The Journal of Open Source Software

scikit-eo: A Python package for Remote Sensing Data Analysis

Yonatan Tarazona [1], **Fernando Benitez-Paez** [2], **Jakub Nowosad** [3], **Fabian Drenkhan** [4], **and Martín E. Timaná** [5]

1 Department of Earth Sciences, Center for Earth and Space Research (CITEUC), University of Coimbra, Portugal **2** The School of Geography and Sustainable Development, University of St Andrews, The UK **3** Adam Mickiewicz University in Poznań **4** Geography and the Environment, Department of Humanities, Pontificia Universidad Católica del Perú, Lima, Peru **5** Applied Geography Research Center, Pontificia Universidad Católica del Perú, Lima, Peru

DOI: 10.21105/joss.06692

Software

- Review
- Repository
- Archive

Editor: Michael Mahoney

Reviewers:

- @dbuscombe-usgs
- @KBodolai

Submitted: 14 March 2024
Published: 16 July 2024

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

Summary

In recent years, a growing body of space-borne and drone imagery has become available with increasing spatial and temporal resolutions. This remotely sensed data has enabled researchers to address and tackle a broader range of challenges effectively by using novel tools and data. However, analysts spend an important amount of time finding the adequate libraries to read and process remotely sensed data.

With an increasing amount of open access data, there is a growing need to account for effective open source tools to read, process and execute analysis that contributes to underpin patterns, changes and trends that are critical for environmental studies. Applications that integrate spatial-temporal data are used to study a variety of complex environmental processes, such as monitoring and assessment of land cover changes (Chaves et al., 2020), crop classifications (Pott et al., 2021), deforestation (Tarazona et al., 2018), impact on urbanization level (Trinder & Liu, 2020) and climate change impacts (Yang et al., 2013). Other complex environmental processes that are monitored by integrating spatial-temporal data are assessments of glacier retreat (Hugonnet et al., 2018), related hydrological change (Huss & Hock, 2018), biodiversity conservation (Cavender-Bares et al., 2022) and disaster management (Kucharczyk & Hugenholtz, 2021).

To bridge the gaps in remotely sensed data processing tools, we here introduce **scikit-eo**, a brand-new Python package for satellite remote sensing analysis. Unlike other tools, it is a centralized, scalable, and open-source toolkit due to its flexibility in being adapted into large dataset processing pipelines. It provides central access to the most commonly used Python functions in remote sensing analysis for environmental studies, including machine learning methods. **scikit-eo** stands out with its ability to be used in various settings, from a lecturer room to a crucial part of any Python environment in a research project. The majority of the tools included in **scikit-eo** are derived from peer-reviewed scientific publications, ensuring their reliability and accuracy.

By integrating this diverse set of tools, **scikit-eo** allows to focus on analyzing the results of data rather than being bogged down by complex lines of code. With its centralized structure, integrated use cases, and example data, **scikit-eo** empowers to optimize resources and dedicate more attention to the meaningful interpretation of findings in a more efficient way.

Statement of Need

As remote sensing data and sophisticated processing tools become increasingly available, there is a growing need for scalable and customized toolkits to help environmental researchers classify satellite and drone imagery quickly and accurately. Our goal is to simplify the identification of patterns, changes, and trends that are crucial for environmental research and analysis.

In this paper, we introduce **scikit-eo**, a specialized library with tailored analysis capabilities designed to meet the unique demands of environmental studies. From statistical methods to machine learning algorithms, **scikit-eo** assists researchers in uncovering intricate spatial patterns, relationships, and trends, while also simplifying the evaluation and calibration of generated outputs.

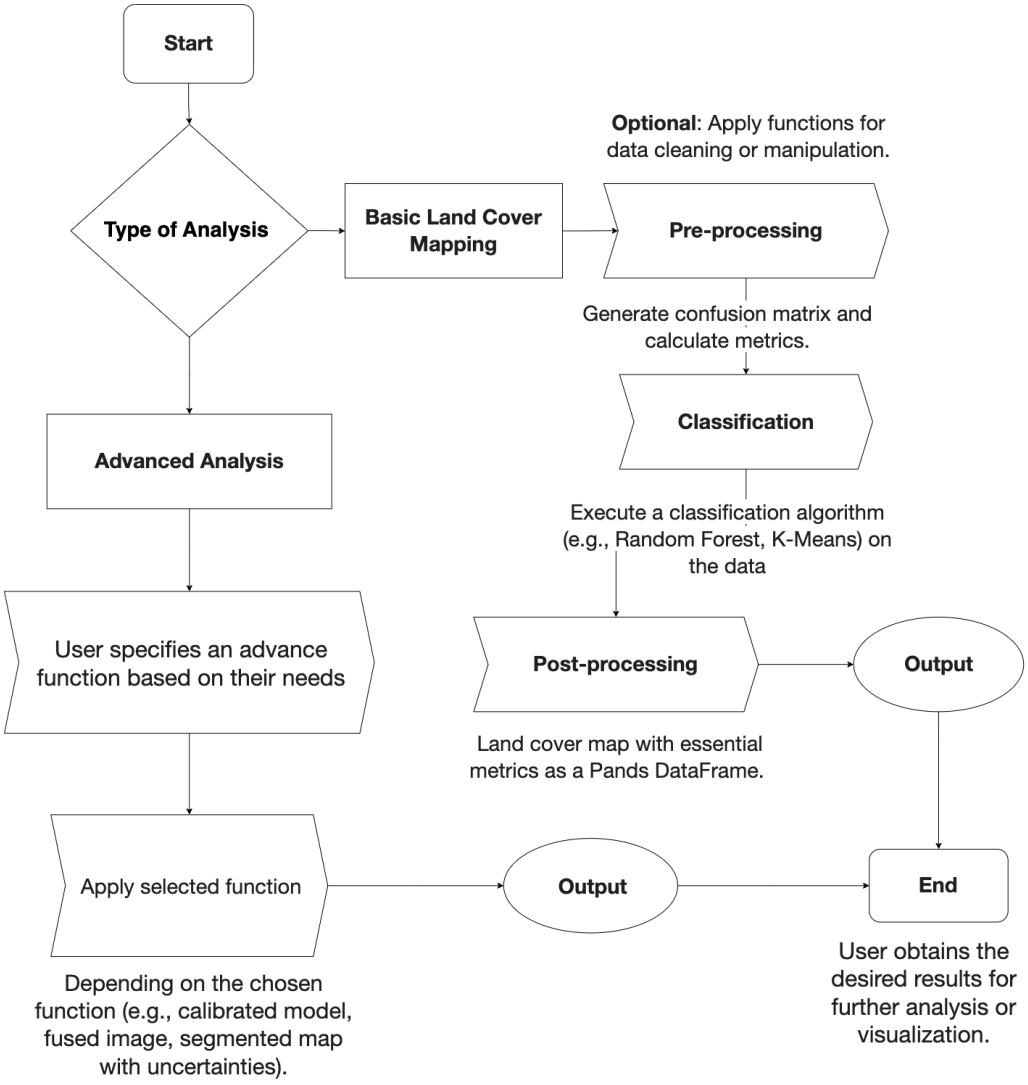

Figure 1: Workflow of main functionalities of the *scikit-eo* python package as well as outputs that can be obtained by using the tools developed.

scikit-eo is an open-source package built entirely in Python through Object-Oriented Programming and Structured Programming that provides a helpful variety of remote sensing tools (see Figure 1), from primary and exploratory functions to more advanced methods to classify,

calibrate, or fuse satellite imagery. Depending on the users' needs, **scikit-eo** can provide the basic but essential land cover characterization mapping, including the confusion matrix and the required metrics such as user's accuracy, producer's accuracy, omission and commission errors. These required metrics can be combined as a pandas `DataFrame` object. Furthermore, a class prediction map is a result of land cover mapping, i.e., a land cover map, which represents the output of the classification algorithm or the output of the segmentation algorithm. These two outcomes must include uncertainties with a confidence level (e.g., at 95% or 90%). All required metrics from the confusion matrix can be easily computed and included confidence levels with **scikit-eo** following guidance proposed by Olofsson et al. (2014). Other useful tools for remote sensing analysis can be found in this package; and for more information about the full list of the supported functions, tutorials as well as how to install and execute the package within a Python setting, visit the scikit-eo website.

Audience

scikit-eo is an adaptable Python package that covers multiple users, including students, remote sensing professionals, environmental analysis researchers, and organizations looking for satellite image analysis. Its tools and algorithms implemented make it well-suited for various applications, such as university teaching, that includes technical and practical sessions and cutting-edge research using the most recent machine learning and deep learning techniques applied to remote sensing.

This python package provides key tools for both students seeking insights from a satellite image analysis or an experienced researcher looking for advanced tools. **scikit-eo** offers a valuable resource to support the most valuable methods for environmental studies.

scikit-eo as a research tool:

In environmental studies, particularly in land cover classification, the availability of scalable but user-friendly software tools is crucial for facilitating research, analysis, and modelling. With the widespread adoption of Python as one of the most popular programming languages, particularly in the GIScience and remote sensing fields, developing specialized packages has massively enhanced the effectiveness of environmental research. **scikit-eo** is a dedicated piece of software tailored to address unique challenges encountered in land cover mapping, forest or land degradation (Tarazona & Miyasiro-López, 2020), and the fusion of multiple satellite imagery from several formats (Tarazona et al., 2021). **scikit-eo** provides the assessment and calibration metrics to evaluate the provided outputs. The current version **scikit-eo** requires that users provide the dataset to process. However, we expect to provide a wide range of functionalities for acquiring environmental data from diverse sources and file formats, enabling researchers to access satellite imagery.

One of **scikit-eo's** key strengths is its advanced analysis capabilities. It provides a rich suite of algorithms specifically designed for environmental studies, including statistical analysis, deep learning, data fusion, and spatial analysis. Researchers can leverage these tools to explore patterns, relationships, and trends within their datasets, uncover complex land or forest degradation or mapping, classify the land cover, and generate insightful visualizations.

As a particular example of these advanced analysis capabilities, we have integrated the **`deepLearning`** function, which includes the *Fully connected layers (FC)*, also known as *dense layers* model. This is one of the most straightforward yet functional neural networks we can apply to remote sensing analysis. The term "fully connected" comes from the fact that each neuron in one layer is connected to every neuron in the preceding layer, creating a highly interconnected network known as the Multi-Layer Perceptron (MLP). The model is trained using a specified dataset, with the bands as *input_shape*, including input features and corresponding class labels. The weights W are initialized and then adjusted during training to ensure the neural network's output is consistent with the class labels. The training process

involves minimizing the error function using *Gradient Descent* and *backpropagation* algorithms. The activation functions used in this model are ReLU (*Rectified Linear Unit*) for the neurons in each hidden layer and *Softmax* for the final classification layer. For more details, see tutorial No 11 In future version, u-net architectures will be implemented within **scikit-eo**.

It's important to note that *ReLU* introduces non-linearity to the model, enabling it to learn complex patterns, while *Softmax* is used for multi-class classification, transforming the output into a probability distribution over multiple classes and better suited for more than two land covers. Unlike traditional machine learning models, such as support vector machine or decision tree, which typically are simpler and don't involve multiple layers, the FC model uses multiple hidden layers, allowing it to learn hierarchical representations of the input data. Deep learning models like this one can automatically learn and extract complex features from the raw input, making them exceptionally powerful for tasks such as land cover classification.

To run an example of how to use the function **`deepLearning`** find a detailed notebook in tutorial No 11 Deep Learning Classification.

scikit-eo in the lecture room:

scikit-eo can be part of a classroom as part of the set of tools for environmental studies where a quantitative approach and computer labs are required. Therefore, first of all an appropriate introduction of Python, the basics of remote sensing, and the relevance of environmental studies to address climate change challenges or impacts of anthropogenic activity are needed. Lectures can use the simplicity of **scikit-eo** routines to execute supervised classification methods, `Principal Components Analysis`, `Spectral Mixture Analysis`, `Mapping forest` or `land degradation` and more types of analysis. By reducing the number of required lines of code, students can focus on the analysis and how the methods work rather than dealing with complex and unnecessary programming tasks. Lecturers can structure their computer labs using open data sources and integrate **scikit-eo** to allow students to understand the importance of the calibration and assessment metrics, get insights about the classification mapping using satellite imagery, and provide an introduction to more advanced methods that include machine learning techniques.

scikit-eo as open source tool:

As open-source software keeps transforming the landscape of scientific research (The Turing Way Community et al., 2019), enabling collaboration, reproducibility and transparency, **scikit-eo** was explicitly developed as an open-source tool. **scikit-eo** integrates most of the popular open source python libraries from the so-called geo-python stack for remote sensing (e.g.,. numpy, pandas, rasterio and few more) to extent and create a centralised package for advanced spatial analysis of remotely sensed data. The package provides researchers and developers with a free, scalable, and community-driven platform to process, analyse, and visualise satellite imagery, specifically, centralising multiple functions' use for classification and mapping land cover.

Functionalities

Main tools

Scikit-eo includes several algorithms to process satellite images to assess complex environmental processes and impacts. These include functions such as atmospheric correction, machine learning and deep learning techniques, estimating area and uncertainty, linear trend analysis, combination of optical and radar images, and classification sub-pixel, and more. The main functions are listed below:

Table 1: Main tools available for **scikit-eo** package.

Name of functions/classes	Description
mla	Supervised Classification in Remote Sensing
calmla	Calibrating Supervised Classification in Remote Sensing
confintervalML	Information of Confusion Matrix by proportions of area, overall accuracy, user's accuracy with confidence interval and estimated area with confidence interval as well.
deepLearning	Deep Learning algorithms
atmosCorr	Radiometric and Atmospheric Correction (currently supporting Landsat)
rkmeans	K-means classification
calkmeans	This function allows to calibrate the kmeans algorithm. It is possible to obtain the best k value and the best embedded algorithm in kmeans.
pca	Principal Components Analysis
linearTrend	Linear trend is useful for mapping forest degradation or land degradation
fusionrs	This algorithm allows to fuse images coming from different spectral sensors (e.g., optical-optical, optical and SAR or SAR-SAR). Among many of the qualities of this function, it is possible to obtain the contribution (%) of each variable in the fused image
sma	Spectral Mixture Analysis - Sub-pixel classification
tassCap	The Tasseled-Cap Transformation

For more information the reader is referred to the scikit-eo website.

State of the field:

Scikit-eo is built upon well-known packages and libraries that support remote sensing analysis in python, but is designed specifically to study land cover classification, and providing tailored functionalities for environmental studies. It aims to simplify the identification of patterns, changes, and trends for environmental research. Like many other Python packages, **Scikit-eo** makes use of *Rasterio* and *GDAL*, which are essential for geospatial data handling. Additionally, **Scikit-eo** is based on **Scikit-learn** and **tensorflow**, a widely-used machine learning libraries in Python that includes several algorithms for classification, regression, clustering, and dimensionality reduction. However, **Scikit-eo** includes machine learning tools to remote sensing analysis. Another well-known library, *Geemap* that provides interactive mapping capabilities using Google Earth Engine and focuses more on visualization and data exploration, while **Scikit-eo** is designed for deeper analytical tasks using remote sensing data for land cover analysis. The *SITS* package is particularly useful for time series, while **Scikit-eo** provides functionalities for land cover studies. *torchgeo* a popular Python library that provides comprehensive deep learning functions for geospatial data. However **Scikit-eo** integrates machine learning methods that are particularly useful for analysing satellite imagery with multiple bands for environmental studies. Lastly, *EO-learn* can be considered the most aligned package with **Scikit-eo** providing efficient processing and analysis of satellite imagery capabilities but lacking the deep learning and tailored functions specifically included for land cover analysis.

Acknowledgments

The authors would like to thank to David Montero Loaiza for the idea of the package name and Qiusheng Wu for the suggestions that helped to improve the package.

References

Cavender-Bares, J., Schneider, F. D., Santos, M. J., Armstrong, A., Carnaval, A., Dahlin, K. M., Fatoyinbo, L., Hurtt, G. C., Schimel, D., Townsend, P. A., Ustin, S. L., Wang, Z., & Wilson, A. M. (2022). Integrating remote sensing with ecology and evolution to advance biodiversity conservation. *Nature Ecology & Evolution 2022 6:5*, *6*, 506–519. https://doi.org/10.1038/s41559-022-01702-5

Chaves, M. E. D., Picoli, M. C. A., & Sanches, I. D. (2020). Recent applications of Landsat 8/OLI and Sentinel-2/MSI for land use and land cover mapping: A systematic review. *Remote Sensing*, *12*(18). https://doi.org/10.3390/rs12183062

Hugonnet, R., McNabb, R., Berthier, E., Menounos, B., Nuth, C., Girod, L., Farinotti, D., Huss, M., Dussaillant, I., Brun, F., & Kääb, A. (2018). Accelerated global glacier mass loss in the early twenty-first century. *Nature*, *592*, 726–731. https://doi.org/10.1038/s41586-021-03436-z

Huss, M., & Hock, R. (2018). Global-scale hydrological response to future glacier mass loss. *Nature Climate Change*, *8*, 135–140. https://doi.org/10.1038/s41558-017-0049-x

Kucharczyk, M., & Hugenholtz, C. H. (2021). Remote sensing of natural hazard-related disasters with small drones: Global trends, biases, and research opportunities. *Remote Sensing of Environment*, *264*, 112577. https://doi.org/10.1016/J.RSE.2021.112577

Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., & Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, *148*, 42–57. https://doi.org/10.1016/j.rse.2014.02.015

Pott, L. P., Amado, T. J. C., Schwalbert, R. A., Corassa, G. M., & Ciampitti, I. A. (2021). Satellite-based data fusion crop type classification and mapping in Rio Grande do Sul, Brazil. *ISPRS Journal of Photogrammetry and Remote Sensing*, *176*, 196–210. https://doi.org/10.1016/j.isprsjprs.2021.04.015

Tarazona, Y., Alaitz, Z., Xavier, P., Antoni, B., Jakub, N., & Hamdi A., Z. (2021). Fusing Landsat and SAR data for mapping tropical deforestation through machine learning classification and the PVts-β non-seasonal detection approach. *Canadian Journal of Remote Sensing*, *47*, 677–696. https://doi.org/10.1080/07038992.2021.1941823

Tarazona, Y., Mantas, V. M., & Pereira, A. J. S. C. (2018). Improving tropical deforestation detection through using photosynthetic vegetation time series – (PVts-β). *Ecological Indicators*, *94*, 367–379. https://doi.org/10.1016/j.ecolind.2018.07.012

Tarazona, Y., & Miyasiro-López, M. (2020). Monitoring tropical forest degradation using remote sensing. Challenges and opportunities in the Madre de Dios region, Peru. *Remote Sensing Applications: Society and Environment*, *19*, 100337. https://doi.org/10.1016/j.rsase.2020.100337

The Turing Way Community, Arnold, B., Bowler, L., Gibson, S., Herterich, P., Higman, R., Krystalli, A., Morley, A., O'Reilly, M., & Whitaker, K. (2019). *The Turing Way: A handbook for reproducible data science*. https://doi.org/10.5281/ZENODO.3233986

Trinder, J., & Liu, Q. (2020). Assessing environmental impacts of urban growth using remote sensing. *Geo-Spatial Information Science*, *23*, 20–39. https://doi.org/10.1080/10095020.2019.1710438

Yang, J., Gong, P., Fu, R., Zhang, M., Chen, J., Liang, S., Xu, B., Shi, J., & Dickinson, R. (2013). The role of satellite remote sensing in climate change studies. *Nature Climate Change*, *3*, 875–883. https://doi.org/10.1038/nclimate1908