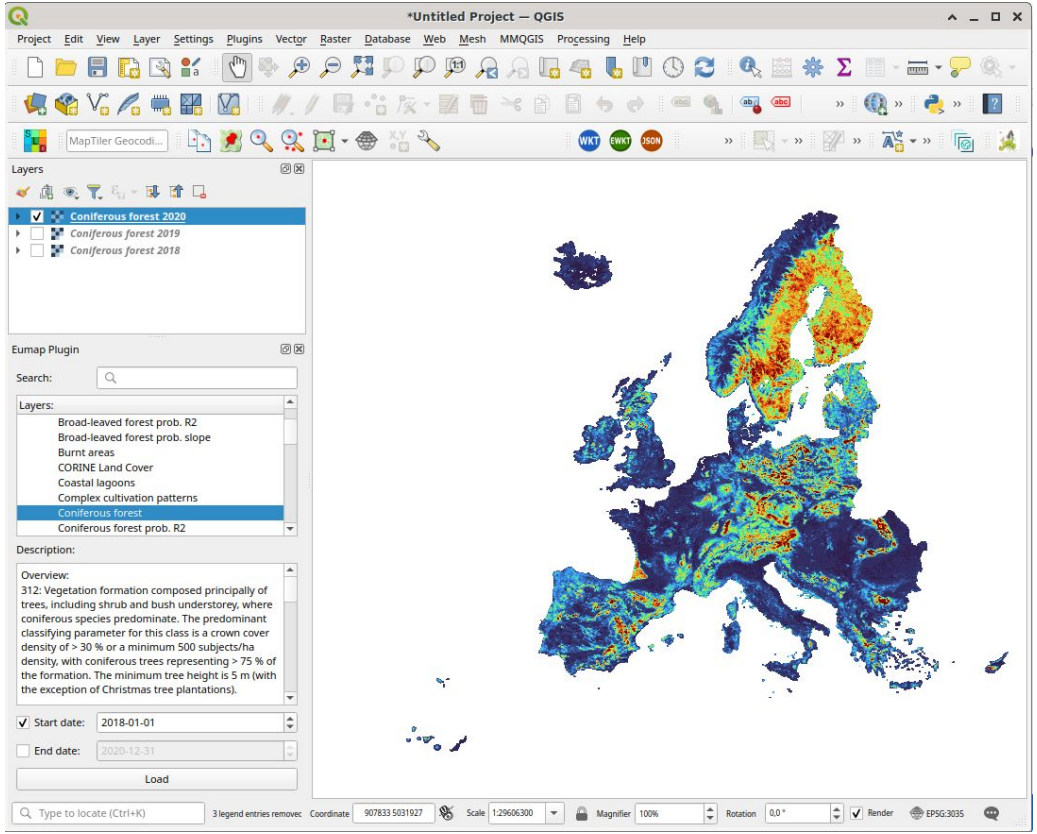
QML Dominant classes

QML Probabilities classes



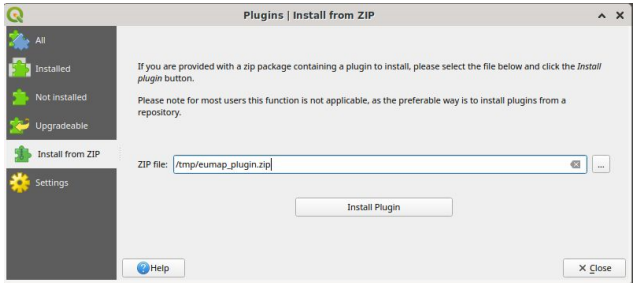
https://gitlab.com/geoharmonizer_inea/spatial-layers

Annual land cover with dominant class, probabilities and uncertainties



QGIS

EUMAP Plugin

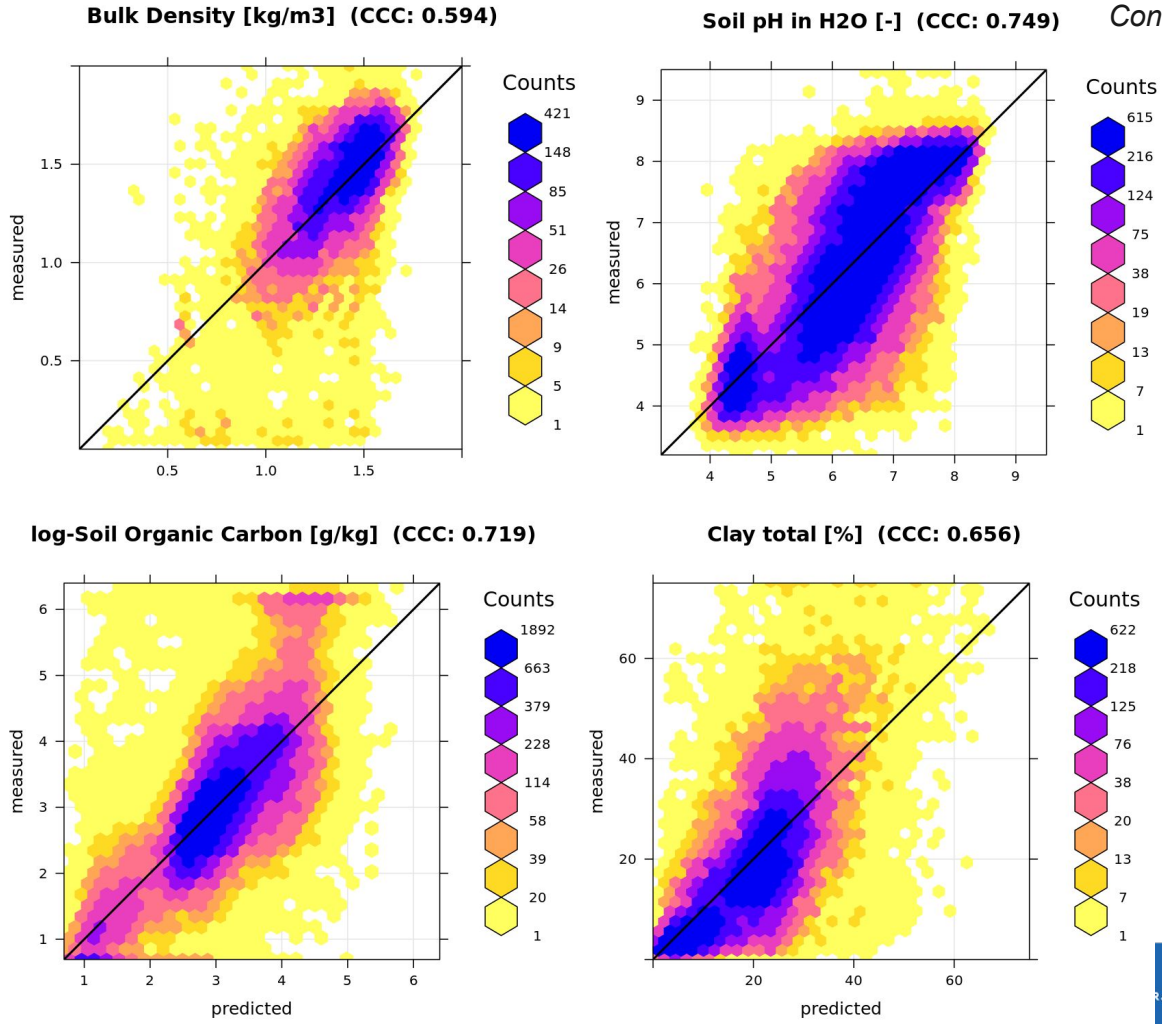


Open Data Science

Multi depth soil organic carbon, pH, nutrients, clay and sand content (2000–2020)

Legacy soil samples & profiles for EU

Open Data Science



Concordance correlation coefficient

Results based on the 5-fold cross-validation (with refitting) and by using 30×30 km blocks to prevent any overfitting

Multi depth soil organic carbon, pH, nutrients, clay and sand content (2000–2020)

```
Variable: log.oc
R-square: 0.562
Fitted values sd: 0.802
RMSE: 0.709

EML model summary:
Call:
stats::lm(formula = f, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2361 -0.3868 -0.0488  0.3169  5.0829

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.17625    0.01046 -16.858   <2e-16 ***
regr.ranger    0.79829    0.01236  64.600   <2e-16 ***
regr.xgboost   0.27624    0.01314  21.021   <2e-16 ***
regr.cvglmnet -0.01951    0.00840  -2.322   0.0202 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7085 on 86986 degrees of freedom
Multiple R-squared:  0.5618,    Adjusted R-squared:  0.5617
F-statistic: 3.717e+04 on 3 and 86986 DF,  p-value: < 2.2e-16
```

N = 87,000 samples

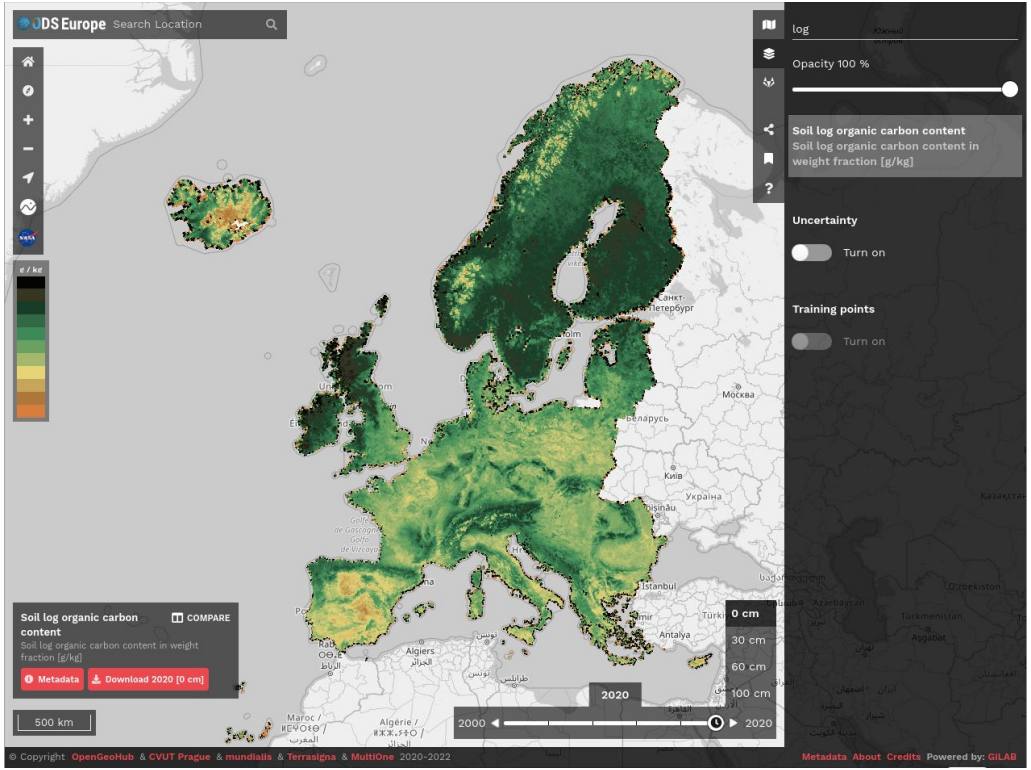
RMSE = 0.70

Example:

Predicted SOC = 1.5%
Lower (1 std) PI = 0.7%
Upper (1 std) PI = 3.1%

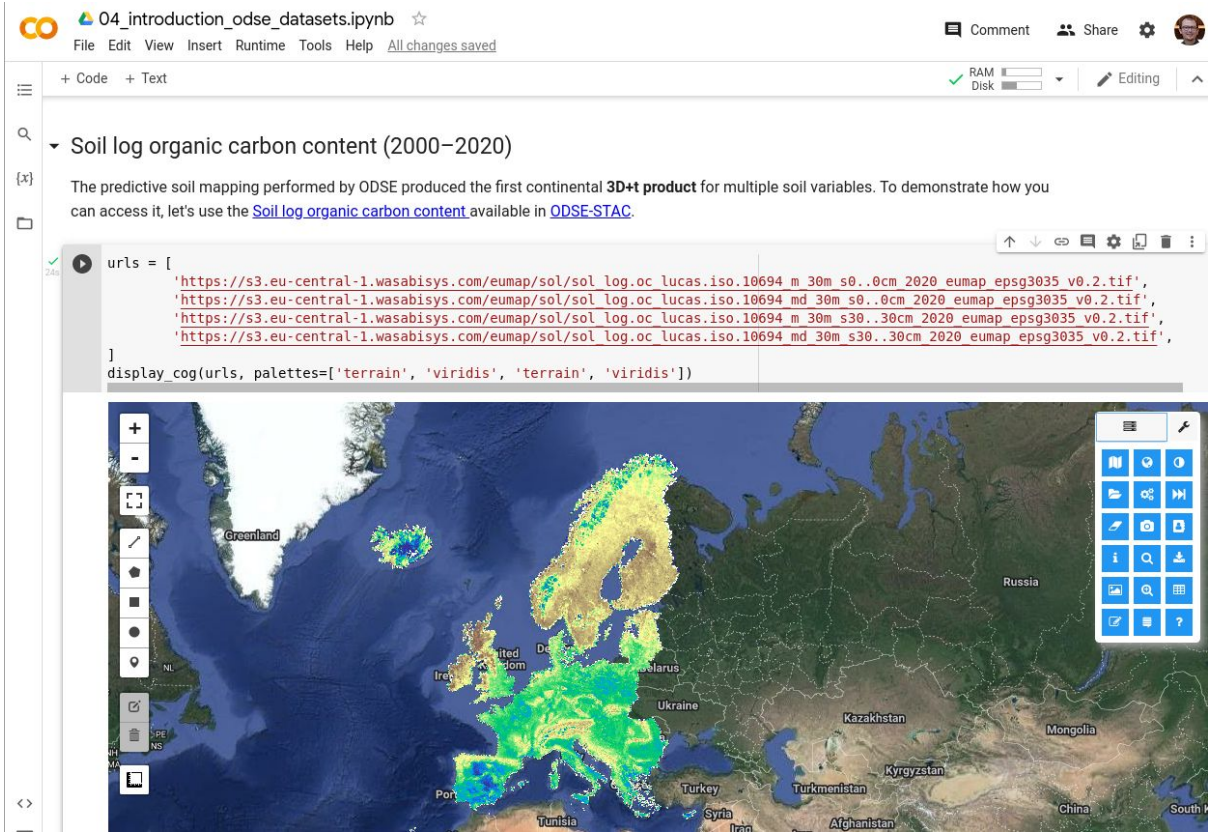
Predicted SOC = 4%
Lower (1 std) PI = 2.0%
Upper (1 std) PI = 8.1%

Multi depth soil organic carbon, pH, nutrients, clay and sand content (2000–2020)



ODSE Viewer Link
https://maps.opendatascience.eu

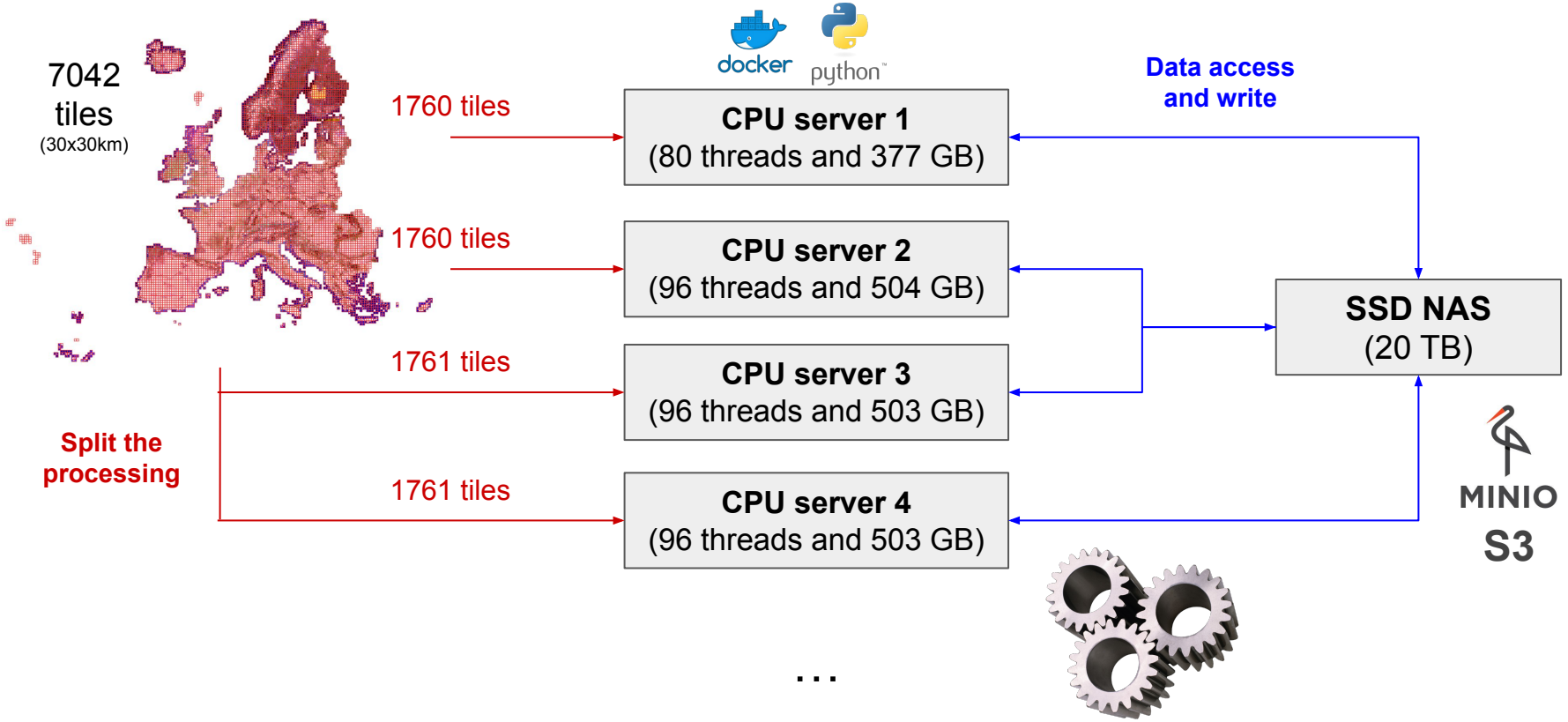
Annual land cover with dominant class, probabilities and uncertainties

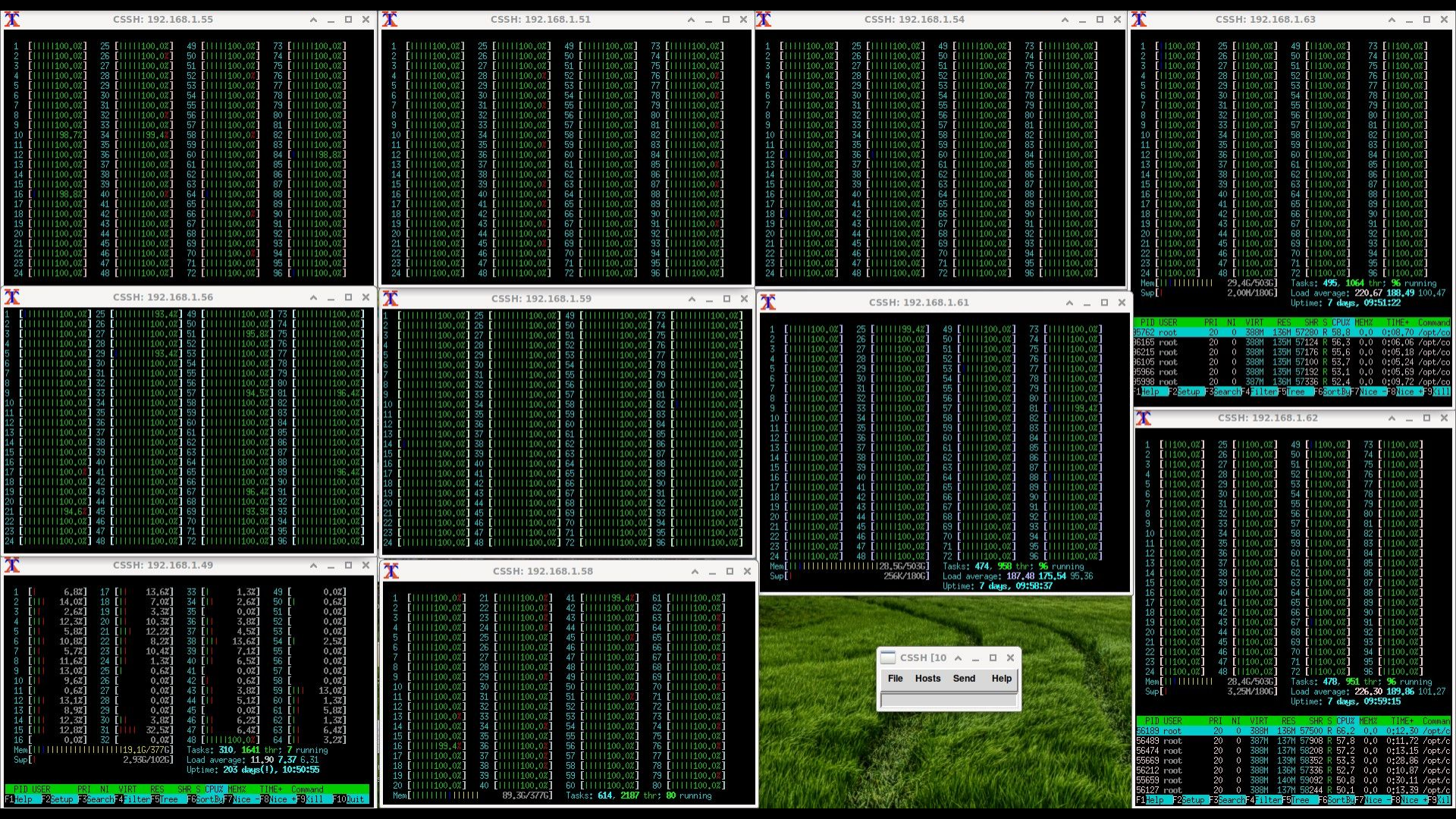


Google colab

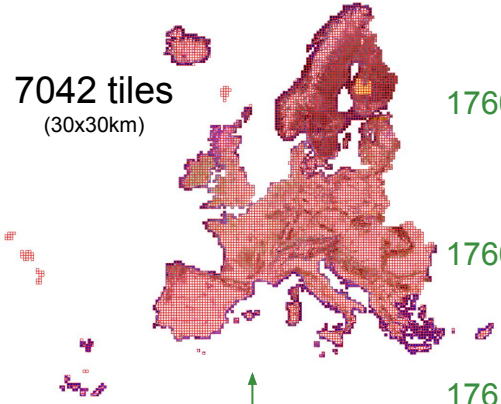
[Colab link]

Distributed processing




CSSH: 192.168.1.55
CSSH: 192.168.1.51
CSSH: 192.168.1.54
CSSH: 192.168.1.63
CSSH: 192.168.1.56
CSSH: 192.168.1.59
CSSH: 192.168.1.61
CSSH: 192.168.1.62
CSSH: 192.168.1.49
CSSH: 192.168.1.58
CSSH [10
File
Hosts
Send
Help

Distributed processing



7042 tiles
(30x30km)

1760 tiles

1760 tiles

1761 tiles

1761 tiles

docker python

CPU server 1
(80 threads and 377 GB)

CPU server 2
(96 threads and 504 GB)

CPU server 3
(96 threads and 503 GB)

CPU server 4
(96 threads and 503 GB)

Data access

SSD NAS
(20 TB)

MINIO
S3

Aggregate the results

GDAL

Cloud optimized tiff generation

gdalbuildvrt
gdal_translate

Open Data Science

WORKSHOP

13. - 16. June 2022

Prague, Czech Republic

Co-financed by the Connecting Europe Facility of the European Union

CTU CZECH TECHNICAL UNIVERSITY IN PRAGUE

mundialis

MultiOne

OpenGeoHUB

TERRASIGNA