Mappatura di variabili forestali basata sul disegno e stima statistica dell'errore per pixel

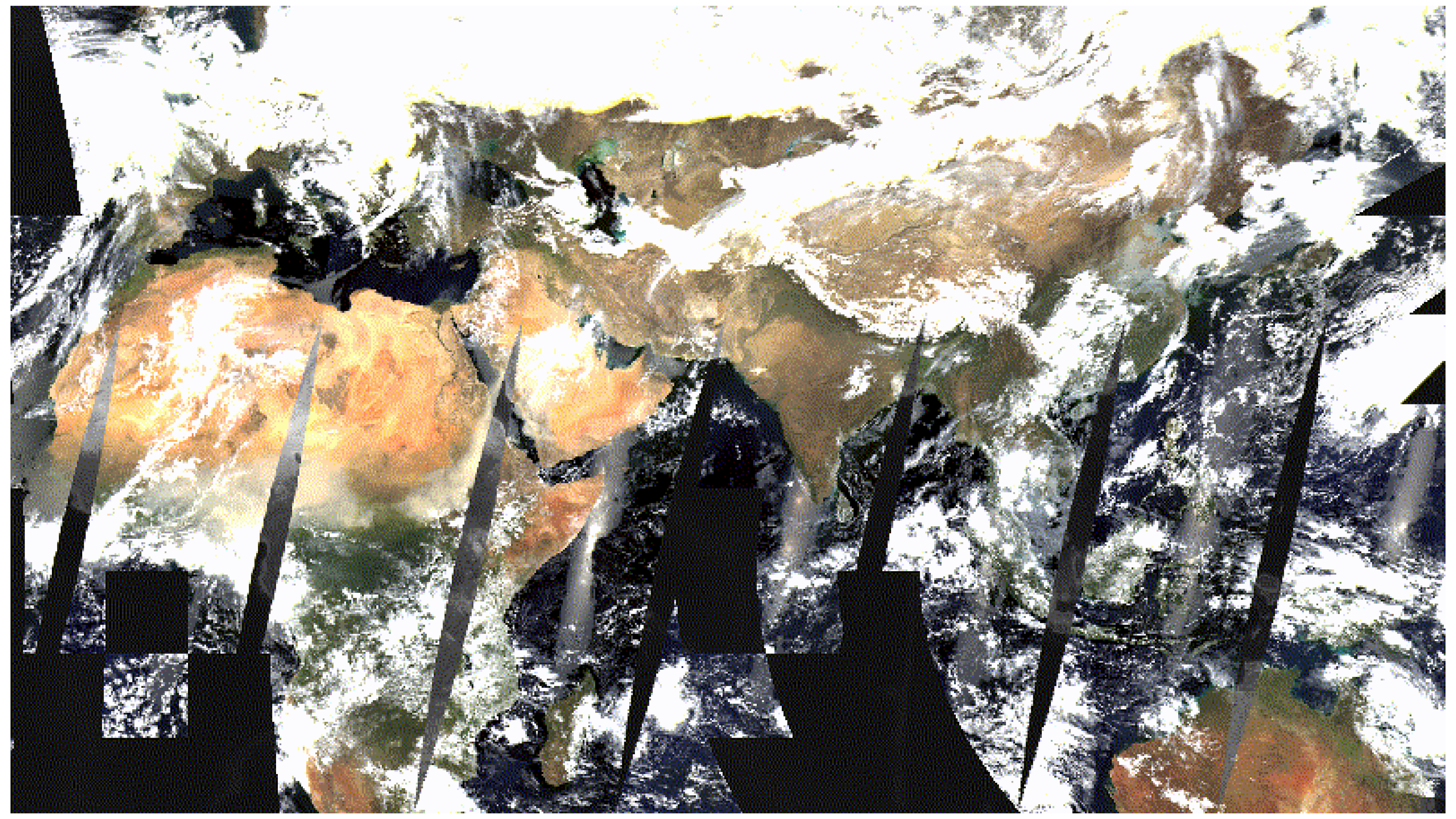
Rosa Maria Di Biase

Joint paper with
Agnese Marcelli, Saverio Francini, Gherardo Chirici, Piermaria Corona, Lorenzo Fattorini

Università degli Studi della Tuscia
Università di Siena 1240
MULTIFOR Multi-scale observations to predict Forest response to pollution and climate change
National Biodiversity Future Center

rosa.dibiase@unisi.it

Introduction



- Mapping forest attributes at pixel level with **per-pixel uncertainty** estimation remains a primary challenge in forest remote sensing

- Huge efforts have been undertaken in recent years to improve the reliability of remote sensing products for providing **statistically rigorous estimates**

- Maps derived from remote sensing products are increasingly accurate and easy to provide, **however they may lack reliable quality indicators**

Clearer and more transparent map validation remains a major challenge in remote sensing for forestry
(Fassnacht et al. 2023)

dataDriven open source tool

- A new statistical model-assisted method for mapping forest attributes and their uncertainty at pixel level
- Sentinel-2 remote sensing data are exploited as auxiliary information combined with field measurements of forest attributes

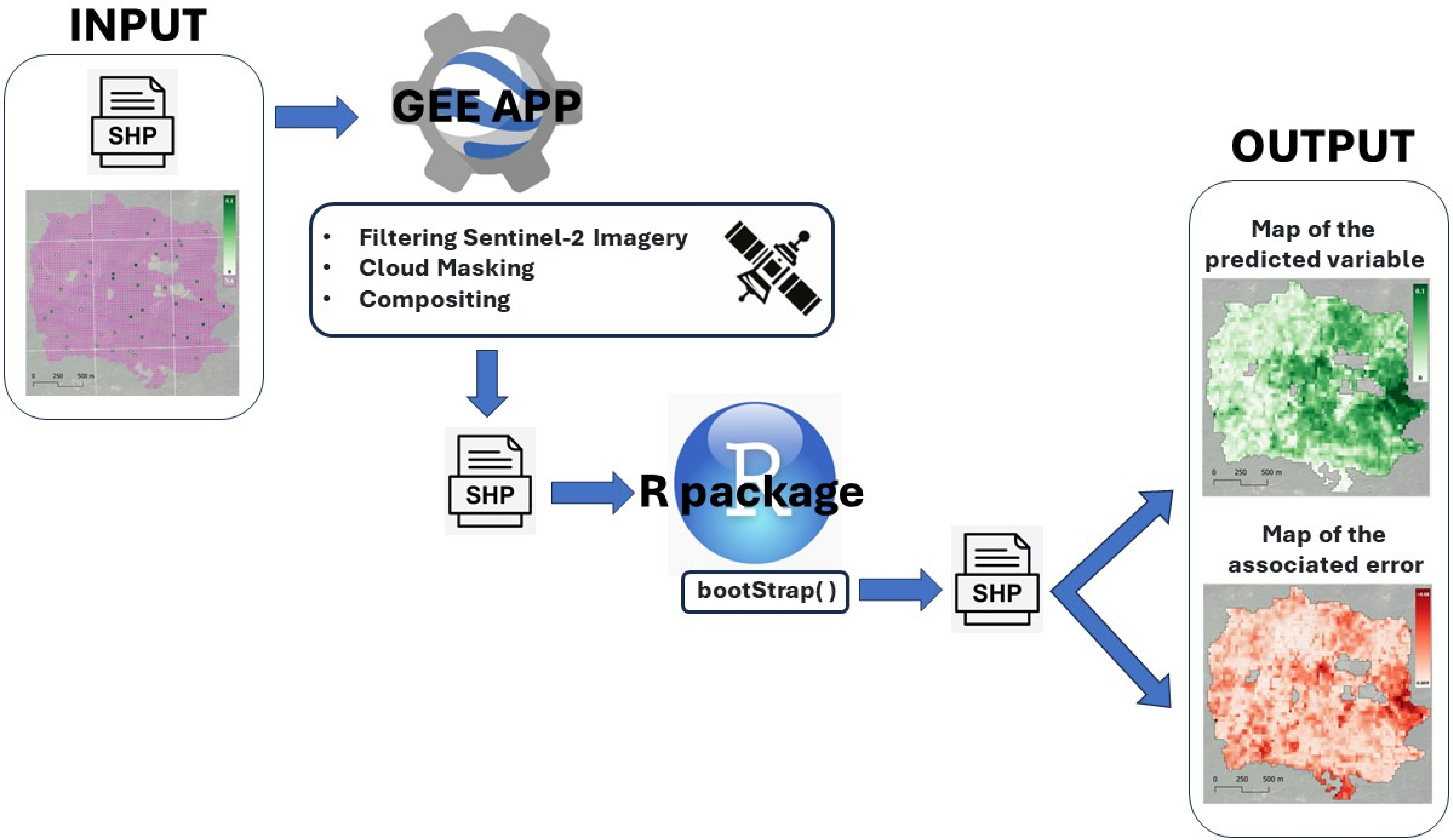
two main steps:

1) **Google Earth Engine (GEE) application** for computing Sentinel-2 predictors as auxiliary information

2) **R-package** for implementing the statistical data-driven procedure using the Sentinel-2 predictors downloaded with the GEE app

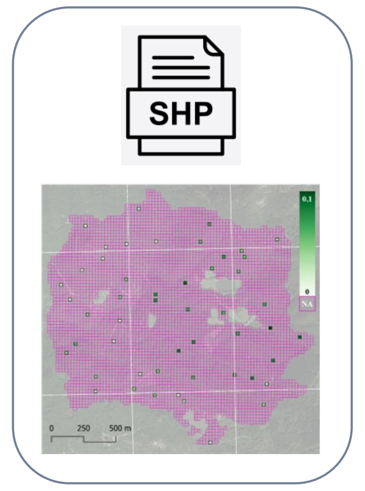
The user is guided from the preliminary choice of the assisting model to the final map and the estimation of its precision

dataDriven workflow



Input data

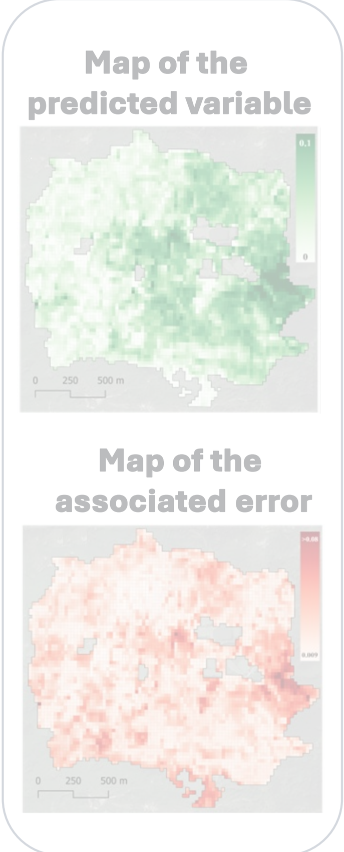
INPUT



GEE APP

R package

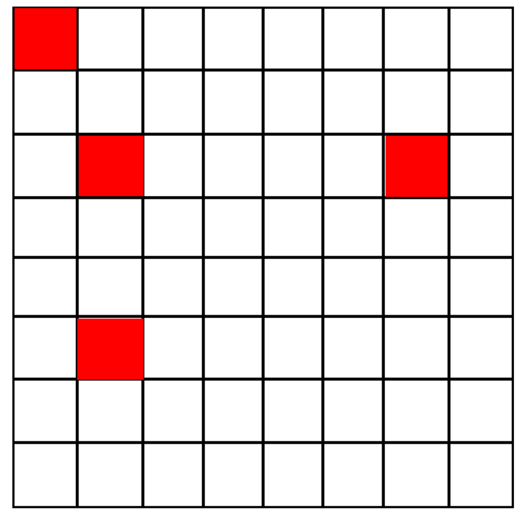
OUTPUT



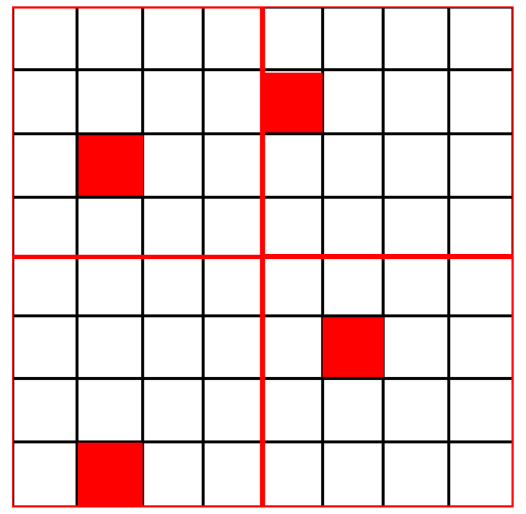
- The study area is tessellated into a population of N pixels

- A sample of $n < N$ pixels is selected to acquire reference data according to one of the following probabilistic sampling schemes:

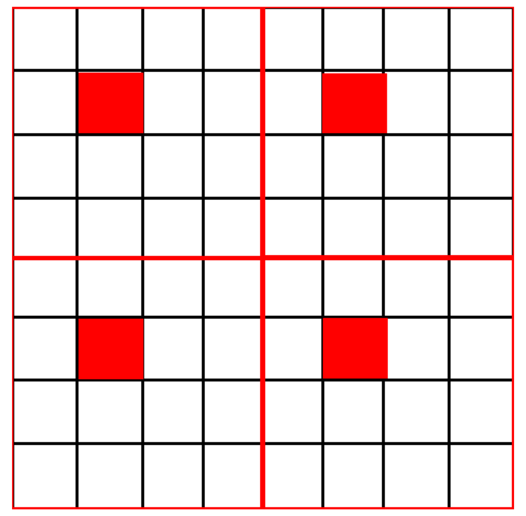
  1. Simple Random Sampling Without Replacement (SRSWoR)



  2. One Per Stratum Stratified Sampling (OPSS)



  3. Systematic Sampling (SYS)

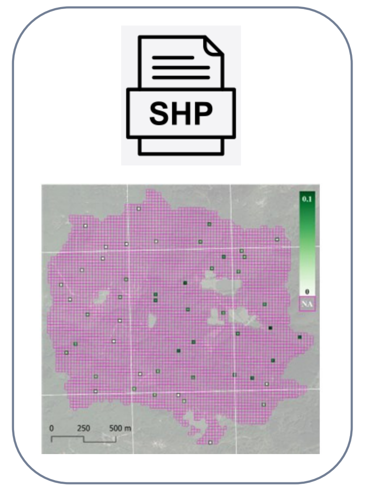


INPUT file → shapefile including two attributes:
(A) ID of the block to which each pixel belongs
(B) attribute of interest measured in the field within sampled pixels (NA for non sampled pixels)

SHP

Input data: Rincine case study

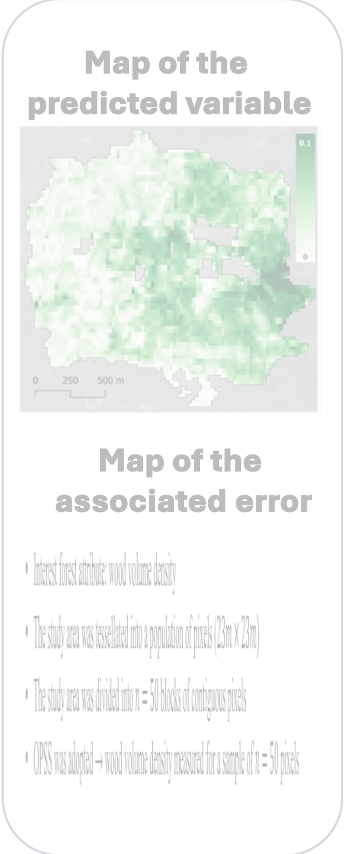
INPUT



GEE APP

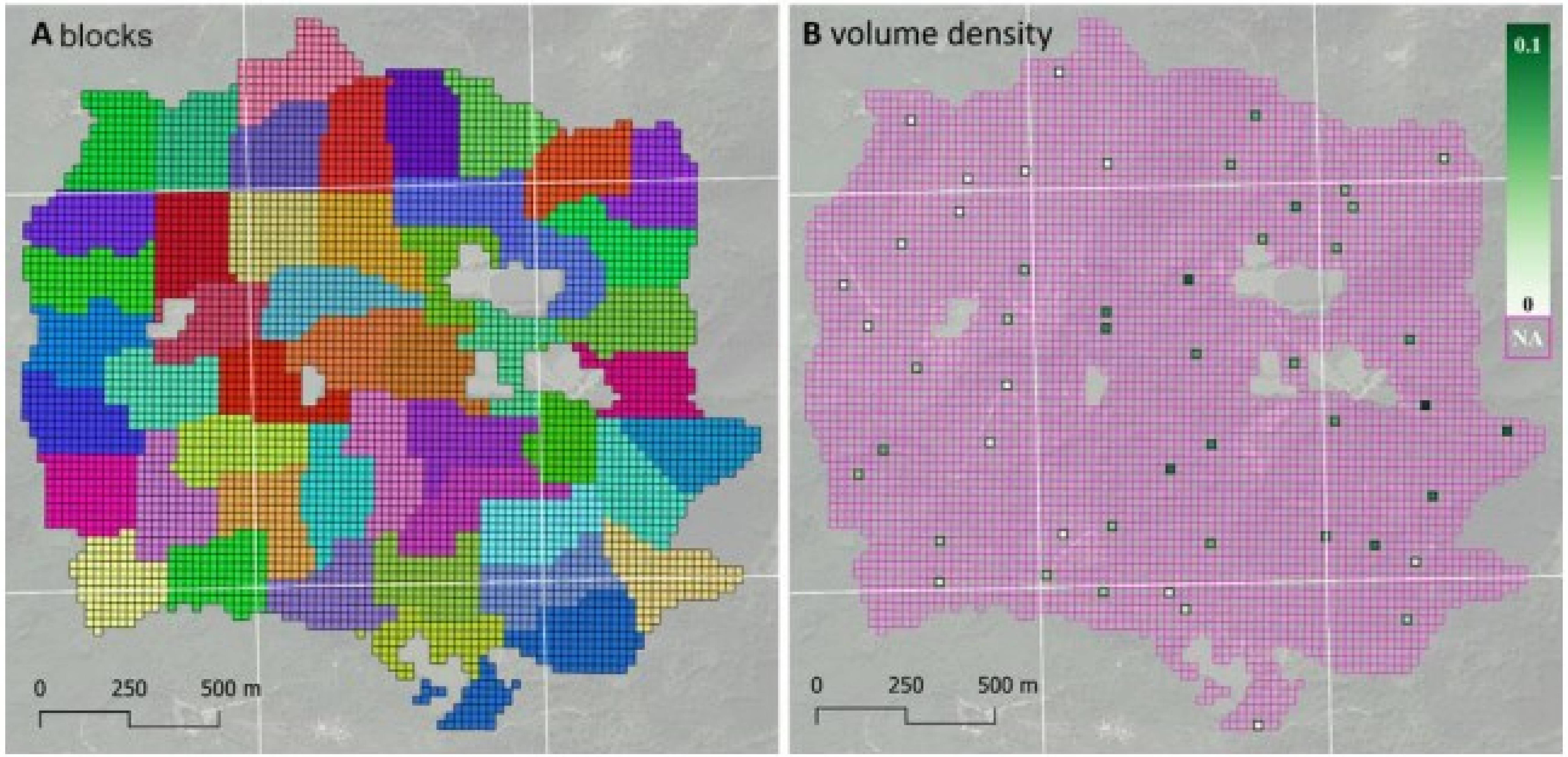
R package

OUTPUT



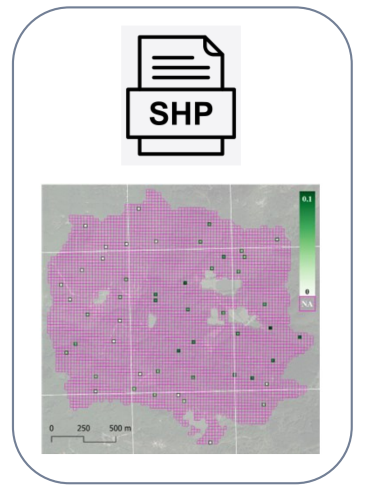
- Interest forest attribute: wood volume density
- The study area was tessellated into a population of pixels ($23m \times 23m$)
- The study area was divided into $n = 50$ blocks of contiguous pixels
- OPSS was adopted → wood volume density measured for a sample of $n = 50$ pixels

INPUT SHAPEFILE:



Step 1: Sentinel-2 predictors

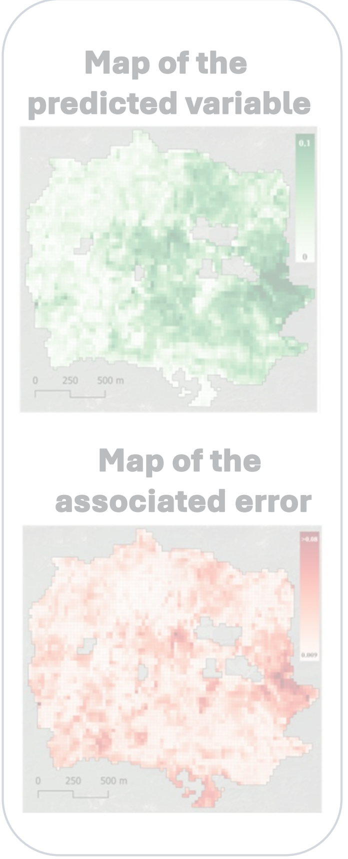
INPUT



GEE APP

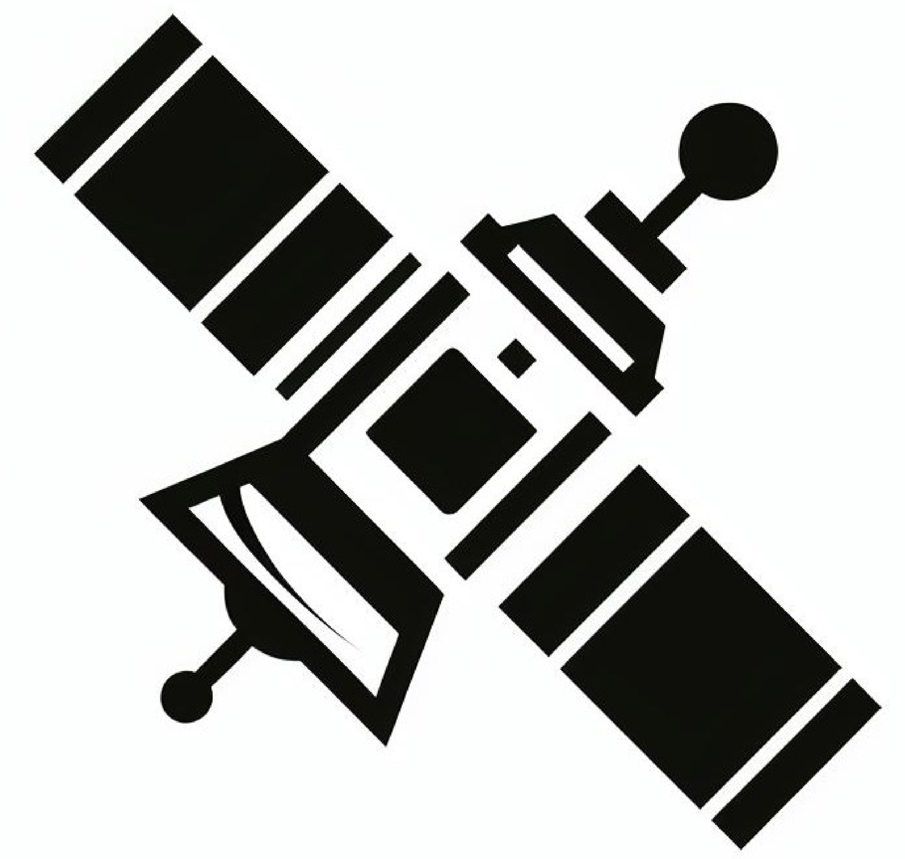
R package

OUTPUT



GEE application exploits:

- visible (blue, green, and red) bands (resolution 10 m)
- near-infrared (nir) bands (resolution 10 m)
- red edge (redE1, redE2, redE3, redE4) bands (resolution 20 m)
- short-wave infrared (swir1, swir2) bands (resolution 20 m)



Three main steps:

1. Filtering of Sentinel-2 imagery
2. Cloud masking →
   - Sentinel-2 cloud probability dataset (Skakun et al., 2022)
   - Final cloud composites generated for the selected years as the *medoid* of all remaining valid observations
3. Calculation of satellite pixel-based composites

GEE app available at: *https://code.earthengine.google.com/?accept_repo=users/saveriofrancini/PRIN*

Step 1. Sentinel-2 predictors

Period to select Sentinel-2 images yearly and construct pixel based medoid composites

Maximum percentage of clouds in the images

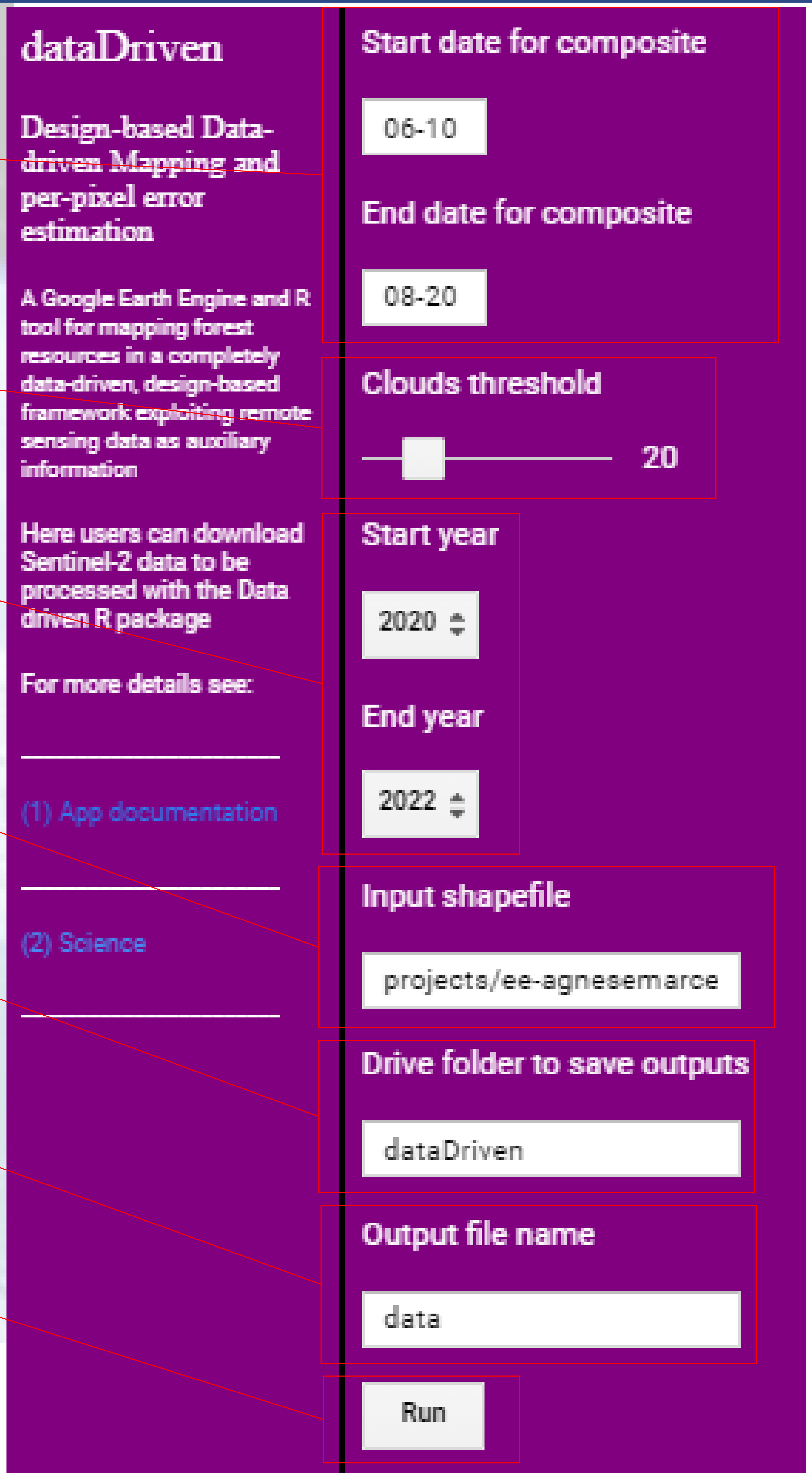
Years for which constructing Sentinel-2 medoid predictors

Input shapefile of the study area

Name of the folder will be generated on user drive

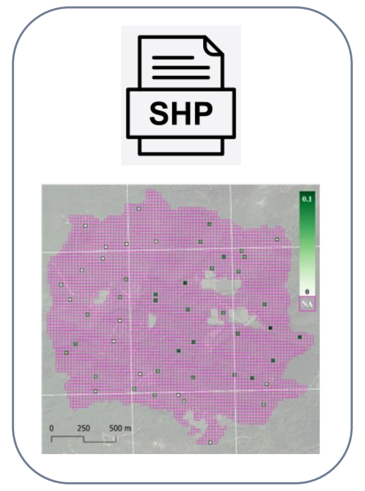
Name of the output file will be generated

Execute the app and get the predictors



Step 1. Sentinel-2 predictors: Rincine case study

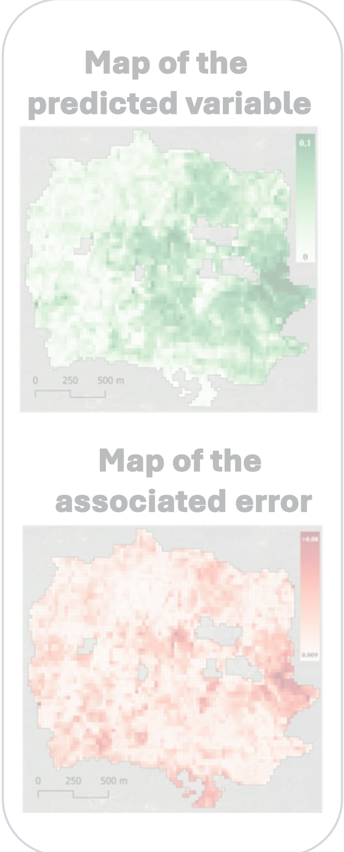
INPUT

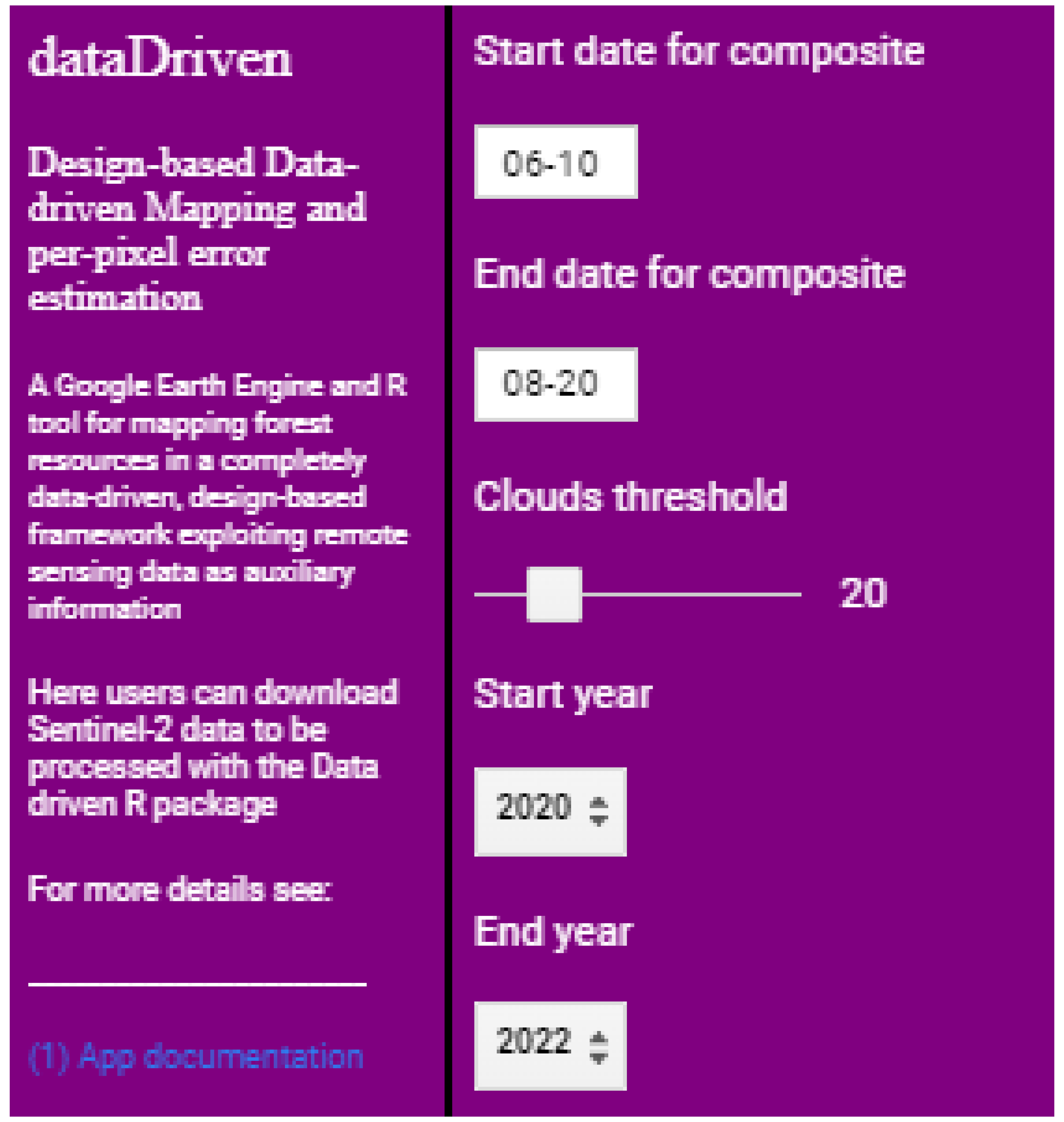


GEE APP

R package

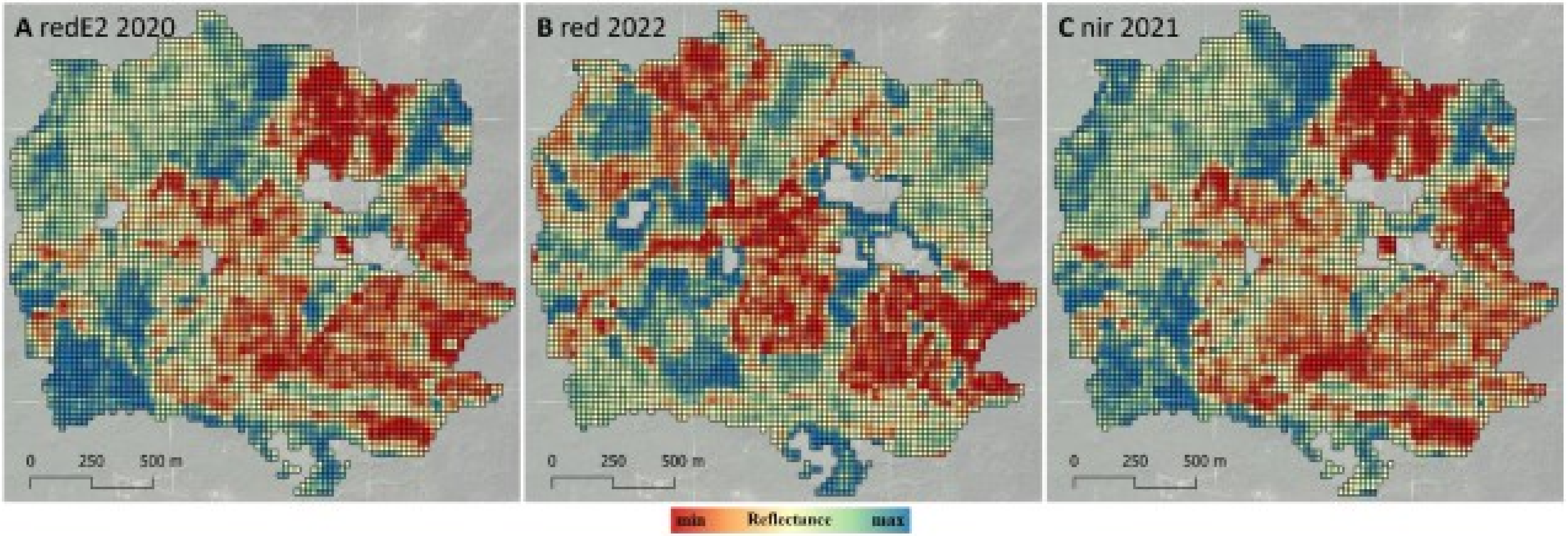
OUTPUT





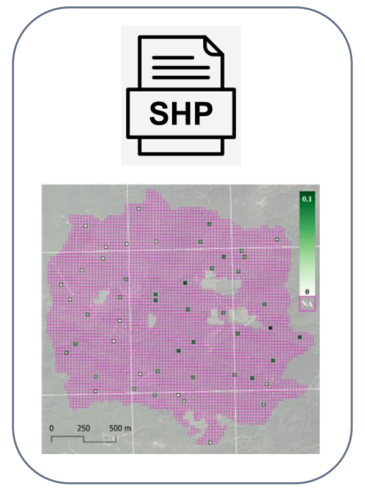
- ID of the block for each population pixel
- wood volume density for sampled pixels visited in the field
- spatial coordinates for each population pixel
- **30 Sentinel-2 predictors for each population pixel**

Example of 3 out of the 30 Sentinel-2 predictors downloaded from GEE



Step 2. statistical data-driven procedure

INPUT



R package implementing the statistical data-driven procedure described in **Di Biase et al. (2022)**

GEE APP

R package

Environmetrics
The official journal of The International Environmetrics Society
An Association of The International Statistical Institute

DOI: 10.1002/env.2750

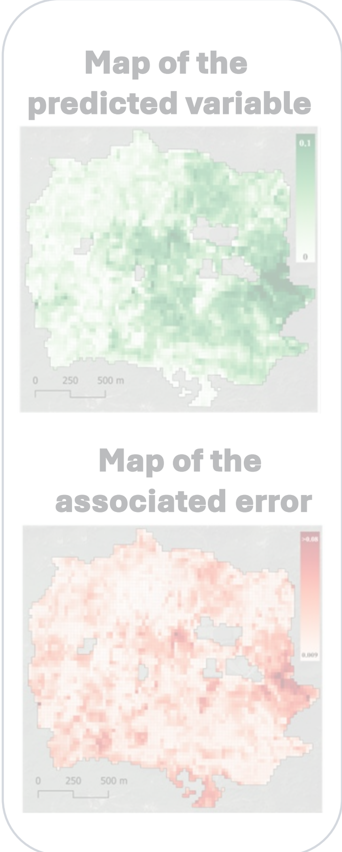
RESEARCH ARTICLE

WILEY

From model selection to maps: A completely design-based data-driven inference for mapping forest resources

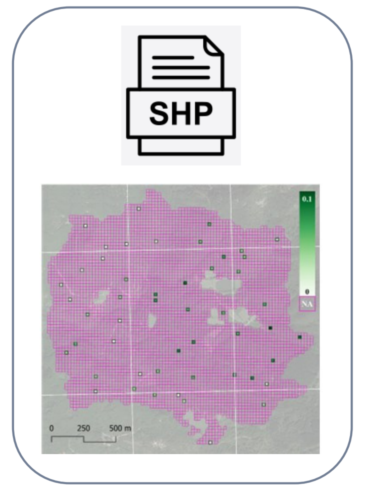
Rosa Maria Di Biase[1,2] | Lorenzo Fattorini[3] | Sara Franceschi[3] | Mirko Grotti[4] | Nicola Puletti[2] | Piermaria Corona[2]

OUTPUT



Step 2. statistical data-driven procedure

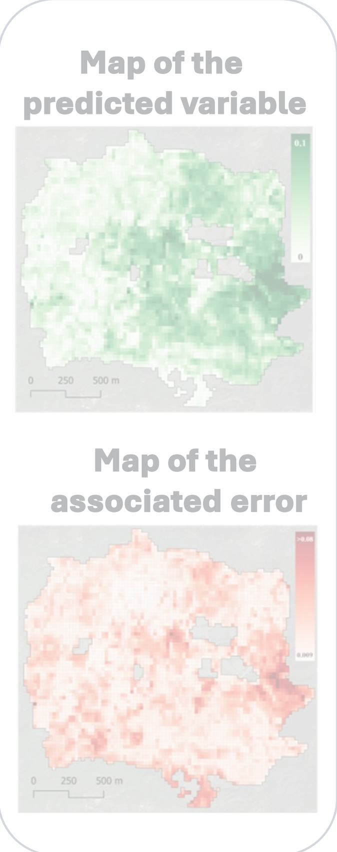
INPUT



GEE APP

R package

OUTPUT



R package implementing the statistical data-driven procedure described in **Di Biase et al. (2022)**

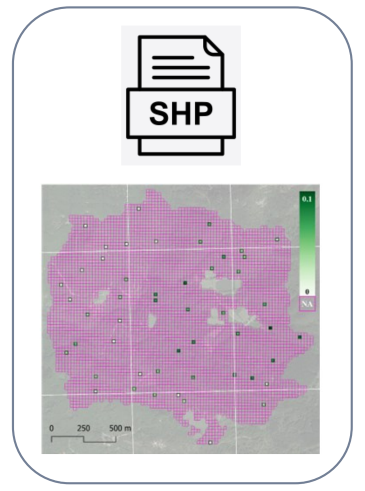
Main steps:

1. Akaike-type criterion for model selection, removing Sentinel-2 predictors poorly correlated with the attribute of interest or strongly correlated with each other

   Strongly correlated auxiliary variables can induce instability in the model or they can exhibit a poor prediction capability when weakly correlated to the interest variable

Step 2. statistical data-driven procedure

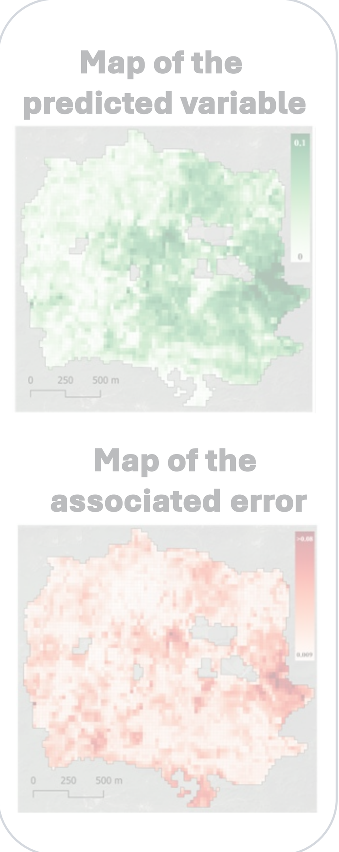
INPUT



GEE APP

R package

OUTPUT



R package implementing the statistical data-driven procedure described in **Di Biase et al. (2022)**

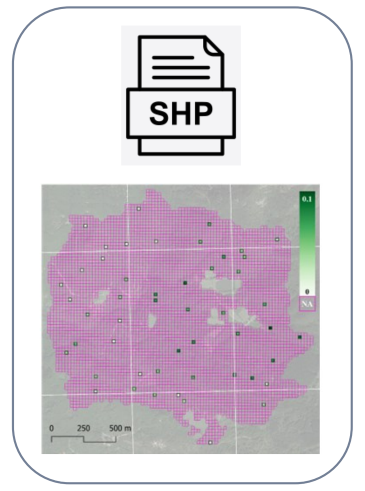
Main steps:

1. Akaike-type criterion for model selection, removing Sentinel-2 predictors poorly correlated with the attribute of interest or strongly correlated with each other
2. Least-squares criterion for the prediction of the values of the interest attribute within the pixels as a linear function of Sentinel-2 predictors selected in step1

- Densities can be expressed as $f_j = \boldsymbol{\beta}^t \boldsymbol{x}_j + \varepsilon_j$
- The **OLS** estimator for $\boldsymbol{\beta}$ is given by: $\mathbf{b} = \left[\sum_{j \in U} \mathbf{x}_j \mathbf{x}_j^{T} \right]^{-1} \sum_{j \in U} f_j \mathbf{x}_j$
- Using HT, $\hat{\mathbf{b}} = \left[\sum_{j \in S} \frac{\mathbf{x}_j \mathbf{x}_j^{T}}{\pi_j} \right]^{-1} \sum_{j \in S} \frac{f_j \mathbf{x}_j}{\pi_j}$
- The residuals are $e_j(\hat{\boldsymbol{b}}) = f_j - \hat{\boldsymbol{b}}^t \boldsymbol{x}_j$ in sampled units, while they are not known in the unsampled ones
- The use of an assisting model is the sole way to solve this problem

Step 2. statistical data-driven procedure

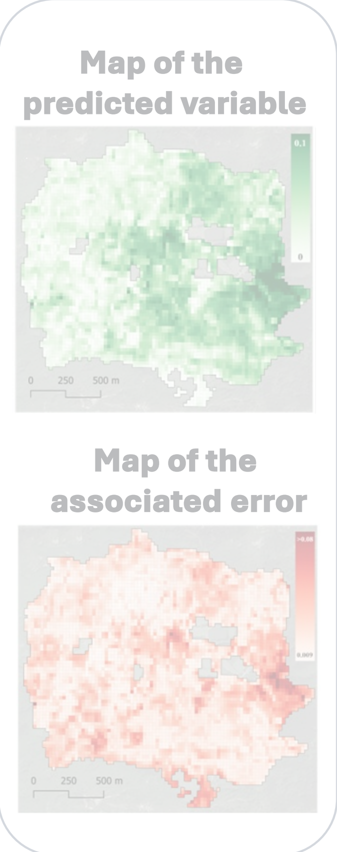
INPUT



GEE APP

R package

OUTPUT



R package implementing the statistical data-driven procedure described in **Di Biase et al. (2022)**

Main steps:

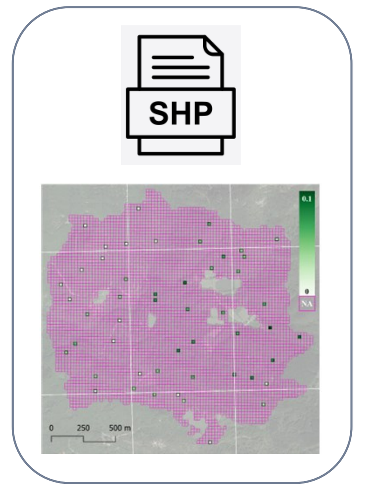
1. Akaike-type criterion for model selection, removing Sentinel-2 predictors poorly correlated with the attribute of interest or strongly correlated with each other
2. Least-squares criterion for the prediction of the values of the interest attribute within the pixels as a linear function of Sentinel-2 predictors selected in step1
3. Inverse distance weighting (IDW) interpolator for interpolation of the residuals in non-sampled pixels
   - Using the **IDW interpolator**

$$\hat{e}_j(\hat{\boldsymbol{b}}) = Z_j e_j(\hat{\boldsymbol{b}}) + (1 - Z_j)\sum_{i=1}^{N} w_{ij}(\alpha) e_j(\hat{\boldsymbol{b}})$$

where Z_j is an indicator variable, $w_{ij}(\alpha) = \frac{Z_i {d_{ij}}^{-\alpha}}{\sum_{i=1}^{N} Z_l {d_{lj}}^{-\alpha}}$ and $\alpha > 2$ the smoothing parameter

Step 2. statistical data-driven procedure

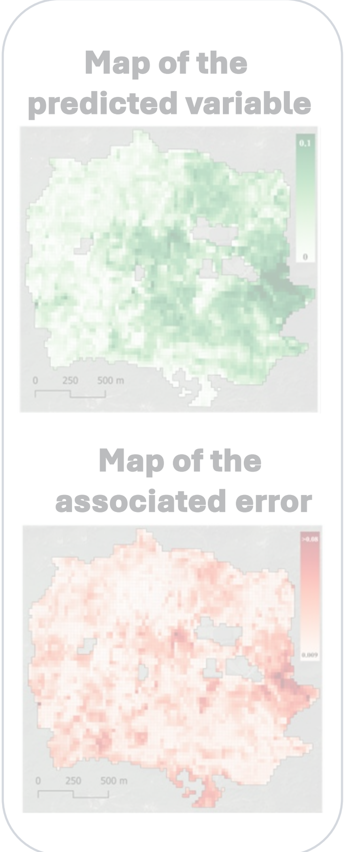
INPUT



GEE APP

R package

OUTPUT



R package implementing the statistical data-driven procedure described in **Di Biase et al. (2022)**

Main steps:

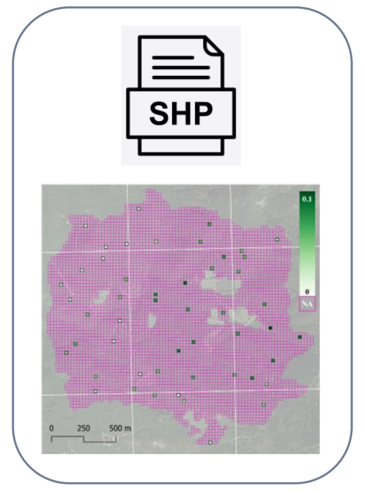
1. Akaike-type criterion for model selection, removing Sentinel-2 predictors poorly correlated with the attribute of interest or strongly correlated with each other
2. Least-squares criterion for the prediction of the values of the interest attribute within the pixels as a linear function of Sentinel-2 predictors selected in step1
3. Inverse distance weighting (IDW) interpolator for interpolation of the residuals in non-sampled pixels

The attribute of interest within pixels is estimated by summing predictions and interpolated residuals

- The resulting interpolated value for the density of unit j is $\hat{f}_j(\widehat{\boldsymbol{b}}) = \widehat{\boldsymbol{b}}^t \boldsymbol{x}_j + \hat{e}_j(\widehat{\boldsymbol{b}})$

Step 2. statistical data-driven procedure

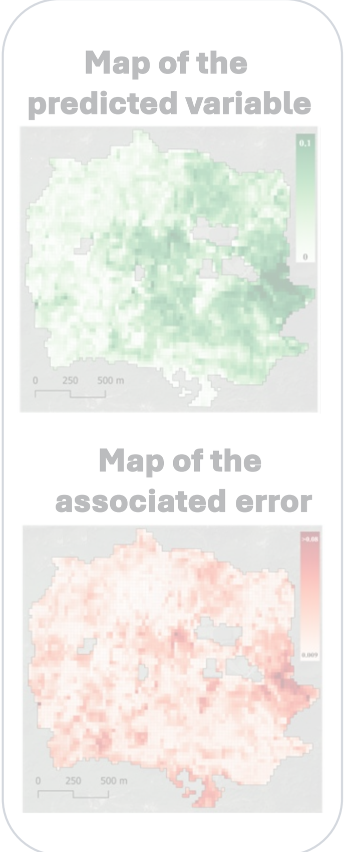
INPUT



GEE APP

R package

OUTPUT



R package implementing the statistical data-driven procedure described in **Di Biase et al. (2022)**

Main steps:

1. Akaike-type criterion for model selection, removing Sentinel-2 predictors poorly correlated with the attribute of interest or strongly correlated with each other
2. Least-squares criterion for the prediction of the values of the interest attribute within the pixels as a linear function of Sentinel-2 predictors selected in step1
3. Inverse distance weighting (IDW) interpolator for interpolation of the residuals in non-sampled pixels

The attribute of interest within pixels is estimated by summing predictions and interpolated residuals

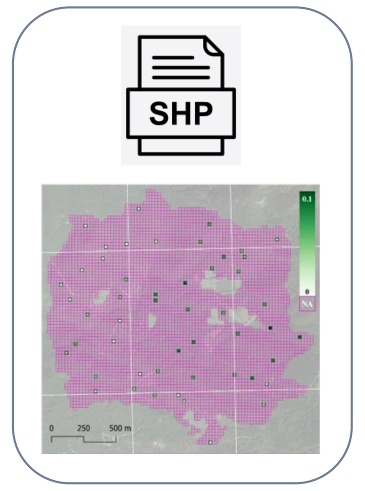
4. Harmonization of the map of the interest attribute to match HT total estimates and map total estimates

- To obtain non-discrepant results, the estimated map is harmonized rescaling density estimates

$$\tilde{f}_j(\hat{\boldsymbol{b}}) = \frac{\hat{T}_{reg}}{\hat{T}(\hat{\boldsymbol{b}})}\hat{f}_j(\hat{\boldsymbol{b}})$$

Step 2. statistical data-driven procedure

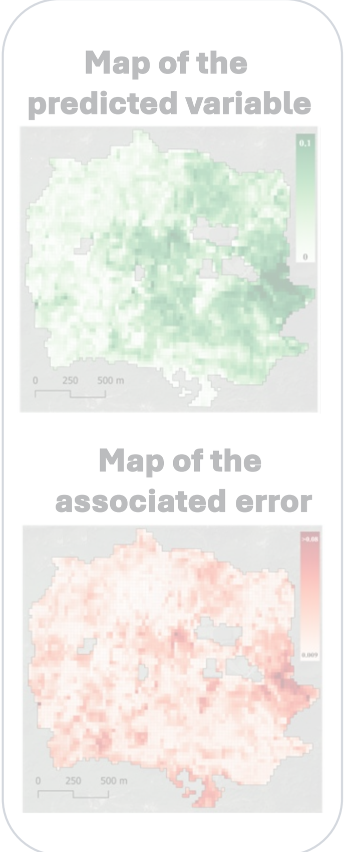
INPUT



GEE APP

R package

OUTPUT



R package implementing the statistical data-driven procedure described in **Di Biase et al. (2022)**

Main steps:

1. Akaike-type criterion for model selection, removing Sentinel-2 predictors poorly correlated with the attribute of interest or strongly correlated with each other
2. Least-squares criterion for the prediction of the values of the interest attribute within the pixels as a linear function of Sentinel-2 predictors selected in step1
3. Inverse distance weighting (IDW) interpolator for interpolation of the residuals in non-sampled pixels

The attribute of interest within pixels is estimated by summing predictions and interpolated residuals

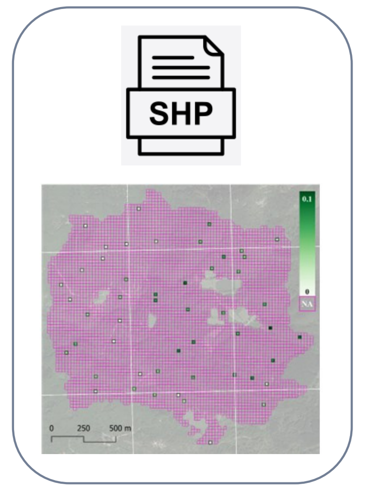
4. Harmonization of the map of the interest attribute to match traditional total estimates and map total estimates
5. Pseudo-population bootstrap procedure for achieving the map of the estimated precision

Steps 1-4 replicated *B* times

•bootStrap() which builds both the shapefile of the estimated attribute and the shapefile of the associated error.

Step 2. statistical data-driven procedure

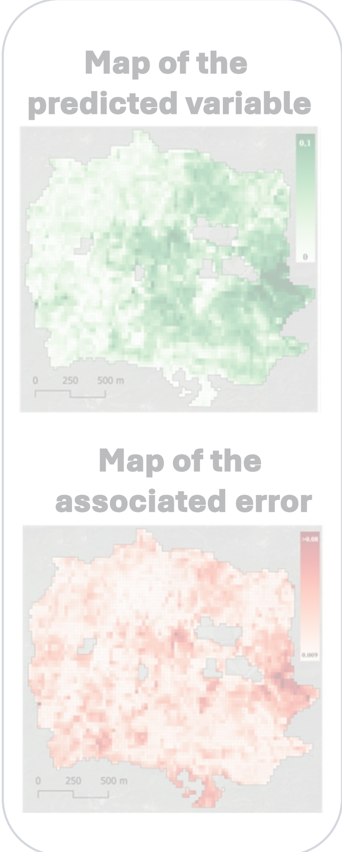
INPUT



GEE APP

R package

OUTPUT



R package available at *https://github.com/saveriofrancini/dataDriven*

main function

bootStrap( ):

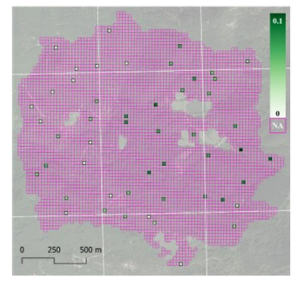
Arguments

`cluster`	The ID strata of the population spatial units. (If simple random sampling is adopted, set cluster = 1)
`inFile`	The path for the input dbf file. By default, it is automatically selected a test file provided within the package. To see it use system.file("data", "data.dbf", package = "dataDriven")
`depVar`	The name of the interest variable column
`varsToRemove`	The name of the columns to do not consider in the analysis (if there are any)
`coordinates`	The spatial coordinates of the population areal units
`resampling`	The number of bootstrap resampling
`outDir`	The directory to save the output shapefile

Final output → SHP shapefile object containing both the estimated forest attribute and the associated error for each population pixel

Final output: Rincine case study

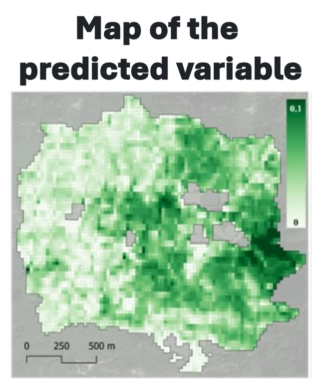
INPUT

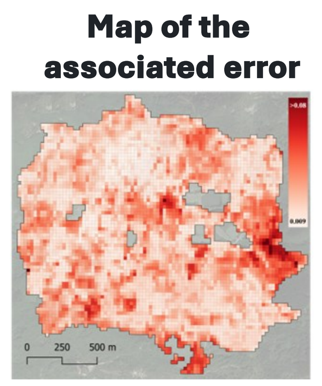


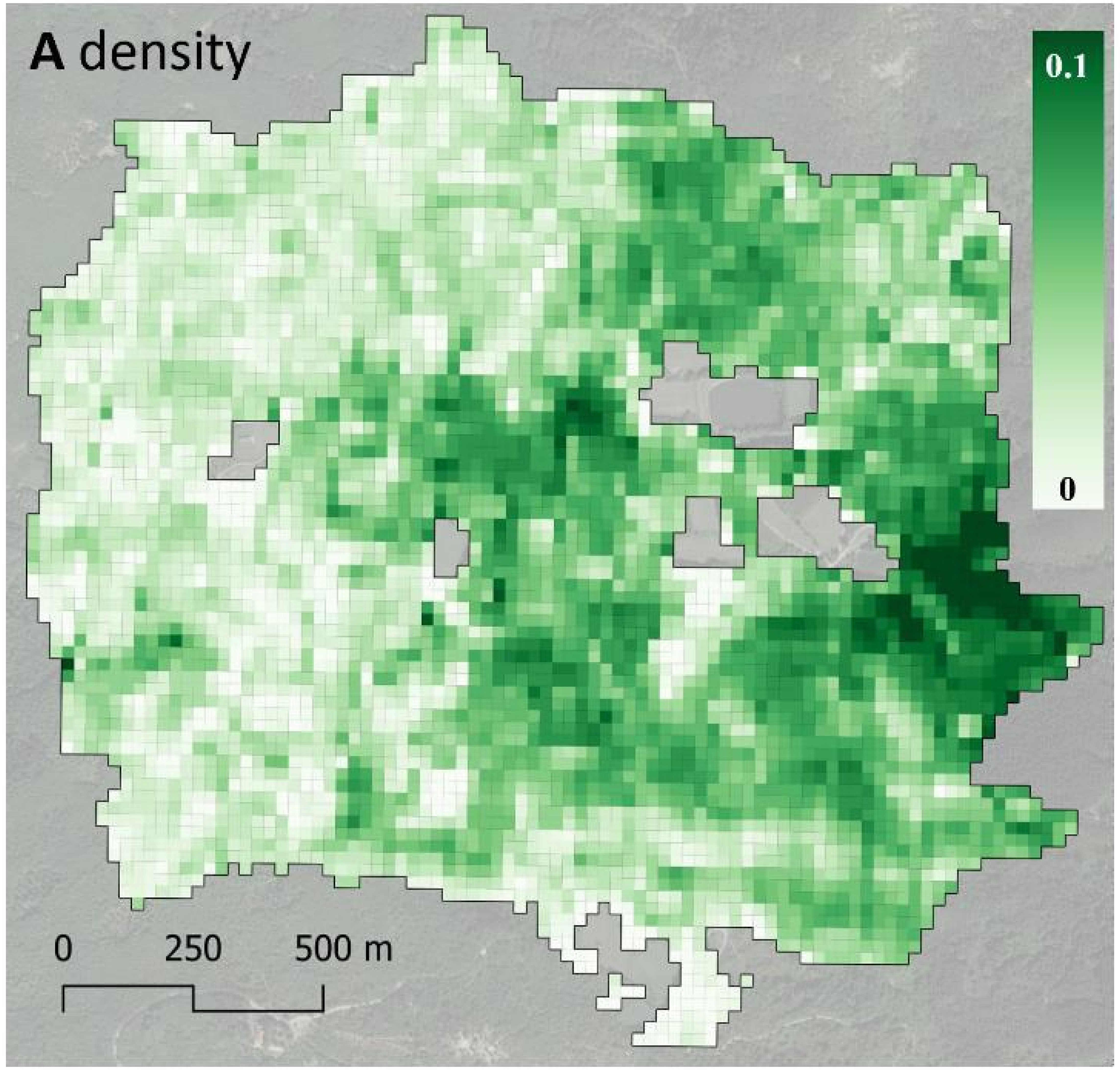
GEE APP

R package

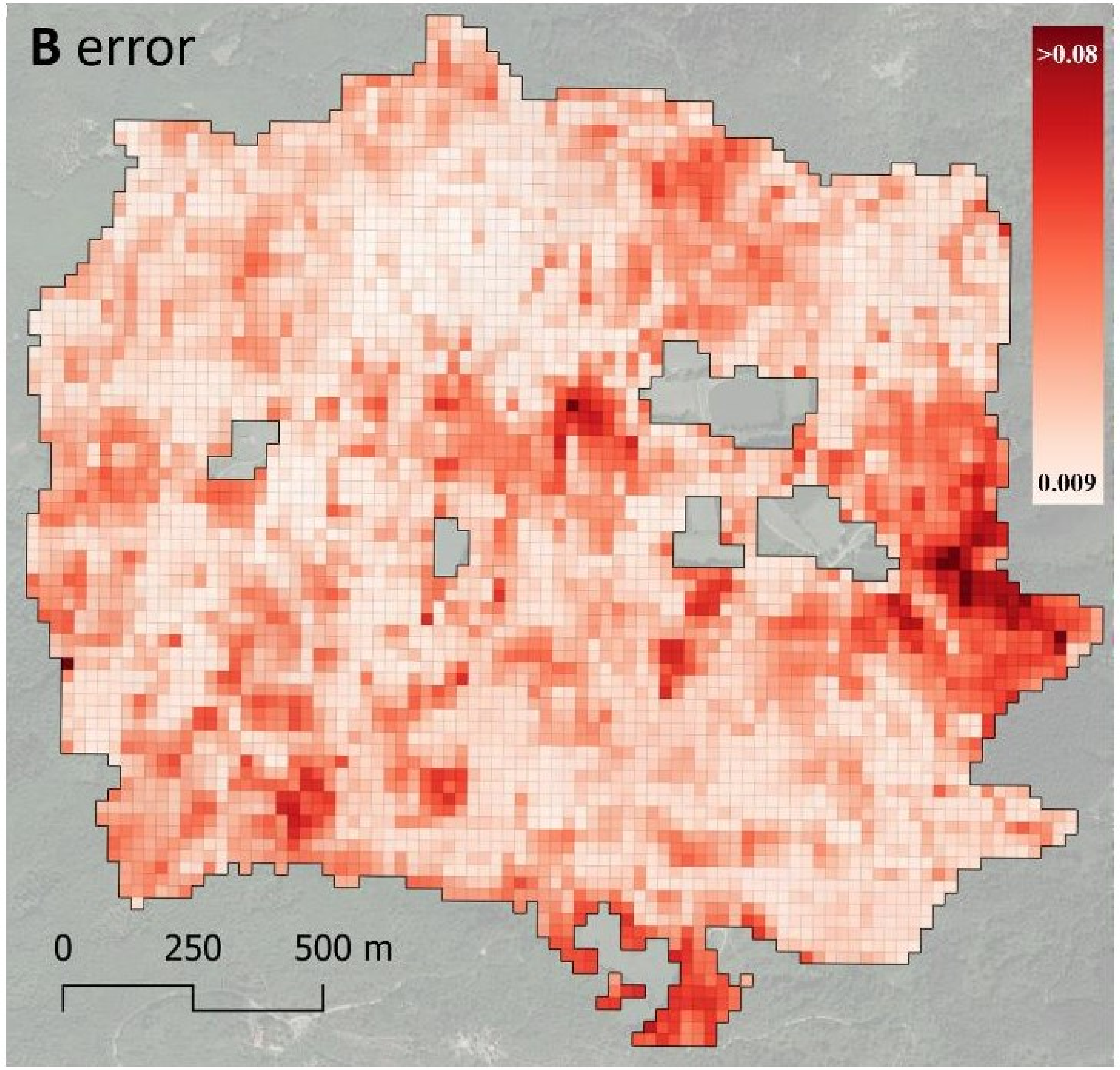
OUTPUT







A. Estimated wood volume density map



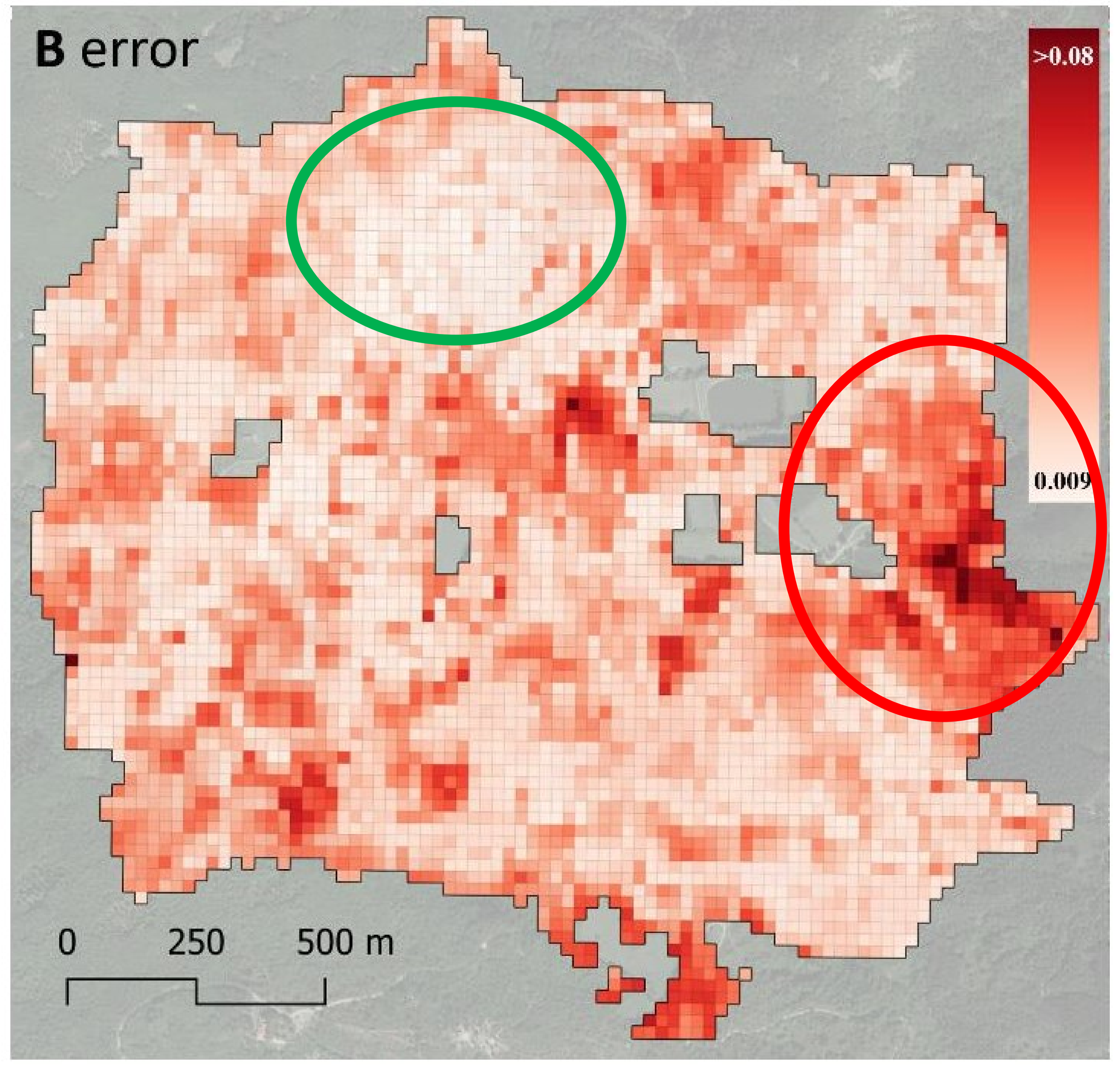
B. Estimated error map

Conclusions

The novel aspect provided by dataDriven is the ability to produce error estimates for each pixel in the map

Map uncertainty

- informs about areas where the map estimates are unreliable and areas in which the information provided by the map is trustworthy



- constitutes a support not only from an analytical point of view but also as a powerful communication tool

References

- Di Biase, R.M., Fattorini, L., Franceschi, S., Grotti, M., Puletti, N., Corona, P. (2022). From model selection to maps: a completely design-based data-driven inference for mapping fores resources. Environmetrics, 33, e2750.

- Fassnacht, F.E., White, J. C., Wulder, M. A., & Næsset, E. (2023). Remote sensing in forestry: current challenges, considerations and directions. Forestry: An International Journal of Forest Research, cpad024

- Skakun, S., Wevers, J., Brockmann et al. (2022). Cloud Mask Intercomparison eXercise (CMIX): An evaluation of cloud masking algorithms for Landsat 8 Sentinel-2. Remote Sensing of Environment, 274, 112990

GEE app available at: *https://code.earthengine.google.com/?accept_repo=users/saveriofrancini/PRIN*
R-package available at: *https://github.com/saveriofrancini/dataDriven*

Thank you!

rosa.dibiase@unisi.it