The Journal of Open Source Software

Museo ToolBox : a python library for remote sensing including a new way to handle rasters.

Nicolas Karasiak[1]

1 Université de Toulouse, INRAE, UMR DYNAFOR, Castanet-Tolosan, France

DOI: 10.21105/joss.01978

Software

- Review
- Repository
- Archive

Editor: Katy Barnhart

Reviewers:

- @cmillion
- @mollenburger

Submitted: 12 December 2019
Published: 20 December 2019

Summary

`Museo ToolBox` is a python library dedicated to the processing of images in remote sensing. Based on the fact that a majority of the needs in machine learning requires knowledge on how to transform your data and since it uses a lot of similar lines of codes on various projects but for the same usage (e.g., for reading and writing the raster block per block, computing a spectral index, fitting a model…), we offer with this library a new approach to compute functions on a raster. For example, as in our field a recurrent usage is to fit a model and predict or to use some functions like one to compute for example a spectral index, `Museo ToolBox` automatically transforms the raster to match your needs (for learning a model, the user needs an array with one line per pixel and its features as columns). Other modules help users to g generate stratified spatial or non-spatial cross-validation, or state-of-the-art learning methods with a automatic grid search and standardized data using algorithms from scikit-learn.

Museo ToolBox's goal is to make working with raster data very easier for scientists or students and to promote the use of spatial cross-validation.

A full documentation is available online on read the docs.

Museo ToolBox functionnalities

`Museo ToolBox` is organized into several modules :

- processing : raster and vector processing.
- cross-validation : stratified cross-validation compatible with scikit-learn
- ai : machine learning module
- charts : plot confusion matrix with F1 score, mean, or producer/user's accuracy.
- stats : compute stats (like Moran's Index, confusion matrix, commision/omission) or extract truth and predicted label from confusion matrix.

Here are some main usages of `Museo ToolBox` :

1. Read and write a raster block per block using your own function.
2. Generate a cross-validation, including spatial cross-validation.
3. Fit models with scikit-learn, extract accuracy from each cross-validation fold, and predict raster.
4. Plot confusion matrix and add f1 score or producer/user accuracy.
5. Get the y_true and and y_predicted labels from a confusion matrix.

The Journal of Open Source Software

RasterMath

Available in `museotoolbox.processing.RasterMath`, RasterMath class is the keystone of `Museo ToolBox`.

The question is simple : How can the transposition of array-compatible functions to raster compatibility be simplified ? The idea behind `RasterMath` is, if your function works with an array, then it will work directly with any raster.

So, what does `RasterMath` really do ? The answer is as simple as the question : the user only works with the array, so he doesn't have to manage the reading and writing process, the no-data management, the compression or the projection.

The objective of RasterMath is to **let the user only focus on his array-compatible function**, and to let RasterMath manage the raster part.

Go to RasterMath documentation and examples

ai

The machine learning module is natively built to manage algorithm from the `scikit-learn` using state of the art methods and good pratices (such as standardizing the input data). `SuperLearner` class optimizes the fit process by a grid search. There is also a Sequential Feature Selection protocol which supports number of components (e.g. a single-date image is composed of four bands, i.e. 4 features, so you want to select the 4 features at once).

Go to SuperLearner documentation and examples

Cross-validation

`Museo ToolBox` produces only stratified cross-validation, which means the split is made by respecting the size per class and not for the whole dataset. For example the Leave-One-Out method will keep one sample of validation per class. As stated by (Olofsson et al., 2014) *"stratified random sampling is a practical design that satisfies the basic accuracy assessment objectives and most of the desirable design criteria"*. For spatial cross-validation, see (Karasiak et al., 2019) inspired from (Roberts et al., 2017).

`Museo ToolBox` offers two differents types of cross-validation :

Non-spatial cross-validation

- Leave-One-Out.
- Leave-One-SubGroup-Out.
- Leave-P-SubGroup-Out (Percentage of subgroup per class).
- Random Stratified K-Fold.

Spatial cross-validation

- Spatial Leave-One-Out (Karasiak et al., 2019).
- Spatial Leave-Aside-Out.
- Spatial Leave-One-SubGroup-Out (using centroids to select one subgroup and remove other subgroups for the same class inside a specified distance buffer).

Go to cross-validation documentation and examples

Acknowledgements

I acknowledge contributions from Mathieu Fauvel, beta-testers (hy Yousra Hamrouni !) and my thesis advisors : Jean-François Dejoux, Claude Monteil and David Sheeren.

Figures

A figure presents how Museo ToolBox is organized per module.

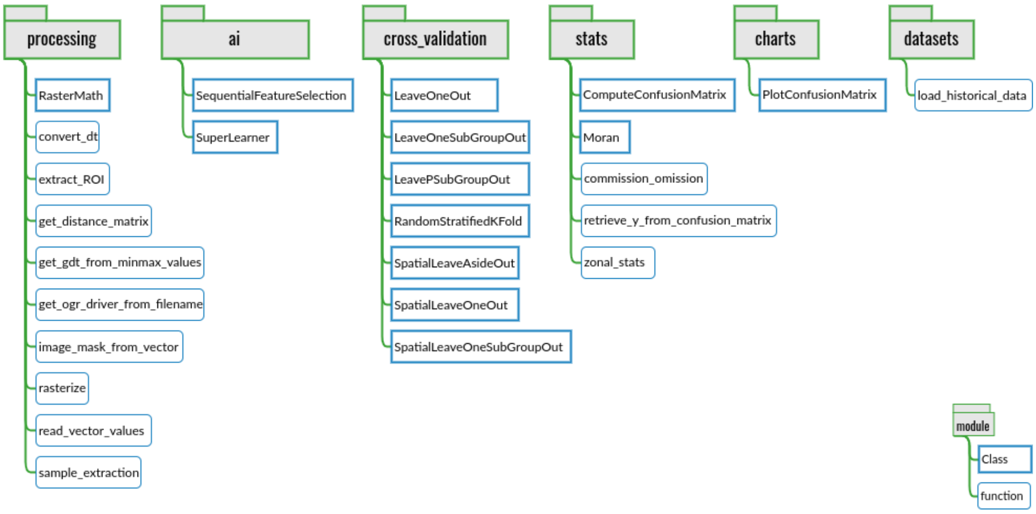

Figure 1: Museo ToolBox schema.

A figure explains how RasterMath manages reading and writing rasters.

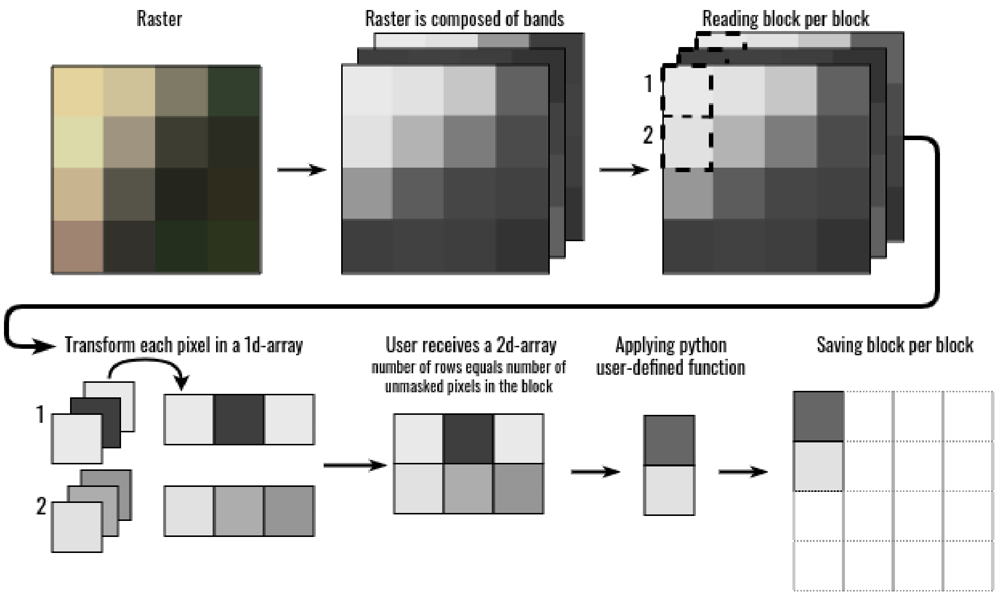

Figure 2: RasterMath under the hood

References

Karasiak, N., Dejoux, J.-F., Fauvel, M., Willm, J., Monteil, C., & Sheeren, D. (2019). Statistical stability and spatial unstability in prediction of forest tree species using satellite image time series. *Remote Sensing*. doi:10.3390/rs11212512

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. doi:10.1016/j.rse.2014.02.015

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. doi:10.1111/ecog.02881