A taxonomy for Random Forest in the spatial regression framework

L. Patelli[1], M. Cameletti[2], N. Golini[3] and R. Ignaccolo[3]

[1]University of Pavia; [2]University of Bergamo; [3]University of Torino.

GRASPA 2023 July 10-11, Palermo

Introduction

Data collected for environmental applications often contain spatial information. Focusing on geostatistical data, different models were proposed in the context of the *Data Modeling Culture* (DMC [1]), but their fitting can become computationally expensive, and in some cases even infeasible, when dealing with complex models, a large number of sites or predictors. In recent years, the *Algorithmic Modeling Culture* (AMC [1]) has gained popularity as an alternative to the DMC. Specifically, supervised Machine Learning and Deep Learning algorithms have emerged as non-parametric predictive techniques that do not require any assumptions about the response-predictors relationship. These algorithms are entirely data-driven, making them highly flexible and capable of modeling complex non-linear relationships. But are they suitable for spatially dependent data?

Aim

This work proposes a new taxonomy for the classification of strategies (described in scientific documents) that deal with the problem of adapting regression Random Forest for geostatistical data.

Standard Regression Random Forest

Random Forest (RF) is a supervised learning algorithm defined as an ensemble of bagged regression trees. RF uses a random subset of covariates during the splits for the construction of each tree [2].

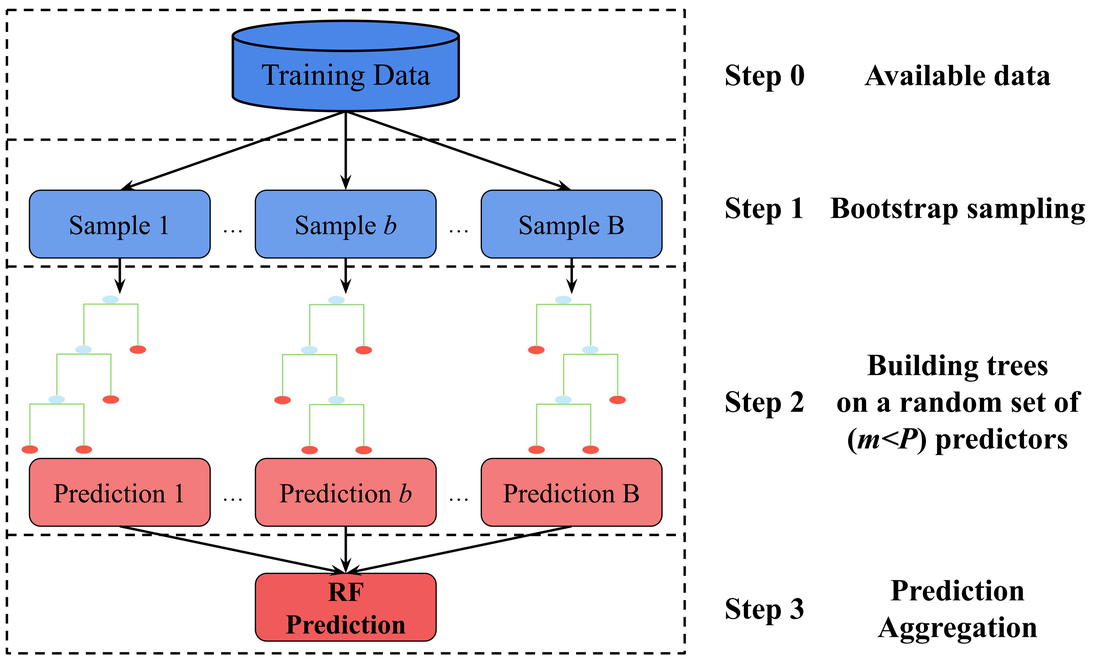

Figure 1: Graphical representation of the RF algorithm.

The limitations of the classic RF are:

- RF is unable to leverage information that is not included in the predictor set [3];
- the re-sampling of correlated data violates the independence assumption required by the standard bootstrap method to generate B datasets;
- the regression tree optimization problem can be written as an Ordinary Least Squares problem, which can lead to sub-optimal results in the presence of dependent data [4].

Taxonomy

We propose a new taxonomy [5] for classifying scientific documents related to the application of regression RF for spatially correlated data. We first define three main categories, according to when RF is adjusted (see the RF steps in Figure 1):

- Pre-processing (Pre): strategies undertaken at Step 0 by augmenting the available data by including predictors which are somehow informative of the spatial autocorrelation existing in the data;
- In-processing (In): strategies which perform a substantial change of RF which can happen at Step 1 (bootstrap sampling) or Step 2 (building trees);
- Post-processing (Post): strategies where the spatial correlation is taken into account after running RF, by adjusting the RF predictions.

Furthermore, there can be combinations of different strategies, which define mixed categories. The full taxonomy can then be represented by a Hasse Diagram of the power set $\mathcal{P}(S)$ composed of 8 elements, where $S = \{Pre, In, Post\}$ is the set of the pure categories (see Figure 2).

Literature review

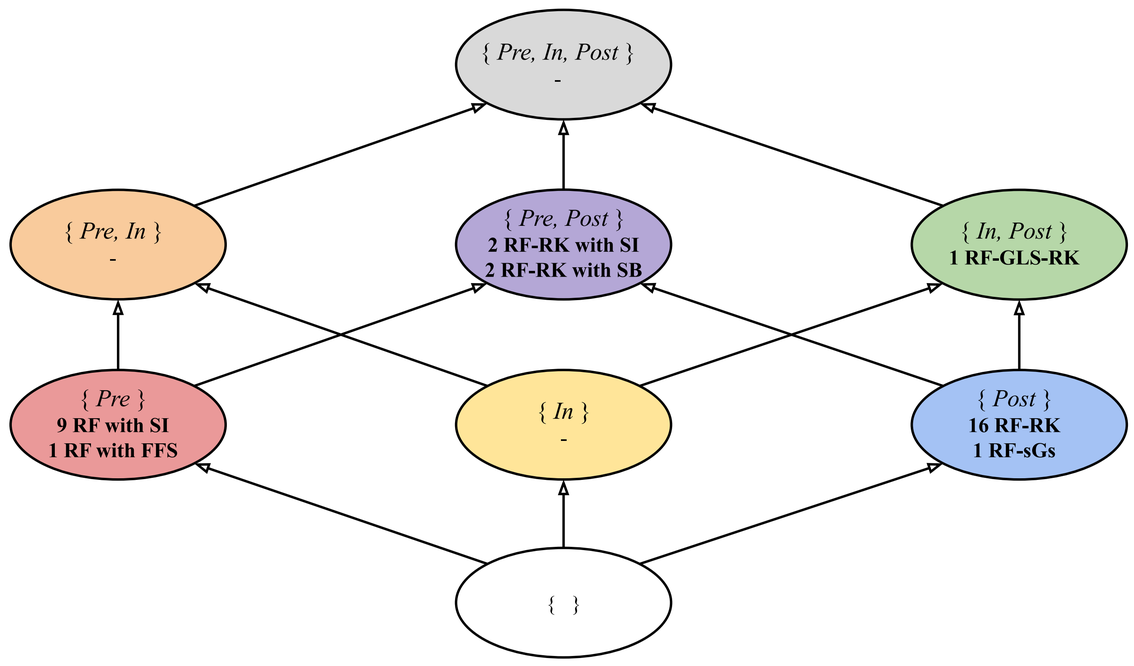

Figure 2: Graphical representation (Hasse diagram) of the power set $\mathcal{P}(S)$ with the strategies adopted in the 32 identified documents. The labels for the adopted strategy are:
RF with SI: Random Forest with Spatial Information;
RF with FFS: Random Forest with Forward Feature Selection;
RF-RK: Random Forest Residual Kriging;
RF-sGs: Random Forest sequential Gaussian simulation;
RF-RK with SI: Random Forest Residual Kriging with Spatial Information;
RF-RK with SB: Random Forest Residual Kriging with Spatial Bootstrap;
RF-GLS-RK: Random Forest based on Generalized Least Squares Residual Kriging.

The proposed taxonomy was used to classify the scientific documents obtained from a systematic literature review regarding the strategies adopted to adjust regression RF in a spatial framework. On the 25^{th} of October 2022 a query on *Scopus* and *Web of Science* was performed using geospatial keywords. By applying the PRISMA approach [6] 32 literature contributions strictly related to the topic were identified. The 32 documents were classified by using the proposed taxonomy and internally grouped based on the specific adopted strategy (see Figure 2). For the taxonomy paper and further details regarding the classified contributions, please refer to the following QR code.

Main references

[1]Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*. **16**, 199–231.
[2]Breiman, L. (2001). Random forests. *Machine Learning*. **45**, 5–32.
[3]Hengl, T., Nussbaum, M., Wright, M., Heuvelink, G. & Gräeler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ*. **6:e5518**.
[4]Saha, A., Basu, S. & Datta, A. (2023). Random forests for spatially dependent data. *Journal of the American Statistical Association*. **118:541**, 665–683.
[5]Patelli, L., Cameletti, M., Golini, N. & Ignaccolo, R. (2023). A path in regression Random Forest looking for spatial dependence: a taxonomy and a systematic review. *arXiv*.**2303.04693**.
[6]Page, M. J., McKenzie, J. E., Bossuyt, P. M. et al. (2021), The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, *Systematic reviews* **10**(1), 1–11.

Acknowledgments

This work was partially funded by Fondazione Cariplo within the "AgrImOnIA" project (https://agrimonia.net/).