INFORSAT: AN ONLINE SENTINEL-2 MULTI-TEMPORAL ANALYSIS TOOL SET USING R CRAN

F. Pirotti[1,2*], E. Kutchartt[1], M. Piragnolo[1]

[1] Department of Land, Environment, Agriculture and Forestry (TESAF), University of Padova, Legnaro, Italy – (francesco.pirotti, erico.kutchartt,marco.piragnolo)@unipd.it
[2] Interdepartmental Research Center of Geomatics (CIRGEO), University of Padova, Legnaro, Italy

Commission III, WG III/10

KEY WORDS: Copernicus, Sentinel-2, earth observation, multitemporal analysis, analytics, web-gis, parallel processing.

ABSTRACT:

Remote sensing via orbiting satellite sensors is today a common tool to monitor numerous aspects related to the Earth surface and the atmosphere. The amount of data from imagery have increased tremendously since the past years, due to the increase in space missions and public and private agencies involved in this activity. A lot of these data are open-data, and academics and stakeholders in general can freely download and use it for any type of application. The bottle-neck is often not data availability anymore, but the processing resources and tools to analyse it. In particular multi-temporal analysis requires stacks of images thus digital space for storage and processing workflows that are tested and validated. Processing image by image is often not a viable approach anymore. Basic tools for multi-temporal analysis are provided via the same web interface, allowing the user to also apply parallel processing for a faster data extraction. A study case over burned areas in the north-eastern region of Italy are reported, to show how the multi-temporal analysis tools provided can be a valid source of data for further analysis. Multitemporal data consisting on the index values of each pixel inside user-defined areas can be downloaded in a spreadsheet that provides the values, the cell ids, the timestamp and the cloud and snow percentage. Also the full-resolution raster with index values that are rendered on screen can be downloaded as GeoTIFF at each specific date.

1. INTRODUCTION

Satellite images are data cubes with large volume with respect to classical images, usually several thousands of rows and columns. They can be provided with different formats, with GeoTIFF being commonly used for simple structures, and NetCDF-derived models like HDf5 for integrating more complex data in a single file. Two of the most commonly used sources of image data are from NASA's Landsat's missions and from the Sentinel-2 missions from the European Space Agency (ESA). The reason being the transparency as open data and the global coverage. Landsat missions are also the most complete set of historical satellite imagery starting with Landsat 1 in 1972.

Images are actually a collection of vectors/tuples in multiple dimensions. The elementary unit, that we can call "pixel", represents reflectance or emittance of energy at a specific wavelength in a unit in space. Space-wise, pixels are organized in a 2D array (matrix) with reflected/emitted energy recorded by the sensor at a specific interval in the electromagnetic (EEM) spectrum, commonly defined with a wavelength (W), at a specific time (T). Sensors commonly cover parts of the EEM spectrum, and thus return images with multiple 2D arrays at i-th wavelength, called bands. Images over the same area are collected many times over the year depending on the orbit configuration, and thus provide multi temporal data. So the value stored in each elementary unit, i.e. pixel (P) collected from satellite imagery can be considered as a function of the following four factors:

$$P = f(X,Y,W,T) \tag{1}$$

where X and Y are the coordinates in space of the center of the pixel, W is the wavelength representing the center of the EEM interval that the sensor is sensible to, and T is the time of recording of the information.

Seen together, by all means these data can be assigned to the big-data paradigm, as they have high volumes, variety and velocity, due to the number of sensors in orbit recording different types of EEM information at different times. It is one of the foreseen future situations that a constant influx of Earth Observation (EO) data will accelerate processes and analyses (Olbrich and Witjes, 2016).

To deal with the volume and complexity of data, big-data arrays have been developed. Some successful examples are available as open source. The RasDaMan "raster data manager" project, which lead to the rasdaman software suite (both with open source and enterprise forks). Rasdaman is an array engine allowing SQL-like access to data. Being SQL a well-known standard, the learning curve is not steep and is attractive to users (Baumann et al., 2018). Another popular project is the open data cube, which, as reported in the website: "The Open Data Cube (ODC) is an Open Source Geospatial Data Management and Analysis Software project that helps you harness the power of Satellite data. At its core, the ODC is a set of Python libraries and PostgreSQL database that helps you work with geospatial raster data.". Again here we see a similar approach, where existing database management systems (DBMS) such as PostgreSQL are harnessed and integrated to allow better management of multi-dimensional data-cubes (Killough, 2018).

Software as a Service (SaaS) and other online services for data and analysis are also becoming the chosen solution to avoid having to download large volumes of data for every project. Google Earth Engine (GEE) is becoming very popular among researchers, due to the large amount of data collections that are easily integrated in users' code (Gorelick et al., 2017). Several solutions have been created to support centralized and automated processing of multiple images. GEE provides users Petabytes of data at their fingertips, access to processing resources and an interface that provides a large number of tools for data processing via Javascript or Python programming environments. What took before days if not months can now be run in a few minutes or hours. GEE is available and free for academics as of today, but it must be noted that it is not to be taken for granted in the future. Other initiatives such as Copernicus RUS project that has closed at the end of 2021 also provided access to data (Copernicus data) and computing resources, to promote uptake of Copernicus data via educational and research activities.

All these solutions are key enabling technologies for Earth Observation. They have been applied by authors in the context of forest monitoring, in particular related to forest applications, like biomass estimation using different sensors (Pirotti et al., 2014), damage assessment from wind throw using ad-hoc indices from remote sensing data (Piragnolo et al., 2021) and comparing open-data sources to assess damaged areas (Laurin et al., 2020; Vaglio Laurin et al., 2016) or mapping wetlends with support of active remote sensing (LaRocque et al., 2020) or even to support spatialization of agricultural practices (Pagliacci et al., 2020).

The project that is reported in the following sections does not intend to compete with existing managers of data-cubes or to substitute larger projects that aim at array analytics. The following work aims at creating a lower-level system for integration of satellite data using straight-forward R functions and a web-interface, without having to migrate data to different data formats, but keeping the data as-is in well-known GeoTIFF or JPEG2000 raster formats. It can be said that the proposed system is a bridge between common visualization with typical geographic information systems (GIS) software and other more process-intensive procedures that require data extraction transform and load (ETL) procedures. The original data is stored in a straight forward file-based structure and existing libraries and software are called via system calls through a server-based R environment.

2. MATERIALS AND METHODS

2.1 Study area

The proposed system, InforSAT, is not limited to a specific area. It automatically picks up the areas that are covered by the images that are detected in the folders that are configured for monitoring via configuration files (see next section). Usually imagery is organized in granules that represent the footprint of the image or, as is most often the case, a number of rows from a push-broom sensor. In the reported implementation, the MultiSpectral Instrument (MSI) from Sentinel-2 is used, and this sensor uses a push-broom technique that increases the number of sensed rows along the orbit direction and across the orbital swath. Each product is provided as a granule that is approximately 100 km by 100 km (Figure 1).

Figure 1. Tiles from Sentinel-2 granules' footprints, with tile coding where the first two digits show the UTM zone and letters the code.

The tiles are then represented in the web interface of InforSAT (Figure 2) so that the user can choose which area to analyse. At this time only Sentinel-2 MSI data are implemented in InforSAT.

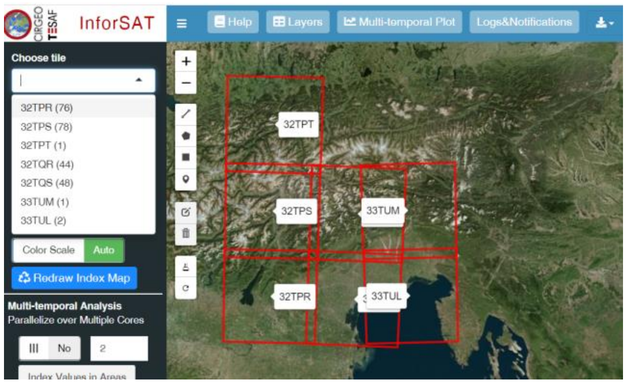


Figure 2. InforSAT's implementation with some S2 tiles available and each tile the number of available images that are found in the folder configuration.

2.2 Software and developer interface

The software is developed in R scripts in an online R environment installed in a Linux-based machine available via remote access. The developer interface is the open-source version of RStudio Server ("RStudio Documentation - RStudio Documentation," 2022). This allows to create a server-based system and to provide tools via web-based widgets. All of the processing is done server-side, and the tools are available to users via a web-page that is also used for visualization of the images and of the derived products of processing. Final products, such as zonal statistics of band-derived indices, can be downloaded for further processing.

2.3 Data structuring

2.3.1 Data storage: all image data are stored in a user-defined folder on the server, here-after called "data container". A list of configuration variables contains the path to the folder. The configuration variables are in a specific file that is read by the R environment, *globals.R*. A specific R script checks weekly (or at other user-defined intervals) for new images. This script simply checks for available imagery and downloads and extracts it from zip archives to the configured folder. Any type of sensor can be simply added by assigning a sub-folder to the main folder defined as “data container”, and the download script must of course take in consideration the APIs of the desired provider. The specific script in this study case downloads S2 Level 2 data, and is saved in a file called “*esaCron.R*”, see github page (Pirotti, 2021).

2.3.2 Metadata mining: another function, stored in the *globals.R* script, crawls the folder structure in the data storage folder and collects information about all the images that are detected. Image detection is done by filtering with extension values (.tif and .jp2). It stores the information in a mapped list structure where the key is the full path to the image file. Image paths, corresponding band wavelengths and also histograms of digital number (DN) values for each band are stored in this object.

This meta-information is used for several procedures in the interface used for rendering the data. For example, the histogram of DN values is an array of the percentiles of the frequency distribution of the DN values. This allows to render images with color balancing via histogram equalization.

The metadata is currently organized as follows. For each granule the following data are stored:

a. a path to a folder containing all files related to that granule;
b. total folder size;
c. UTM tile code;
d. list of bands with the following information:
    i. Central wavelength
    ii. Full path to image file
    iii. All percentiles of DN (100 floating point digits)
    iv. Band name according to sensor (e.g. B1, B2 …)

As mentioned, all of the above is automatically stored once the image is downloaded. Periodical daily validation is carried out to ensure that there is a match between this list and the available images. For example, a user might remove a folder with a granule, and this missing image will be detected and removed from the metadata list via this validation procedure. This ensures that there is a coherent list of images in the interface that corresponds to the real available imagery.

2.4 Data visualization

Satellite data cubes in InforSAT are accessible via a web interface that is created via the Shiny R package which provides interactive widgets that create a bridge between the online user and a server-based R environment. Figure 2 shows the overall interface that the user can access via a web url. Figure 3 and Figure 4 show some of the aspects of the interface. First the user must choose the area of interest by choosing a granule via the tile code. Then the user can choose a specific date for which to render the image (Figure 3).

Once the user chooses the image date, a raster with index values is automatically created and rendered. The index to be calculated can be chosen from a specific toolbar or the user can provide its own band math using the interface to insert the formula (Figure 3 right). Each index raster is calculated every time the user calls for rendering the raster. The procedure consists in sampling the original image with points that correspond to the center of the screen pixels, reprojected from screen coordinates to image coordinates. Depending on the screen size and on the area, these are around one million points. These points are then converted to an image and rendered on screen with a fixed scale that depends on the expected minimum and maximum values of the index (e.g. for the normalized vegetation index that would be between -1 and 1) or a scale that automatically stretches between the 10^{th} and 90^{th} percentile of the frequency distribution of the index values of the currently rendered image.

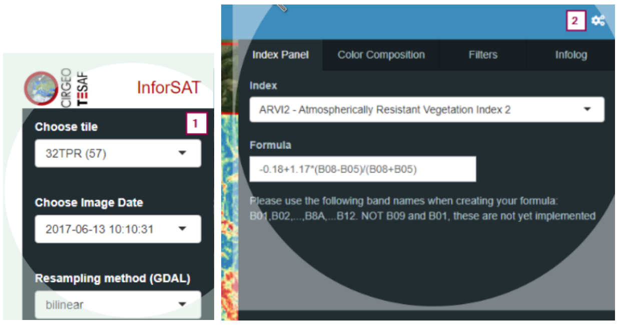


Figure 3. Left: drop down menus for choosing the tile (see Figure 2) and the specific date. The resampling menu on the bottom provides choices for resampling the output raster. Right: dropdown menu for choosing the index to be calculated, or the space for creating the user-defined index.

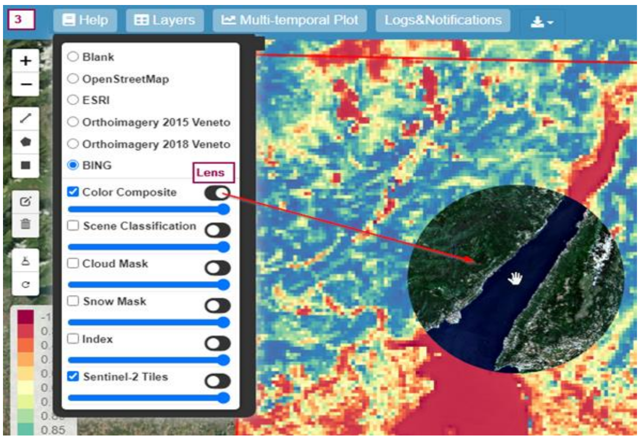


Figure 4. The result from rendering the chosen index raster (see Figure 3). It is possible to overlap another layer for interactive visualization via the lens activation.

Users can also render real-color and false-color composites defining their own band combinations. The images to be rendered on the user browser are processed on-the-fly from the original JPEG2000 format of Sentinel-2. The color-composites are automatically drawn at any scale using the intrinsic overviews for each Sentinel-2 band that are present from the JPEG2000 format. Index rasters and color composites can be interactively mixed also by combining their visualization using a transparency bar or a lens clip that allows to compare specific areas (Figure 4).

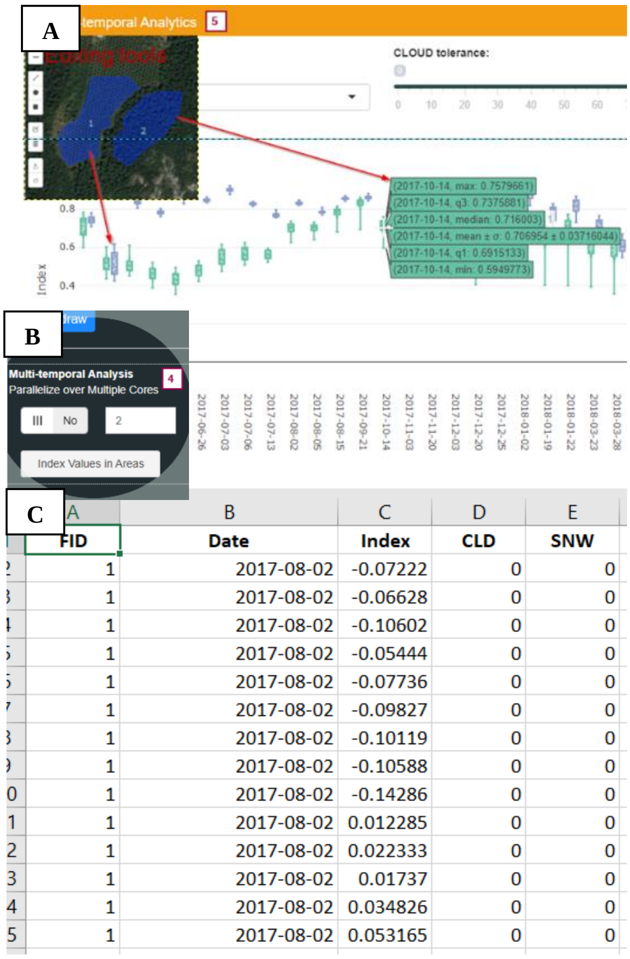
Under the hood the R functions call internal R packages (*terra*, 2022) but also calls directly GDAL (*GDAL - Geospatial Data Abstraction Library*, 2022) and mapserver (*MapServer*, 2022) executables. Mapserver was used to create a dynamic mapfile that is used to stream color composites and index files via a WMS service to the client. The mapserver file is created for each user session and is populated with raster layers according to the chosen granule. Index raster are created on the fly and stored in a temporary file and addressed in the mapfile for rendering.

2.5 Data analysis

Data analysis in InforSAT consists in providing users with zonal statistics that can drill through all the timeline of available images and extract values of each pixel inside the area. This aggregated statistic is then plotted in a time-series plot, using boxplots or points and error bars representing standard deviation, in case of polygons (Figure 5). The results section will report some analyses over areas that were burned with fires in different dates. Fire areas and dates were extracted from the EFFIS fire database (San Miguel Ayanz et al., 2003).

3. RESULTS

Regarding multi-temporal analysis, users can define one ore more polygons over the area and for each polygon extract single pixel values (digital numbers – DN) and aggregated zonal statistics for each and all available images in a few seconds, with or without using parallel processing mode. Users can download the multi-temporal data, i.e. the DN values, in table format for further analysis. The table is in long format and has a column with a timestamp, one with polygon ID and one column for each band with values (Figure 5C).



FID	Date	Index	CLD	SNW
1	2017-08-02	-0.07222	0	0
1	2017-08-02	-0.06628	0	0
1	2017-08-02	-0.10602	0	0
1	2017-08-02	-0.05444	0	0
1	2017-08-02	-0.07736	0	0
1	2017-08-02	-0.09827	0	0
1	2017-08-02	-0.10119	0	0
1	2017-08-02	-0.10588	0	0
1	2017-08-02	-0.14286	0	0
1	2017-08-02	0.012285	0	0
1	2017-08-02	0.022333	0	0
1	2017-08-02	0.01737	0	0
1	2017-08-02	0.034826	0	0
1	2017-08-02	0.053165	0	0

Figure 5. (A) panel area for defining the number of parallel clusters used for running the zonal statistic. (B): results of the analysis, plotted as boxplots for two areas. (C) downloaded table with polygon/point ID and index, cloud and snow probability values.

In both visualization and multi-temporal analysis, users can decide a threshold for masking according to cloud and snow probability, which are available products from the sen2cor processing of Sentinel-2 to level 2C. An example can be seen in Figure 6D were the map has null data from filtering out pixels with cloud or snow probability above zero.

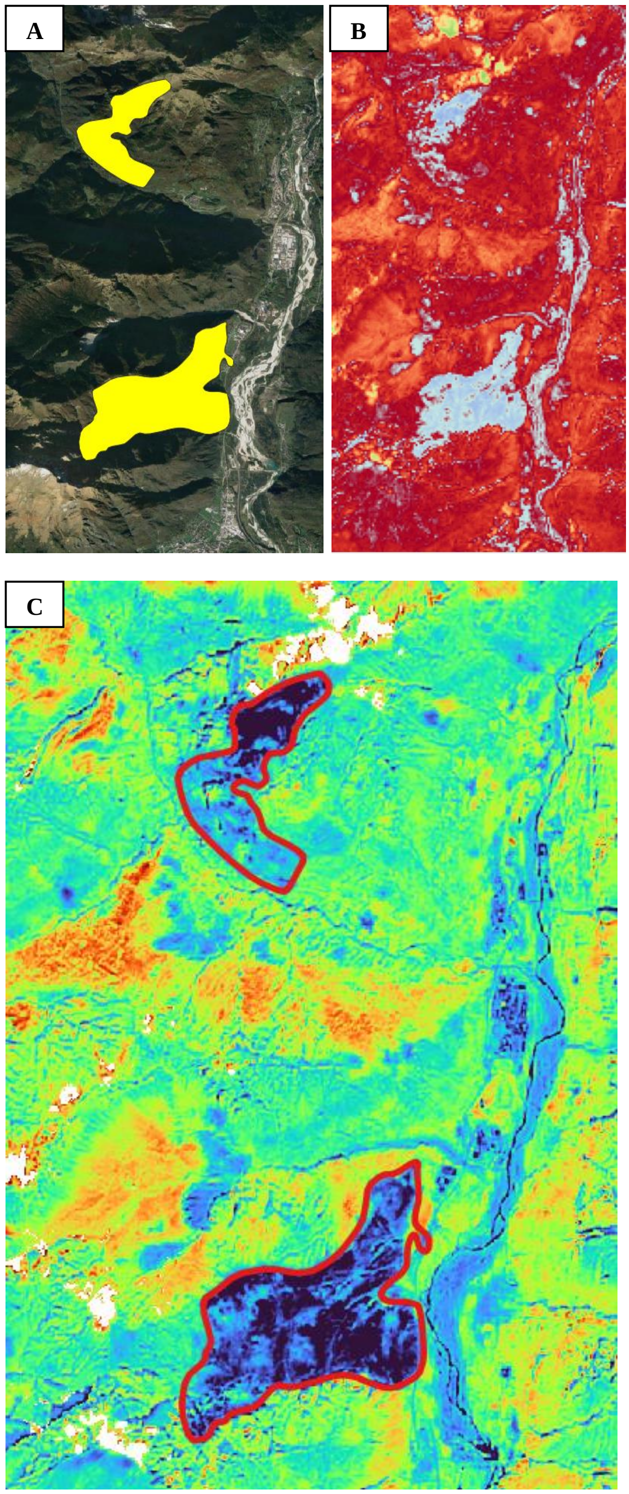


Figure 6. (A) burned areas from EFFIS database that occurred in 23 March 2022, Ponte delle Alpi and Longarone locations; (B) NBR index in GIS environment of the same areas; (C) dNBR index calculated in GIS environment after downloading rasters of NBR.

https://doi.org/10.5194/isprs-archives-XLVIII-4-W1-2022-365-2022 | © Author(s) 2022. CC BY 4.0 License.

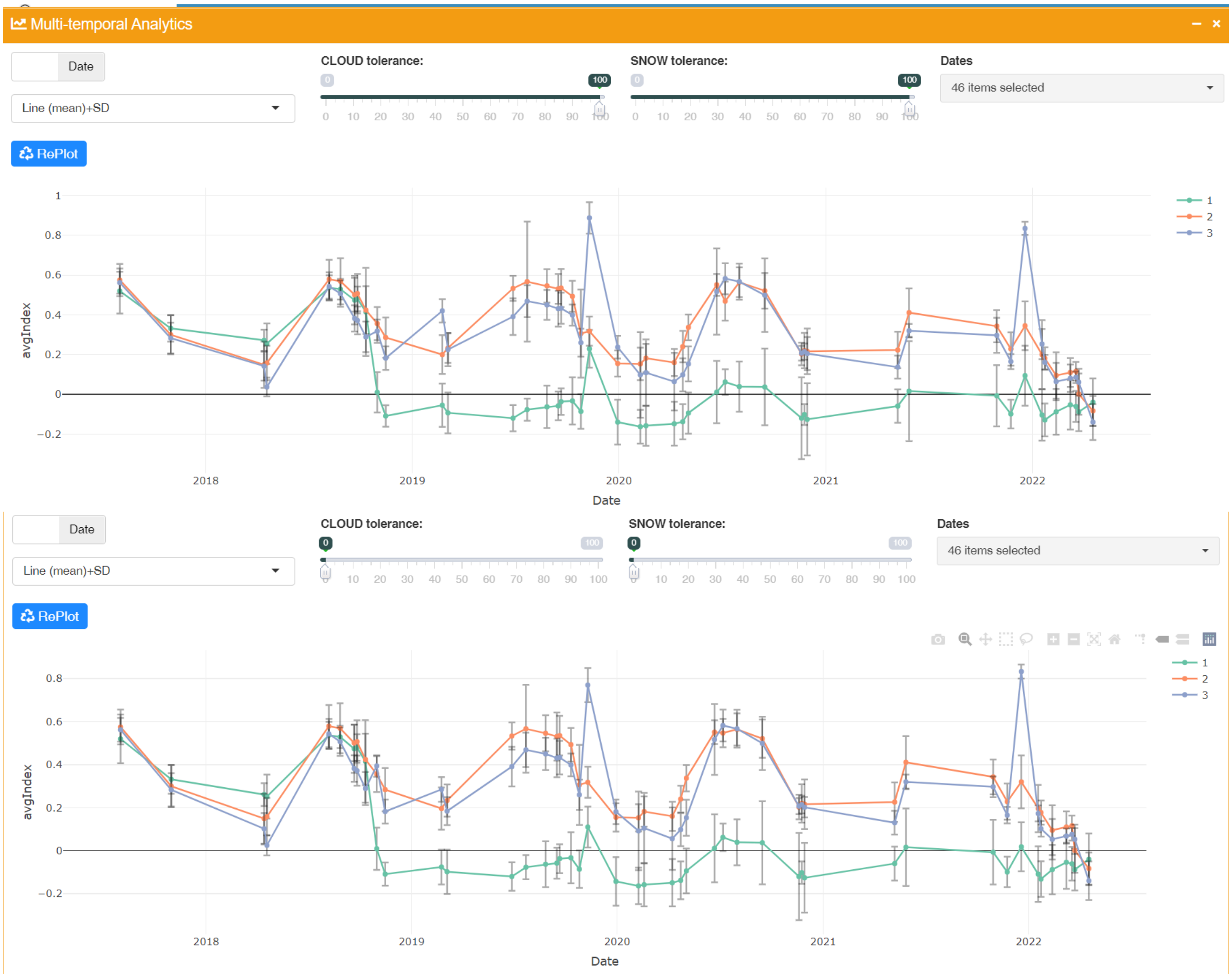


Figure 7. Zonal statistics of three burned areas: 2 and 3 are the areas shown in figure 5, and line 1 is the area of Taibon Agordino that burned the 24th October 2018. The top plot is without filtering cloud/snow, the bottom one is with full filtering of any pixel above zero probability. Notice the different scale in the Y axis.

4. DISCUSSION

Figure 6 shows two areas that were burned by fires in the same date, 23 March 2022. Different indices, NDVI and NBR can be calculated in InforSAT and the index raster can be exported in a local GIS environment as a GeoTIFF file. This supports further processing in a local environment. The raster that is rendered in InforSAT is usually resampled to the screen's resolution, i.e. if the screen is viewing a whole Sentinel-2 image in a 500 rows by 500 columns screen space, then the raster that is created server side by calling GDAL's gdal_warp executable is sample to a resolution that will produce a similar number of rows and columns. The user can bypass this configuration by selecting to force the full resolution of the original image, which is set to 20 m for Sentinel-2. This allows users to download their own data-cubes with a specific index at each desired date that is available for that area.

Figure 7 is the result of plotting the two areas plus another area which had a fire event in October 2018. The plot shows the average and standard deviation, without and with the cloud/snow filter applied. These data can be downloaded in an Excel spreadsheet for further analysis. The downloaded data consists in the following columns: ID of polygon, id of cell(pixel), timestamp,,index value, cloud probability, snow probability. It is clearly seen from the plots that after the fire event, NBR values drop below zero. It is also worth noting the effect of the phenology stage on NBR values, driven by seasonality, that is clearly seen in the plot for the two areas that burned in 2022.

The InforSAT project is still in development, and it aims to support researchers but also to support educational efforts. It can be used to explain basic concepts of remote sensing such as band math for extracting indices, band false color composites for image interpretation, image resampling and visualization, zonal statistics and the impact of phenology on temporal series of indices. Other authors have discussed and implemented online learning tools, but most are for specific applications, such as (Maggioni et al., 2020) that focus on hydrological applications and also use a web-based portal. The authors in (Joyce et al., 2014)

5. CONCLUSIONS

This is a proposed solution for a system that can be used for access and analysis of imagery directly from the original raster formats as they are available from the missions' APIs. The advantage is that the scripts automatically collect data that is necessary for the web interface. The web interface allows access from any client browser to a large volume of data that is serviced online with little internet traffic as the processing is all optimized on the server. The final product can be viewed as styled rasters or downloaded as a raster or an excel data table with zonal statistics from user-defined areas or points.

In the near future this solution will be integrated in an R package, allowing users to easily download, install and replicate their own portal locally or in their own server.

ACKNOWLEDGEMENTS

This activity was partly funded by the FIRE-RES project, Grant agreement ID: 101037419
https://cordis.europa.eu/project/id/101037419

The server was supported by University of Padova's VSIX Neutral Access Point that hosts CIRGEO's computational infrastructure from HyperEarths project https://vsix.unipd.it/

REFERENCES

Baumann, P., Misev, D., Merticariu, V., Huu, B.P., Bell, B., 2018. rasdaman: Spatio-temporal datacubes on steroids, in: Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. Presented at the SIGSPATIAL '18: 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, Seattle Washington, pp. 604–607. https://doi.org/10.1145/3274895.3274988

GDAL - Geospatial Data Abstraction Library, 2022. . Open Source Geospatial Foundation.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment, Big Remotely Sensed Data: tools, applications and experiences 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Joyce, K.E., Boitshwarelo, B., Phinn, S.R., Hill, G.J., Kelly, G.D., 2014. Interactive online tools for enhancing student learning experiences in remote sensing. Journal of Geography in Higher Education 38, 431–439.

Killough, B., 2018. Overview of the Open Data Cube Initiative, in: IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium. Presented at the IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, IEEE, Valencia, pp. 8629–8632. https://doi.org/10.1109/IGARSS.2018.8517694

LaRocque, A., Phiri, C., Leblon, B., Pirotti, F., Connor, K., Hanson, A., 2020. Wetland Mapping with Landsat 8 OLI, Sentinel-1, ALOS-1 PALSAR, and LiDAR Data in Southern New Brunswick, Canada. Remote Sensing 12, 2095. https://doi.org/10.3390/rs12132095

Laurin, G.V., Francini, S., Luti, T., Chirici, G., Pirotti, F., Papale, D., 2020. Satellite open data to monitor forest damage caused by extreme climate-induced events: a case study of the Vaia storm in Northern Italy. Forestry: An International Journal of Forest Research. https://doi.org/10.1093/forestry/cpaa043

Maggioni, V., Girotto, M., Habib, E., Gallagher, M.A., 2020. Building an Online Learning Module for Satellite Remote Sensing Applications in Hydrologic Science. Remote Sensing 12, 3009. https://doi.org/10.3390/rs12183009

MapServer, 2022. . MapServer.

Olbrich, P., Witjes, N., 2016. Sociotechnical Imaginaries of Big Data: Commercial Satellite Imagery and Its Promise of Speed and Transparency, in: Bunnik, A., Cawley, A., Mulqueen, M., Zwitter, A. (Eds.), Big Data Challenges: Society, Security, Innovation and Ethics. Palgrave Macmillan UK, London, pp. 115–126. https://doi.org/10.1057/978-1-349-94885-7_10

Pagliacci, F., Defrancesco, E., Mozzato, D., Bortolini, L., Pezzuolo, A., Pirotti, F., Pisani, E., Gatto, P., 2020. Drivers of farmers' adoption and continuation of climate-smart agricultural practices. A study from northeastern Italy. Science of The Total Environment 710, 136345. https://doi.org/10.1016/j.scitotenv.2019.136345

Piragnolo, M., Pirotti, F., Zanrosso, C., Lingua, E., Grigolato, S., 2021. Responding to Large-Scale Forest Damage in an Alpine Environment with Remote Sensing, Machine Learning, and Web-GIS. Remote Sensing 13, 1541. https://doi.org/10.3390/rs13081541

Pirotti, F., 2021. InforSAT https://github.com/fpirotti/inforsat .

Pirotti, F., Laurin, G., Vettore, A., Masiero, A., Valentini, R., 2014. Small Footprint Full-Waveform Metrics Contribution to the Prediction of Biomass in Tropical Forests. Remote Sensing 6, 9576–9599. https://doi.org/10.3390/rs6109576

RStudio Documentation - RStudio Documentation [WWW Document], 2022. URL https://docs.rstudio.com/ (accessed 6.1.22).

San Miguel Ayanz, J., Barbosa, P., Schmuck, G., Liberta, G., Schulte, E., Gitas, I., 2003. Towards a coherent forest fire information system in Europe: the European Forest Fire Information System (EFFIS). Environmental Monitoring in the South-Eastern Mediterranean Region Using RS/GIS Techniques; Gitas, IZ, San Miguel Ayanz, J., Eds 5–16.

terra, 2022. . rspatial.

Vaglio Laurin, G., Hawthorne, W., Chiti, T., Di Paola, A., Cazzolla Gatti, R., Marconi, S., Noce, S., Grieco, E., Pirotti, F., Valentini, R., 2016. Does degradation from selective logging and illegal activities differently impact forest resources? A case study in Ghana. iForest - Biogeosciences and Forestry e1–e9. https://doi.org/10.3832/ifor1779-008