Contribution of Random Forest and Deep Neural Network Algorithms with Environmental Covariates for the Spatial SOC Stock Modelling

Mounir Oukhattar, Sébastien Gadal, Yannick Robert, Catherine Keller

[ISBN : 978-2-910545-13-X, Mai 2025, Avignon]

Contribution of Random Forest and Deep Neural Network Algorithms with Environmental Covariates for the Spatial SOC Stock Modelling

Example of Aix-Marseille Provence Metropolis

Mounir Oukhattar [1,2], Sébastien Gadal [1,4], Yannick Robert [3], Catherine Keller [2]

1. Aix-Marseille Univ., CNRS, ESPACE UMR 7300, Univ., Nice Sophia Antipolis, Avignon Univ., 13545 Aix-en-Provence, France
mounir.oukhattar@etu.univ-amu.fr; sebastien.gadal@univ-amu.fr
2. Aix-Marseille Univ, CNRS, IRD, INRAE, CEREGE, Technopole de l'Environnement Arbois-Méditerranée, BP80, 13545 Aix-en-Provence, Cedex 4, France
keller@cerege.fr
3. Service Observatoire et lutte contre les pollutions Direction Expertise et Médiation environnementale Pôle Transition Ecologique et Energétique DGD Transition environnementale, Culture, Sport et Equipements, Métropole Aix-Marseille Provence, France
yannick.robert@ampmetropole.fr
4. Institute of Mathematical Computer Sciences, I.A. Remote Sensing Team, Vilnius University, Lithuania

ABSTRACT. This study aims to compare the effectiveness of random forest (RF) and deep neural network (DNN) algorithms in conjunction with variable selection methods for modelling soil organic carbon (SOC) stocks in the Aix-Marseille-Provence (AMP) Metropolis. Using a total of 51 soil samples and 29 different environmental factors, which include climate data, parent material, topographic details, land cover, human impact, remote sensing information, and soil characteristics, the examination has demonstrated that the Deep Neural Network (DNN) generally performs better than Random Forest (RF), particularly when all variables or those chosen through the Boruta technique or the variance inflation factor (VIF) are integrated. Nevertheless, RF excels over DNN when only non-redundant variables are considered. The resulting visual representations illustrate a higher SOC Stock in hilly forested regions and lower levels in maritime wetlands and farming areas. These conclusions offer a valuable tool for sustainable soil management and spatial organization, helping in achieving AMP Metropolis's climate goals.
KEYWORDS: SOC Stock; spatial modelling; RF; DNN; Environmental covariates; AMP Metropolis.

1. Introduction

Variations in global SOC stocks, impacted by both intensive agricultural practices and natural processes like organic matter decomposition, are of significant concern within

the scientific community (FAO, 2015; Prăvălie et al., 2021). Accurately modelling the spatial distribution of SOC is essential for efficient soil management and for optimizing sampling efforts (Radočaj et al., 2024), which can be costly and time-consuming. Intensive agriculture, erosion, and climate change alter carbon storage in soils, leading to reduced soil fertility and increased CO_2 emissions (Lal, 2004; Ontl & Schulte, 2012). Sophisticated algorithms like RF and DNN are renowned for their ability to accurately model SOC levels. RF is particularly adept at handling non-linear relationships (Pouladi et al., 2023), while DNN excels at capturing intricate patterns within datasets (Odebiri et al., 2022). An example of this is evident in a study by Odebiri et al. (2022), which showed that DNN outperformed RF by achieving a coefficient of determination (R^2) of 0.67 for national SOC mapping in South Africa. This study employed RF and DNN algorithms to model the spatial distribution of SOC stocks at a depth of 0–30 cm in the Aix-Marseille-Provence Metropolis. Data from 51 sampled sites were combined with environmental variables sourced from remote sensing, topo-climatic conditions, and soil physical properties. Various approaches were tested, incorporating methods such as Boruta and VIF for variable selection. The resulting maps generated from this research serve as a valuable tool for guiding sustainable soil management practices at the metropolitan level.

2. Study area

The AMP Metropolis is located in the southeastern part of France within the PACA region, as depicted in Figure 1. Comprising an expansive area of 3,173 square kilometres, with a perimeter spanning 773.6 kilometres, it includes a diverse landscape of dense urban centres, forests, agricultural plains, valleys, and plateaus. Natural areas make up 61% of the territory, while urban zones and agricultural land account for 15% and 24%, respectively. By 2019, the population had grown to 1.9 million inhabitants, concentrated in Marseille and Aix-en-Provence. The elevation in the area varies from -20 meters to 1,042 meters. The Mediterranean climate, characterised by abundant sunshine, sporadic rainfall, and prevailing winds, significantly influences this region, which is characterised by limestone reliefs in the east and bounded by the Rhône and Durance rivers. The average annual precipitation during the period of 1981–2010 falls within the range of 515 to 586 millimetres.

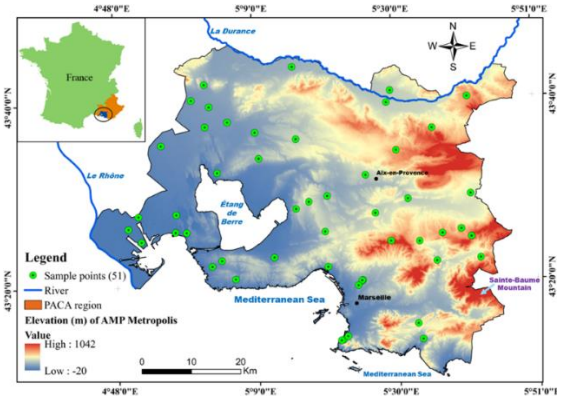

FIGURE 1. Geographical location of the AMP Metropolis with 51 sampling points.

[ISBN : 978-2-910545-13-X, Mai 2025, Avignon]

3. Methods

3.1. Soil sampling

During the 2022–2023 period, a total of 51 sites were surveyed as part of this research project. The methodology employed for sampling took into consideration the proportional size of each land use category to guarantee a consistent spatial representation. More specifically, land use categories with larger extents were allocated a greater number of sampling locations, with 25 sites designated for forests and semi-natural environments, 16 sites for agricultural areas, 6 sites for artificial zones, and 4 sites for wetlands. Before the commencement of fieldwork, the specific locations for sampling were predetermined to ensure comprehensive coverage across the entire study area. The sampling was performed according to the soil horizons down to the parent material.

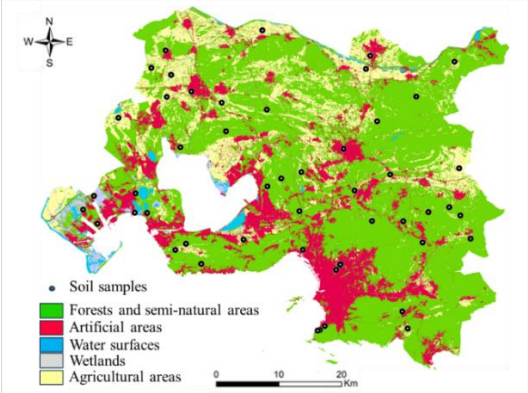

FIGURE 2. Sampling points overlaid on the 2021 land cover map.

3.2. Spatial modelling of SOC stock

The methodology employed in this research for the spatial modelling of SOC stock in the AMP Metropolis was conducted through a series of steps. Initially, the target variable (SOC stock) was randomly split into two subsets: one for training the models (80%) and the other for validation (20%). To streamline the variables, redundant elements were reduced using Spearman's correlation coefficient. The selection of relevant variables in the regression models to enhance the reliability and precision of predictions was carried out through Multicollinearity testing (VIF) and the Boruta algorithm. VIF was utilised to pinpoint and eliminate highly collinear variables, while Boruta identified and retained significantly important variables. The input database was categorised into four groups: Group 1 (full variable set), Group 2 (datasets with non-redundant variables), Group 3 (datasets with variables selected using the Boruta algorithm), and Group 4 (datasets without collinear variables). The RF and DNN models were calibrated and validated. RF is an algorithm based on multiple decision trees, where each tree is trained on a random sample of the data. The final prediction results from aggregating all the trees' predictions, improving accuracy and reducing overfitting (Breiman, 2001). DNN is a neural network with several hidden layers, where each neuron applies transformations to the data. Learning occurs by progressively adjusting the weights through backpropagation, allowing the capture of complex relationships (Hinton et al., 2006). The models were then evaluated using

three metrics: R^2, root mean square error (RMSE) and mean absolute error (MAE). R^2 indicates the proportion of explained variance, while RMSE and MAE directly measure error, highlighting the magnitude of discrepancies between predictions and observations. Finally, spatial distribution maps for SOC stock were created.

4. Results

Table 1 presents a comparison of the performance of two models, RF and DNN, in the spatial modelling of SOC stocks at a depth of 0–30 cm using four different sets of variables. In general, DNN demonstrates superior performance over RF in most cases, particularly in Group 1, which consists of 29 variables, achieving an R^2 value of 0.57 compared to RF's 0.45, as well as exhibiting lower errors in terms of RMSE and MAE. However, RF exhibits better performance in Group 2 (non-redundant variables), with an R^2 value of 0.59 compared to 0.51 for DNN. DNN once again surpasses RF in performance in Group 3 (variables selected by Boruta) and Group 4 (variables without multicollinearity), displaying higher R^2 values and diminished errors. These findings suggest that DNN outperforms RF due to its superior capability in elucidating variance, while RF proves to be more effective when there is low redundancy among variables. Despite the utilisation of selection methods such as Boruta and VIF, neither model enhances the explained variance of the dataset.

The R^2 values show a more pronounced difference between the RF and DNN models, while RMSE and MAE reveal more subtle discrepancies, indicating a trend but not significant discrimination, except for Group 2. This highlights the importance of considering R^2, RMSE, and MAE to assess the performance of the models fully. Additionally, the variability in SOC stock data (with a range of 0.4 to 11.5 kg/m² and a standard deviation of 2.4 kg/m², with an average of 6 kg/m²) may influence the RMSE and MAE results, explaining the differences observed between RF and DNN and emphasizing the complexity of spatial modelling of SOC stocks.

TABLE 1. Evaluation of the Accuracy of RF and DNN Algorithms.

Groups	R^2		RMSE (kg/m^2)		MAE (kg/m^2)	
	RF	DNN	RF	DNN	RF	DNN
Group 1: all variables	0.45	0.57	1.7	1.6	1.4	1.3
Group 2: non-redundant variables	0.59	0.51	1.5	1.8	1.2	1.7
Group 3: Boruta algorithm	0.23	0.33	2.1	1.9	1.8	1.7
Group 4: VIF test	0.39	0.54	1.8	1.6	1.5	1.4

5. Conclusion

This study conducted a comparative analysis of the efficiency of RF and DNN algorithms in mapping SOC stock in the AMP Metropolis. Despite only having 51 samples and a variety of environmental data, the DNN algorithm demonstrated superior performance to the RF model, particularly when utilizing all 29 variables due to its ability to capture intricate relationships. Conversely, the RF model exhibited better performance when using non-redundant variables, showcasing its effectiveness in specific scenarios. The resulting maps illustrated elevated SOC stock levels in

forested regions and lower levels in maritime wetlands and agricultural areas (refer to Figure 3). These findings offer a significant resource for sustainable soil management, helping in the advancement of carbon neutrality objectives. The research underscores the significance of employing sophisticated methodologies and advocates for the enhancement of data optimisation and variable selection to enhance model precision.

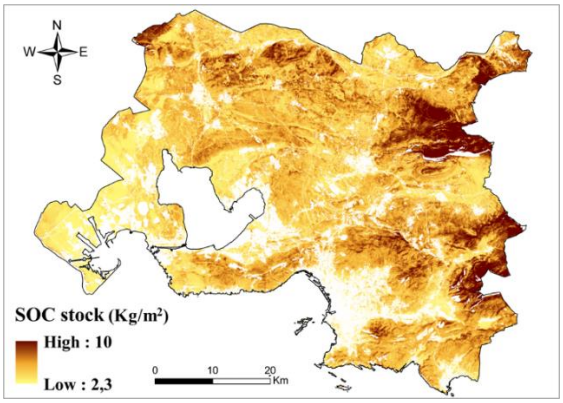

FIGURE 3. Spatial distribution of SOC stock using the DNN model.

6. Acknowledgement

This project was granted funding by the French government within the framework of the France 2030 investment plan, as a component of the Initiative d'Excellence of Aix-Marseille University - A*Midex, via the Mediterranean Institute for the Environmental Transition ITEM, AMX-19-IET-012, and through the Research Federation ECCOREV (FR3098). In addition, financial support was provided by the MAMP-CNRS SOM and the LMT NEUTRINO research initiatives.

7. Bibliography

Breiman, L. (2001). *Random forests*. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/a:1010933404324.

FAO, Bot, A., Benites, J. (2015). *The importance of soil organic matter. Key to drought-resistant soil and sustained food and production*. Retrieved from https://www.fao.org/3/a0100e/a0100e00.htm#Contents.

Hinton, G. E., Osindero, S., Teh, Y. W. (2006). *A fast learning algorithm for deep belief nets*. Neural Computation, 18(7), 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527.

Lal, R. (2004). *Soil carbon sequestration to mitigate climate change*. Geoderma. https://doi.org/10.1016/j.geoderma.2004.01.032.

Odebiri, O., Mutanga, O., Odindi, J. (2022). *Deep learning-based national scale soil organic carbon mapping with Sentinel-3 data*. Geoderma, 411. https://doi.org/10.1016/j.geoderma.2022.115695.

Ontl, T. A. Schulte, L. A. (2012). *Soil Carbon Storage*. Nature Education Knowledge 3(10):35.

Pouladi, N., Gholizadeh, A., Khosravi, V., Borůvka, L. (2023, November 1). *Digital mapping of soil organic carbon using remote sensing data: A systematic review*. Catena. Elsevier B.V. https://doi.org/10.1016/j.catena.2023.107409.

Prăvălie, R., Nita, I. A., Patriche, C., Niculiță, M., Birsan, M. V., Roșca, B., Bandoc, G. (2021). *Global changes in soil organic carbon and implications for land degradation neutrality and climate stability*. Environmental Research, 201. https://doi.org/10.1016/j.envres.2021.111580.