




An improved digital soil mapping approach to predict total N by combining machine learning algorithms and open environmental data

Alessandro Auzzas¹ · Gian Franco Capra¹ · Arun Dilipkumar Jani² · Antonio Ganga¹ 

Received: 16 May 2024 / Accepted: 2 August 2024 / Published online: 20 August 2024
© The Author(s) 2024

Abstract

Digital Soil Mapping (DSM) is fundamental for soil monitoring, as it is limited and strategic for human activities. The availability of high temporal and spatial resolution data and robust algorithms is essential to map and predict soil properties and characteristics with adequate accuracy, especially at a time when the scientific community, legislators and land managers are increasingly interested in the protection and rational management of soil.

Proximity and remote sensing, efficient data sampling and open public environmental data allow the use of innovative tools to create spatial databases and digital soil maps with high spatial and temporal accuracy. Applying machine learning (ML) to soil data prediction can improve the accuracy of maps, especially at scales where geostatistics may be inefficient. The aim of this research was to map the nitrogen (N) levels in the soils of the Nurra sub-region (north-western Sardinia, Italy), testing the performance of the Ranger, Random Forest Regression (RFR) and Support Vector Regression (SVR) models, using only open source and open access data. According to the literature, the models include soil chemical-physical characteristics, environmental and topographic parameters as independent variables. Our results showed that predictive models are reliable tools for mapping N in soils, with an accuracy in line with the literature. The average accuracy of the models is high ($R^2=0.76$) and the highest accuracy in predicting N content in surface horizons was obtained with RFR ($R^2=0.79$; RMSE=0.32; MAE=0.18). Among the predictors, SOM has the highest importance. Our results show that predictive models are reliable tools in mapping N in soils, with an accuracy in line with the literature. The results obtained could encourage the integration of this type of approach in the policy and decision-making process carried out at regional scale for land management.

Keywords Machine learning · Random Forest Regression · Support Vector Regression · Digital soil mapping · Open data

Introduction

Digital Soil Mapping (DSM) has been the main spatial information practice in soil science for many years. This sub-discipline of soil science received international recognition in 2005 with the establishment of a dedicated working group led by IUSS (Arrouays et al. 2017). Today, the main processes of DSM are based on geostatistical methods, machine learning (ML) models, and algorithms (Heung et al. 2016;

Khaledian and Miller 2020; Padarian et al. 2019; Wadoux et al. 2020). Geostatistics refers to methods of studying environmental phenomena based on their spatial variability, starting from real data collected in the field (Hoffmann et al. 2021). These tools are widely used for drafting prediction maps, especially through different Kriging algorithms (Keskin and Grunwald 2018; Santra et al. 2017; Zhang et al. 2020). Alongside them, however, ML (i.e., tools obtaining comparable results), is increasingly being used (Taghizadeh-Mehrjardi et al. 2021; Wadoux et al. 2020).

Indeed, ML is applied in several fields, such as monitoring of hydrogeological risk (Jain et al. 2020; Ma et al. 2021), wildfire prevention (Elia et al. 2020), the prediction of soil physical–chemical parameters (Li et al. 2023a, b; Li et al. 2022; Wang et al. 2021, 2022; Xu et al. 2021), and human health (Aghazadeh et al. 2019; Piunti 2019). Consequently, the number of algorithms to reference is as numerous as the

✉ Antonio Ganga
aganga@uniss.it

¹ Dipartimento Di Architettura, Design E Urbanistica,
Università Di Sassari, Via Piandanna 4, 07100 Sassari, Italy

² Department of Biology and Chemistry, California State
University, Monterey Bay, Seaside, CA 93955, USA

fields of application. Depending on the objective, the sampling characteristics and the dataset, it is necessary to choose one algorithm over another (Li et al. 2023a, b; Wadoux et al. 2020). A relevant aspect in the application of ML is the abundance and quality of databases (Chen et al. 2022). In environmental science, the application of ML requires extensive and costly surveying campaigns, which can be supported by existing databases, often shared by institutions and governmental bodies according to the logic of open data (Hengl et al. 2017). It is precisely in the environmental field that we are witnessing in recent years the proliferation of open databases, especially by public institutions (Worthy 2015), and in the field of soil science (Orgiazzi et al. 2018). Furthermore, the increased use of open data in digital soil mapping is recent and strictly related to the use of new spatial analysis tools, such as Google Earth Engine (GEE), and the availability of large datasets of remote sensing data acquired by satellite missions (Copernicus, Landsat) (Poppiel et al. 2021). National and international agencies are developing policies and tools to share soil data, also for scientific purposes, such as the LUCAS soil project implemented by the EU Environment Agency (Orgiazzi et al. 2018). Indeed, today almost all medium/large scale studies focused on digital soil mapping integrate field data with updated, publicly managed, high-resolution open data (Radočaj et al. 2024; Searle et al. 2021). This type of data, coupled by a ML algorithm, appears to be more efficient, also in terms of cost–benefit, than the traditional approach using a geostatistical algorithm (Radočaj et al. 2022a).

Soil mapping can have two main purposes: i) assignment of a class associated with observed soil, or ii) identification of one or more soil features (Zhang et al. 2017). Among these, physical–chemical parameters were extensively investigated to create regional (Brungard et al. 2021; Maleki et al. 2023), local and field scale distribution maps (Chlingaryan et al. 2018; Söderström et al. 2016; Zhou et al. 2023). Among the chemical parameters, the map elaboration for soil macronutrients (N, P, and K) represents a pivotal step, for environmental and agricultural development agencies, farmers, etc., to understand their spatial distribution and consequently improve nutrient input management while avoiding soil water pollution. Nitrogen is a fundamental macronutrient for the development of plant species, not the least because of the quantities that plants require for sustenance (Högberg et al. 2017). In fact, plant species accumulate N in different forms and through different modalities, throughout their life cycle and predominantly during the growth phases (Das et al. 2022). The continuous input of N needed by crops has a significant impact on production cycles and markets (Dimkpa et al. 2020). Use of N fertilizers has a significant economic weight; this entails careful and constant monitoring over time, to highlight the spatial distribution dynamics of N deficits and surpluses (Singh 2018; Wang et al. 2019).

The Nurra subregion (northwestern Sardinia) provides an excellent paradigmatic case to explore previously reported questions. Indeed, it encompasses several environmental conditions, passing from natural areas (Parks protected and ruled by laws) to highly productive enterprises, mainly located in plains, and represented by: the production of famous, high-quality wines that are exported around the world; from intensive to semi-intensive agricultural activities; cattle and sheep farming for meat and milk-derived products. Additionally, the area has undergone extensive urbanization due to the presence of extended urban areas (Sassari and Alghero) and famous tourist locations (Arru et al. 2019).

However, the objectives of this research were to: i) assess the effectiveness and performance of some ML models using only open access environmental databases; ii) predict N values in soil surface horizons of the Nurra sub-region (Sardinia, Italy) and iii) based on the predicted values, draw up a sub-regional scale map. Only open-access data were used, provided, and implemented by different bodies and organizations at different hierarchical levels. Variables under investigation have been selected through data exploration, i.e., an in-depth analysis of the dataset to study its distribution and main characteristics from a statistical point of view. Random tree models were used since they are in common use and integrated, as algorithms, in several statistical software packages, such as “CART”, “RF” and “Ranger,” of Rstudio (RStudio Team 2011). Furthermore, this approach has three important characteristics. It is: i) easy to reproduce with open-source software; ii) powered by public open data; iii) oriented to produce outputs that can be easily integrated into decision-making processes (Fig. 1).

Materials and methods

Study area

The study area, which covers 1,330 km², is located in NW Sardinia (Italy, Fig. 2), in the Nurra sub-region (40°48′28.8″N 8°15′14.4″E). Different geological substrates are featured in the area. The most extensive is the limestone formation, followed by pyroclastic flow deposits (south), aeolian sandstones, and gravel (Carmignani et al. 2015). The study area is characterized by high pedodiversity, (Aru e Baldaccini 1983) with Alfisols (Rhodoxeralfs, Palexeralfs, Haploxeralf), Inceptisols (Xerochrepts) and Entisols (Fluvents—Xerofluvents, Aquepts—Fluvaquepts, Psamments—Xeropsamments, (*Keys to Soil Taxonomy, 13th Edition* 2022)) dominating. The main land uses are: agriculture (65%), urban settlements (5%), and natural areas (30%, CORINE Land Cover Copernicus Land Monitoring Service). The vegetative cover is mainly divided into forest

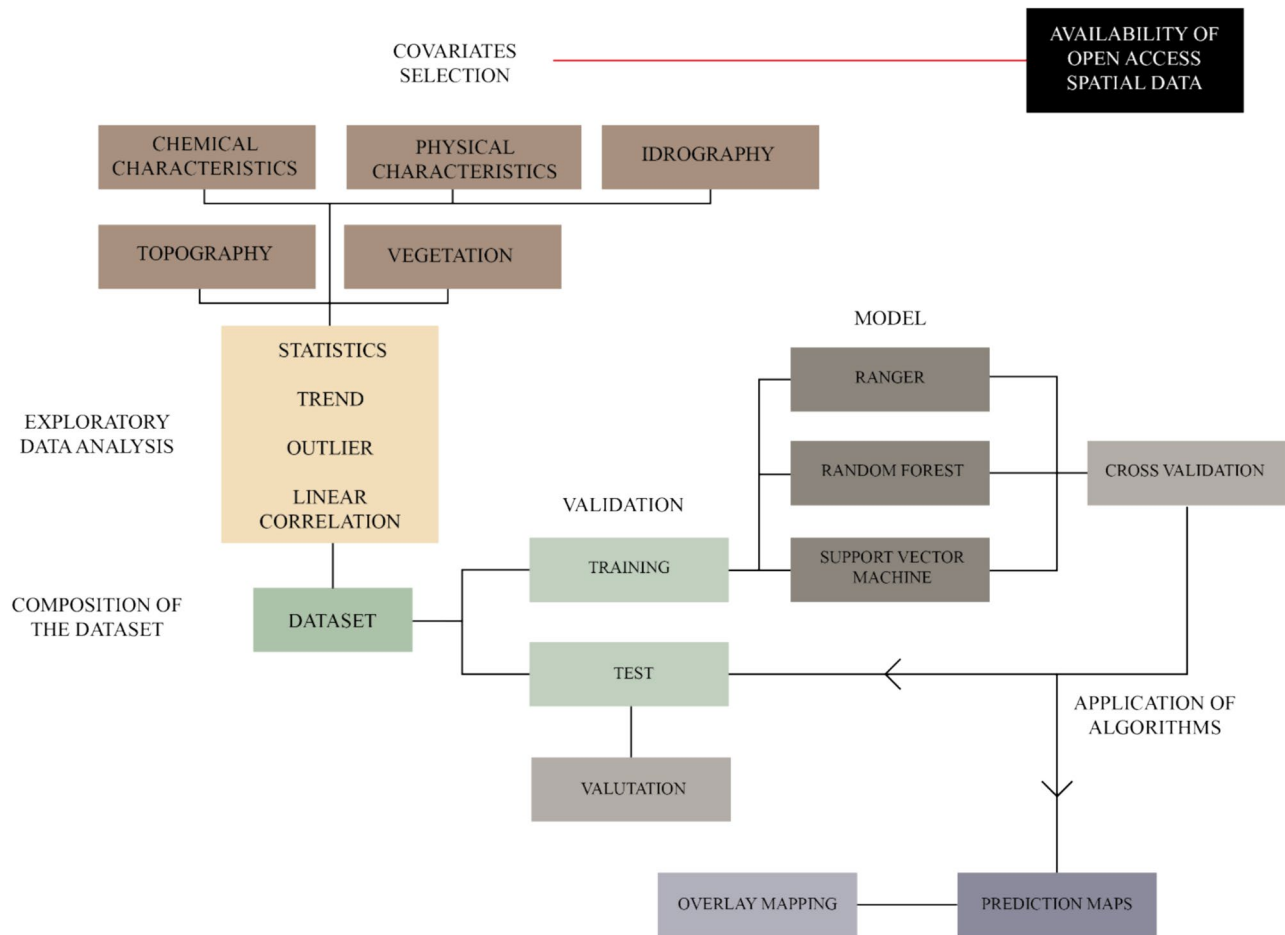


Fig. 1 Workflow Diagram

vegetation (30%), such as hardwood and coniferous trees, and arable crops (40%), as described by Corine Land Cover (CLC). A part of the forest is located on the coastline of Asinara's Bay. These are relatively recent conifer plantations placed behind the dunes. Approximately 10% of the surface is occupied by olive trees. The central part of the study area is characterized by irrigated, arable lands.

Data collection

The construction, implementation, and validation of the dataset is a pivotal part of the mapping process; the predictive results of the model depend on its characteristics and composition. The availability of quality data determines the accuracy of the model; therefore, it is necessary to build a general dataset that includes a carefully selected range of variables that, as a whole, influence the values of the variable we want to predict (Wadoux et al. 2020). Only open sources have been used in this work. The use of open sources increases the level of replicability of this research, thus providing the possibility to compare results. Furthermore, as

shown by several authors (Ferreira et al. 2022; Nussbaum et al. 2018; Wadoux et al. 2020), the availability of data, especially those related to soil characteristics, stimulates research regarding the conditions of this resource. At the same time, the existence and availability of freely accessible data increases society's awareness of soil resource issues (Gorelick et al. 2017; Orgiazzi et al. 2018). In this work, chemical, physical, topographical, and land-use-related predictors are used. In Table 1, the main characteristics of the predictors are reported (type, source and resolution).

Soil chemical-physical features

Soil data used in the study are available on the official website of the Sardinian Soil Survey.¹ These data are provided in ESRI shapefile format with a geometric punctual structure. Each one of these points represents a sample collected by

¹ Available on: <http://www.sardegnaportalesuolo.it/opendata>, redacted by Agris Sardegna.

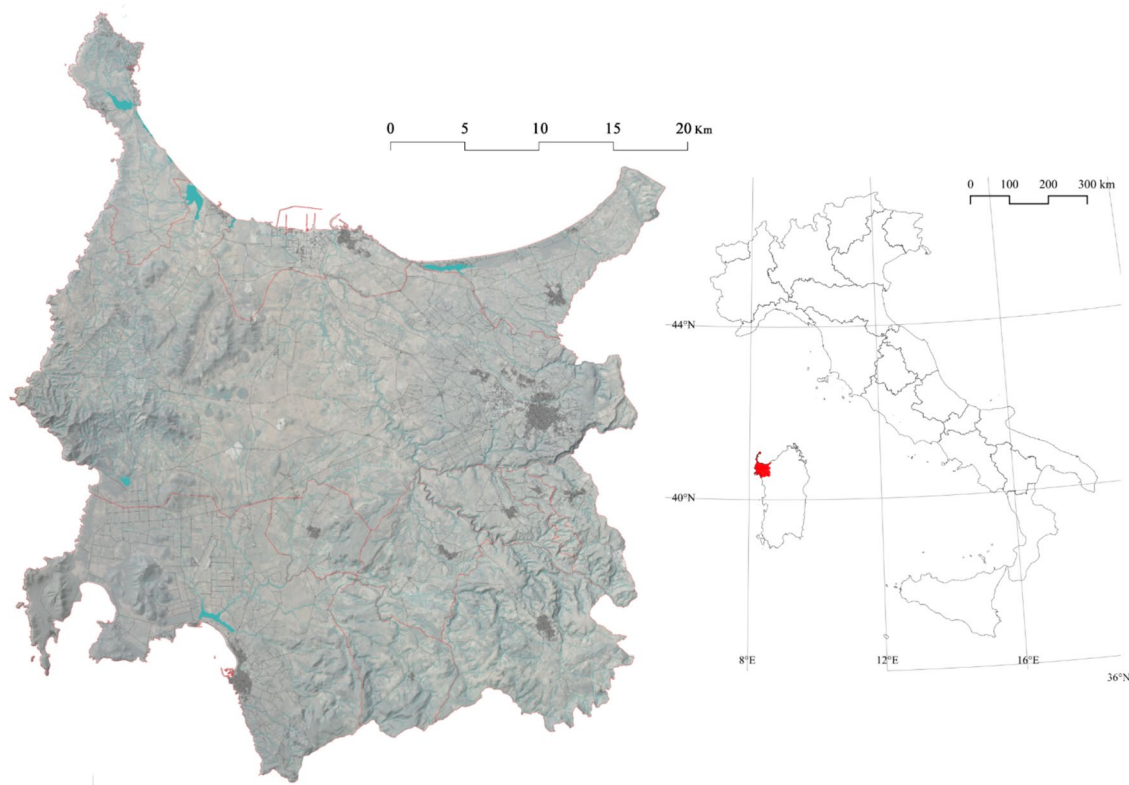


Fig. 2 Study area framework

Table 1 Predictors source and classification table

Predictors	Source	Theme	Resolution
Slope (%)	Sardegna Geoportale (https://www.sardegna-geoportale.it/index.html)	Topography	10 × 10 m
Altitude (m a.l.s.)			10 × 10 m
Topographic Position Index (TPI)			10 × 10 m
Aspect			10 × 10 m
pH	Portale dei Suoli (Regional Soil Survey, http://www.sardegna-geoportale.it/)	Chemical characteristics	100 × 100 m
SOM (Soil Organic Matter)			100 × 100 m
Ca (Calcium)			100 × 100 m
Mg (Magnesium)			100 × 100 m
Na (Sodium)			100 × 100 m
K (Potassium)			100 × 100 m
Ece (μs/cm)			100 × 100 m
P (Phosphorus)	European Soil Data Centre (ESDAC) (https://esdac.jrc.ec.europa.eu/)	Idrography	100 × 100 m
Soil loss rate (RUSLE) ($T \cdot ha^{-1} \cdot Y^{-1}$)			100 × 100 m
Waterbody Distance (m)	Sardegna Geoportale (https://www.sardegna-geoportale.it/index.html)		10 × 10 m
CLAY (%)	SOILGRIDS PROJECT (https://soilgrids.org/)	Soil Texture	195 × 299 m
SAND (%)			
SILT (%)			
NDVI	United States Geological Survey (https://earthexplorer.usgs.gov/)	Cover land/Vegetation	30 × 30 m

different institutions involved in several projects: Regional Agency (AGRI, LAORE), University of Sassari and Cagliari. There are 1511 samplings in the study area, each point is associated with the prosaic card's code and the relative link that contains the profile description and chemical and physical parameters. Unfortunately, 981 of the 1511 maps contained only physical property data, reducing the number of observations available to apply the models. Further data will be added by LUCAS.²

Topography

Topography directly and indirectly affects the dynamics of soil N concentrations (Weintraub et al. 2017). In this research, we studied the spatial variation of the Topographic Position Index (TPI), which expresses the shape of the space making up the landscape. We demonstrated the relationship between topographic index and N concentration in soil, especially in forest watersheds (Dai et al. 2022; Li et al. 2020). The data relating to the topography were explored using the Digital Terrain Model (DTM), developed by the cartographic office of the Sardinian Region, available on the Regional Geoportal (Regione Autonoma della Sardegna 2023) at the resolution 10 × 10 m. The TPI values were calculated through the SAGGIS tool (Conrad et al. 2015).

Erosion by water and distance to waterbody

Nitrogen is one of the essential macronutrients in vegetation. The color and vigour of the plant depend on the soil N concentration. Soil N is susceptible to runoff due to water-induced soil erosion (Sequi et al. 2017). A covariate related to the hydrography of the study area consisted of an estimation of soil water erosion. This estimate was made available by the European Soil Data Centre (ESDAC) and was achieved using the Revised Universal Soil Loss Equation (RUSLE) model. This empirical model is defined by the following equation:

$$A = K * R * C * l * s * P \quad (1)$$

where,

- K = Soil Erodibility, (Panagos et al. 2014);
- R = Erosivity, (Panagos et al. 2015a, b, c, d);
- C = Vegetation Cover, (Panagos et al. 2015a, b, c, d);
- l = Slope length, s = Steepness (Panagos et al. 2015b);
- P = Support Practices, (Panagos et al. 2015c).

This model estimates soil loss per year (t/ha⁻¹). Another important dataset, related to hydrography, is the Euclidean

distance between the cell and the waterbodies. The presence of water affects N concentration in the surface horizons of soil (Amicabile 2016) and is, therefore, included in the data set. Our aim was to assess the influence of these and other predictors to improve the accuracy of the predictions.

NDVI

Soil N concentrations in the surface horizons are intrinsically linked to vegetation cover conditions (Chen et al. 2014), so vegetation data contribute to assessing land degradation processes (Ridwan et al. 2024). Therefore, the vegetation index could help to detect and describe soil conditions. The vegetation spectral indices were obtained by combining several satellite images (Chlingaryan et al. 2018). One of the covariates selected to represent the vegetation cover was the Normalized Difference Vegetations Index (NDVI), which represents the vigour of the vegetation with a range of values from [−1; +1], interpreted by the color of the leaves (Antognelli 2018). This index estimates the vigour of the vegetation by photosynthesis and is found by the satellite image combination, product by Landsat 8,³ through the elaborations of the following band:

- n° 4 Red (0.64–0.67 μm).
- n° 5 Near-Infrared (0.85–0.88 μm).

the band is elaborated through the following equation:

$$NDVI = \frac{(NIR - VIS)}{(NIR + VIS)} \quad (2)$$

where,

- NIR corresponds to the band 5;
- VIS corresponds to band 4.

The final NDVI reading is the average of the values and the image detected in the summer and winter seasons in the years from 2016 to 2020. Data of the images are as follows (Table 2):

Exploratory data and spatial analysis

The Exploratory Data and Spatial Analysis (EDA) was implemented using R software. In this study, EDA consisted of analysing the distribution and composition of any predictors, through use of descriptive statistics. It was articulated in five parts: i) data collection, ii) data cleaning, iii)

² Available on: <https://esdac.jrc.ec.europa.eu/projects/lucas>

³ Available on: <https://earthexplorer.usgs.gov/>

Table 2 Satellite images data

Dates	Image Name
08/18/2016	LC08_L1TP_193032_20160818_20200906_02_T1
12/08/2016	LC08_L1TP_193032_20161208_20200905_02_T1
08/21/2017	LC08_L1TP_193032_20170821_20200903_02_T1
10/24/2017	LC08_L1TP_193032_20171024_20200902_02_T1
07/07/2018	LC08_L1TP_193032_20180707_20200831_02_T1
12/11/2018	LC08_L1TP_193032_20181112_20200830_02_T1
08/11/2019	LC08_L1TP_193032_20190811_20200827_02_T1
10/14/2019	LC08_L1TP_193032_20191014_20200825_02_T1
08/29/2020	LC08_L1TP_193032_20200829_20200906_02_T1
11/01/2020	LC08_L1TP_193032_20201101_20210317_02_T1

univariate statistics, iv) multivariate statistics, and v) spatial distribution analysis.

- i) Once collected, all data selected in a vectorial dataset in the QGIS workspace (QGIS Development Team 2023) covered a wide study area with 100×100-m cell grids. The matrix associated with the vectorial grid showed the cell as the row and the variable as the column. The raster dataset was appropriately re-scaled and transformed into a vector dataset using the QGIS raster statistics procedure. The Raster dataset was re-scaled and incorporated into a vectorial dataset using the QGIS raster statistics procedure (QGIS Development Team 2023).
- ii) In the final dataset, a general check was carried out to identify and remove the null values (NA) and outliers.
- iii) Univariate statistics were used to describe the distribution of the values of the predictor and dependent variable.
- iv) To detect multicollinearity, we created a correlation matrix. Multicollinearity is a phenomenon that arises during regression analysis when multiple variables exhibit significant correlations not only with the dependent variable but also with each other (Shrestha 2020). If two covariates are correlated, it increases the absolute error of the predictions (Daoud 2017). Therefore, this analysis helped identify variables that had no impact on prediction quality or, worse, adversely affected it. According to the literature (Chan et al. 2022; Lindner et al. 2022), we removed the covariates with a correlation coefficient >0.80, because if the value of Pearson correlation coefficient is close to 0.8, collinearity is probable (Shrestha 2020).
- v) Another analysis that we conducted on the N value point dataset was the study of spatial autocorrelations, which is the phenomenon associated with the presence of a systematic spatial variation in a variable. A positive spatial autocorrelation is the trend of a site or nearby space to have similar values (Chlingaryan et al. 2018; Li et al.

2016; Nguyen and Vu 2019). The Moran index (Moran 1948) enables an estimation of the grade of global spatial autocorrelation. The index is given by:

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \quad (3)$$

where,

N is the number of the events;

X_i and X_j are the values taken from the intensity at the points i and j with $i \neq j$;

\bar{X} is the average of the covariate considered;

w_{ij} is an element of the matrix containing arbitrary event weights.

The weights are determined according to the contiguity of the events. The range values of the index I are $[-1; +1]$ (Tybl 2016). The values closest to 1 and -1 indicate the presence of clustering. While values close to zero indicate a random spatial distribution. This approach could be useful for strengthening model selection. In the absence of high spatial correlation, it is preferable to use multivariate statistical methods rather than geostatistical methods.

Machine learning algorithms

This type of model has been used widely in both classification and regression problems. (Wadoux et al. 2020) analysed a large amount of peer-reviewed literature and found that, in the case of classification, 80% of the articles contained the application of at least one random tree model. More than one model was chosen in this research, as it is common to use several models of different types to compare results (Wadoux et al. 2020; Zhou et al. 2023).

The selection of algorithms was based on the results of previous applications in this field. As described by several authors (Wadoux et al. 2020), ML tools have not previously considered soil mechanics, phenomena, and properties, but rather learn from the data on which they are trained. For this reason, it can be useful to understand the results of the model applications in similar situations. In this case, to select the models, we search for a similar case study, where the goal is to predict the values of chemical components in the soil (Dai et al. 2022; Flynn et al. 2023; Forkuor et al. 2017; Hengl et al. 2017; Li et al. 2023a, b; Li et al. 2022; Prado Osco et al. 2019; van der Westhuizen et al. 2023; Wadoux et al. 2020; Wang et al. 2022; Xiaorui et al. 2023; Xu et al. 2021; Zhou et al. 2023). Following the bibliography analysis, the algorithms selected were Random Forest Regression (RFR), Ranger, and Support Vector Machine Regression (SVR).

Random forest regression and ranger

While the RF and model is often used in fields, such as medicine (Sarica et al. 2017), it is also widely used in soil mapping (Wadoux et al. 2020).

This method is based on the creation of forests of decision trees to improve the accuracy of predictions, and is, therefore, classified as an ensemble algorithm, i.e. one that includes a number of other models (Zhou et al. 2023). Unlike other ML models, RF randomly selects the subset of independent variables to subdivide the nodes (leaves), making it more accurate and further minimising the instability of the trees (Forkuor et al. 2017; van der Westhuizen et al. 2023). It is possible to choose the number of trees that make up the forest (Tree Number = 500), each of which is created independently using a single sample of the training data.

Ranger is a fast implementation of RF mostly used for large datasets (Wright and Ziegler 2017). Both belong to the class of tree models. The Ranger package, implemented in the R workspace, enables managing some other aspects in the model realisation phase.

Specifically, the parameters to be handled in the function are different from those of RF and allow the implementation

of model management and refinement. The main ones used in the model training phase are:

- *Quantreg*, if enabled it performs a quantile prediction through a regression forest;
- *Num.trees*, which adjusts the quantity of trees in the forest;
- *Write.forest*, to store the results of the model;
- *Min.node.size*, which is the minimum size of the leaves, the value 5 is recommended for this parameter if a regression is performed.
- *Importance*, which makes a ranking of the importance of the independent variables in the prediction, for regression the importance is based on the value of the variance of the results and is coded with the terminology “*impurity*” (Xu et al. 2016).

This makes this phase more refined compared to other models. We demonstrated the computational and memory efficiency of a ranger in the implementation done in R software, the algorithm manages many more values and variables in less time than RF, making it very effective and fast compared to other models (Wright and Ziegler 2017).

Algorithm 1 RFR Program Code

```

1:  train(x= traindata_noout[3:19],
2:      y= traindata_noout$N,
3:      method = "rf",
4:      ntree = 500,
5:      trControl = ctrl,
6:      importance = TRUE)

```

Algorithm 2 Ranger Program Code

```

1:  ranger(x= traindata_noout[3:19],
2:      y= traindata_noout$N,
3:      quantreg=TRUE,
4:      num.trees=200,
5:      importance ="impurity",
6:      write.forest = TRUE,

```

Support vector regression

SVR, an extension of Support Vector Machine for Regression issues (Lee et al. 2020; Ramedani et al. 2014) is not a widely used model in this field, but there are some examples of its application in regression issues to predict the values of different soil properties (Li et al. 2023a, b; Wang et al. 2021; Xu et al. 2021; Zhou et al. 2023). This algorithm implements a function whose purpose is to predict the dependent

variable. One of the reasons we chose this algorithm is the difference in the inner workings of the tree models. SVR formulations are analogous to common linear regression, but there are some differences concerning it (Ramedani et al. 2014). This algorithm projects the data into a high-dimension space, through the Kernel function (the choice of kernel depends on the characteristics of the data and can have a significant impact on the performance of the model (Forkuor et al. 2017)), to identify a separation hyperplane due to the

Table 3 Descriptive statistics

Covariates	Min	Max	Average	Stand. Dev	Var	Outlier*	NA
Slope	0	62.61	8.50	9.44	89.14	62	0
Altitude	0.75	420.03	85.23	69.53	4834.76	33	0
Aspect	5.20	9.40	7.57	0.70	0.49	0	1
pH	0.00	67.00	1.68	1.35	1.82	35	1
SOM	78.0	7707.0	2654	1360.15	1,849,996.48	2	125
Ca	14.0	2968.2	273	179.44	32,197.98	28	125
Mg	2.30	2905.0	148.5	278.14	77,361.81	23	125
Na	11.7	1595.3	243.9	148.33	22,001.13	26	125
K	0.00	111.00	1.07	6.89	47.46	23	172
EC	0.00	270.00	18.82	14.35	206.03	40	145
P	−6.96	5.66	−0.11	0.91	0.83	117	0
TPI	0	359.90	194.10	94.71	8970.38	0	20
Waterbody Distance	112	61.27	3.77	5.30	28.07	40	16
Soil loss rate (RUSLE method)	0	2000	387.09	417.20	174,053.20	36	0
Sand	0	532	372.2	67.85	4604.17	65	0
Silt	0	435	354.8	61.70	3807.09	34	0
Clay	0	338	239	47.20	3807.09	65	0
NDVI	0.15	0.74	0.47	0.08	0.01	7	0

support vector. Into the limit of the vector, managed by the cost parameter (C), the prediction occurs, i.e., the value predicted is located in this range (Adwad and Khanna 2015).

Algorithm 3 Ranger Program Code

```

1: svm(N ~ Slope + Altitude + Ph_H2o + So + Ca + Mg + Na + K + Ece + P + Tpi + Aspect + Rusle + Clay +
   Sand + Silt,
2: data = traindata_noout,
3: kernel = "radial",
4: cost = 1,
5: epsilon = 0.001)

```

Validation and assessment models

Two different techniques were used to validate the models. The first divided the model into two parts, in random mode. The larger part of the dataset was used to train the models (training dataset). The second part was used to test the performance of the model on unknown data (test dataset). The split of the dataset was 75% for the training dataset and the rest for the test dataset. The cross-validation, or *k-fold cross-validation* (CV), is a statistical technique that consists of dividing the training dataset into *k* parts to limit the overfitting phenomenon. The overfitting problems are essential when one wants to use ML tools, both in the case of classification and regression issues (Berrar 2019; Wang et al. 2021). According to the bibliography (Aghazadeh

et al. 2019; Berrar 2019; Dharumarajan 2019; Hounkpatin et al. 2022; Khaledian and Miller 2020; Li et al. 2023a, b; Liu et al. 2022; Maleki et al. 2023; Mashaba-Mungomezulu et al. 2021; Nolan et al. 2018; Radočaj et al. 2022b; Rahman et al. 2020; Uddameri et al. 2020; Van Der Westhuizen et al. 2022, 2023; Wadoux et al. 2020; Wang et al. 2021; Xu et al. 2021; Zhang et al. 2021; Zhou et al. 2023), the most widely used and efficient CVs are those with *K* = 5 and *K* = 10. In this paper, we have chosen a CV of *K* = 10.

The metrics used to assess the accuracy of the performance can be different according to the issue at hand. In this paper, we use the metrics that assess the residual of the prediction, i.e., the difference between actual and predicted values. The most common are the coefficient of

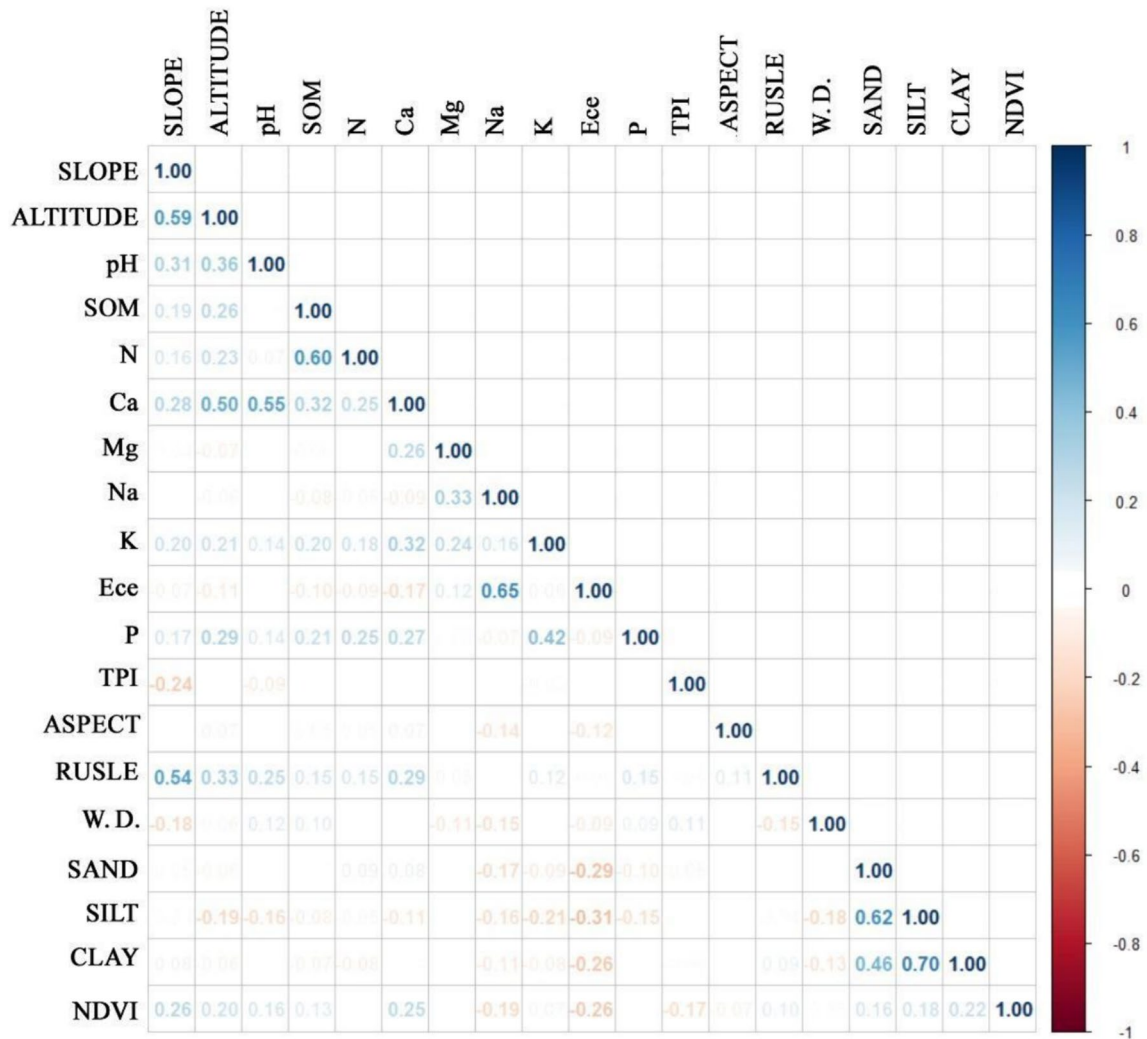


Fig. 3 Correlation matrix

determination (R^2), the root-mean-square error (RMSE), and the mean absolute error (MAE). These metrics are used in several soil mapping cases to compare the performance of the different mapping models chosen (Chlingaryan et al. 2018; Dai et al. 2022; Lee et al. 2020; Liang et al. 2018; Prado Osco et al. 2019; Wadoux et al. 2020; Zhang et al. 2019). The formulas are as follows:

– (R^2)

$$1 - \frac{\sum_{i=1}^n (P_i - O_i)^2}{\sum_{i=1}^n (O_i - \bar{O})^2} \quad (4)$$

– (RMSE)

$$\left[\frac{1}{n} \sum_{i=1}^n (P_i - O_i)^2 \right]^{1/2} \quad (5)$$

– (MAE)

$$\frac{1}{n} \sum_{i=1}^n (|O_i - P_i|) \quad (6)$$

where,

O is the real value of N;

P is the prediction.

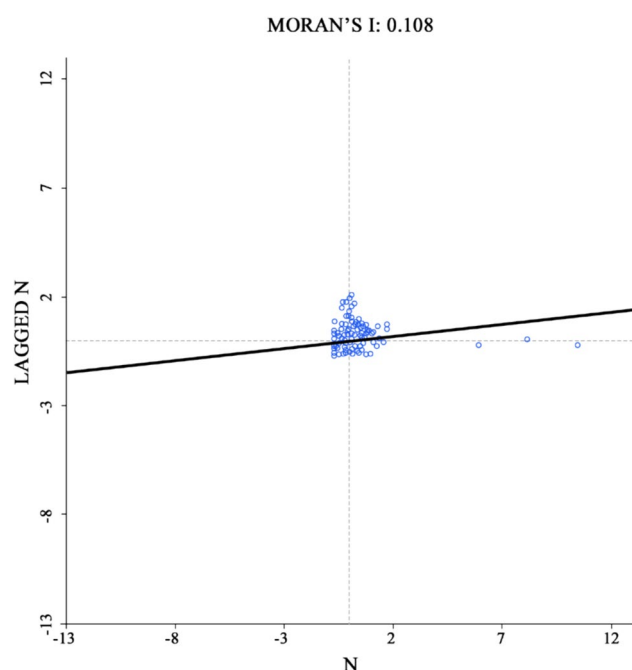


Fig. 4 Moran I scatterplot

Results and discussion

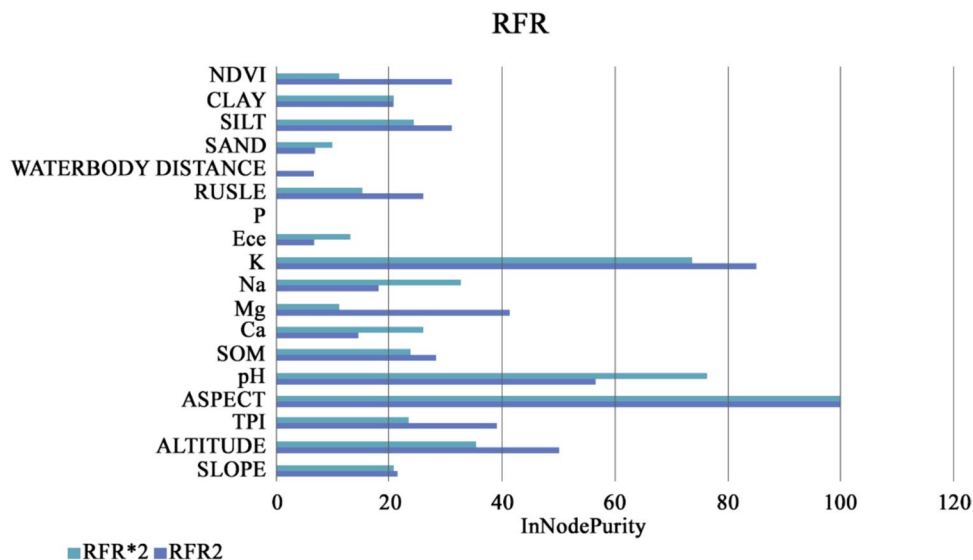
EDA

The following table shows the results of the descriptive statistical analysis (Table 3):

The final dataset consisted of 300 observations and 18 predictors.

The correlation matrix (Fig. 3) did not indicate a high association between the predictors, so we excluded the potential presence of the phenomenon of multicollinearity.

Fig. 5 Plot of covariates importance in RFR model (RFR2 = RFR standard run; RFR*2 = RFR with *tenfold CV*)



Results from the spatial autocorrelations (Fig. 4) indicated a value of 0.108. These relationships were, therefore, like random spatial phenomena; in these cases, it might be more appropriate to apply a multivariate statistical algorithm to study the distribution of variables, rather than using a 'traditional' geostatistical approach.

Covariates importance

In the tree models, it is possible to verify the importance of the variables in the predictions (Figs. 5 and 6). The importance of the variables is defined in models such as RFR and Ranger; that is why the evaluation of the importance is based on the deep mechanics of the model when it creates the tree that will compose the random forest in the regression process. The statistics analysed by the function are *InNodePurity* (Increase in Node Purity), which assesses how the purity of the node (detected by a metric such as the Gini index or the entropy) increases when a node is split based on a specific variable. High values in this case indicate a greater influence of the variable in the node splitting, in this case, process.

The SOM represents the principal source of organic N in the soil, which amounts to approximately 97–98%. Vegetation accumulated N in the ammoniacal and nitrate forms and returned it to the soil as organic N after death (Sequi et al. 2017).

For this reason, we justified the high relevance of SOM. It is important to verify, in a subsequent phase, if there is a spatial relationship between the distribution of the prediction and the values of SO. The class of variables that had the most influence in the prediction of N values were the same in both models (Table 4). It was possible to say that the predictors with more influence belonged to the class of chemical characteristics of the soil. The topography, especially altitude, also is important.

Fig. 6 Plot of covariates importance in Ranger model (Ranger2 = RFR standard run; Ranger*2 = RFR with *tenfold* CV)

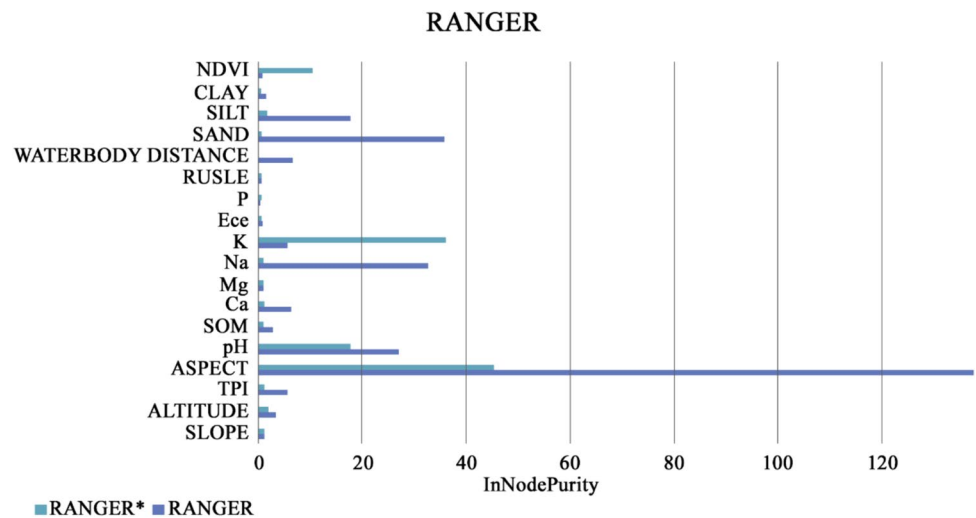


Table 4 Variable importance

Variable	Ranger (%)	RFR (%)	Ranger (%)	RFR (%)	Classes
Slope	4.39	3.64	15	13	Topography
Altitude	6.28	6.96			
Aspect	2.93	1.82			
pH	1.26	0.13			Chemical Characteristics
SOM (Soil Organic Matter)	1.58	4.62	62	68	
Ca (Calcium)	22.54	20.62			
Mg (Magnesium)	9.54	11.52			
Na (Sodium)	1.86	3.84			
K (Potassium)	5.55	3.45			
EC	2.38	3.90			
P (Phosphorus)	8.34	5.86			
TPI	10.63	13.92			
Waterbody Distance	2.46	3.19	7	5	Hydrography
Soil loss rate (RUSLE method)	4.41	1.89			
Sand	6.32	2.99	14	11	Soil Texture
Silt	5.51	5.34			
Clay	1.91	3.12			
NDVI	2.12	3.19	2	3	Cover land/Vegetation

Residual analysis

Residual analysis was performed on the predictions made in the test phase to assess the performance of the models and their accuracy when working with unknown data. SOM contributes approximately 98% of organic nitrogen in the soil. Most plant accumulate N directly from the soil as ammonium and nitrate. After death, plant N is returned to the soil in organic form (Sequi et al. 2017). For this reason and because of the importance of the variable in the prediction, we chose to relate the residuals of the results and the value of the SOM.

The greater density of values in the Ranger prediction corresponded to fewer residual values (Fig. 7). This shows

that the model generated relatively accurate results, with less deviation from the real value. Most of the results are located in the negative component of the plots, i.e., the model tends to underestimate the prediction relative to the real value. The values that were aligned in the first row of the graph were instead overlapped in the second row, corresponding to the zero value of the y-axis. The model was, therefore, able to predict these specific values without error.

In the model without CV validation, there is an inherent tendency to overestimate values in the range from 0 to 0.5. As can be seen in Fig. 7, this tendency is eliminated in the model to which tenfold CV validation has been applied. The values that are aligned in the first line in the graph are superimposed at the zero value of the y-axis in the second.

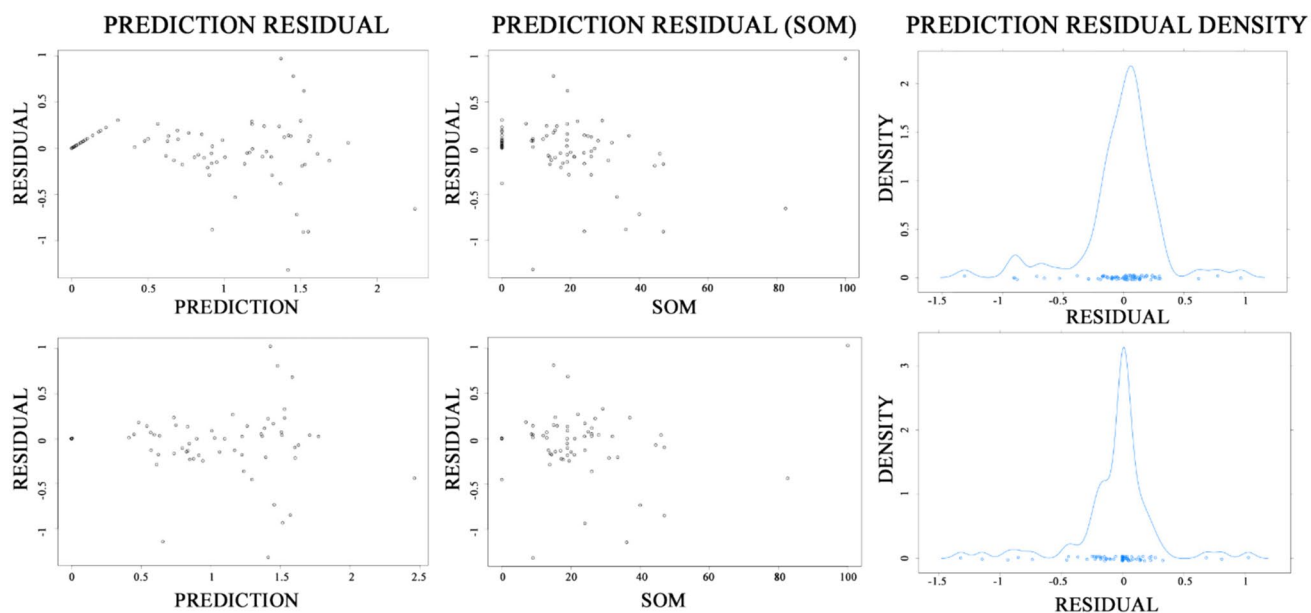


Fig. 7 Plot of residual in Ranger model (first row: application without CV *tenfold*; second row: application with CV *tenfold*)

Table 5 Residual statistics in Ranger

Models	Min	Max	Mean	Variance
RANGER	-1.31	0.97	-0.02	0.11
RANGER CV	-1.31	1.02	-0.05	0.11

The model was therefore able to predict these certain values without committing any errors. The statistics on the residuals of the two applications of the model are shown in Table 5.

The residual from the RFR model was very similar to the Ranger result. Again, the model showed the previously

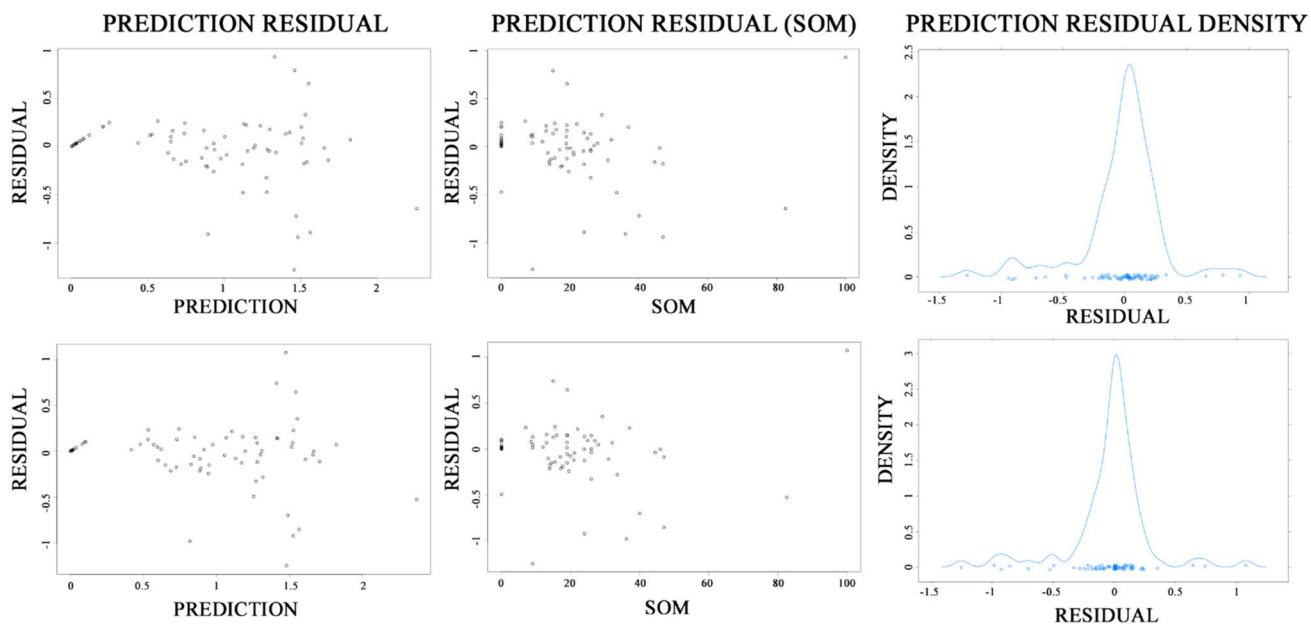


Fig. 8 Plot of residual in RFR model (first row: application without CV *tenfold*; second row: application with CV *tenfold*)

Table 6 Residual statistics in RFR

Models	Min	Max	Mean	Variance
RFR	−1.27	0.92	−0.02	0.11
RFR CV	−1.25	1.07	−0.03	0.10

observed trend, but with greater moderation compared to the Ranger results. Contrary to Ranger, the application of the model with CV did not eliminate all the trends, resulting in an overestimated prediction corresponding to a real value of 0. The density of the predicted values was concentrated near the zero value in both RFR applications with and without CV (Fig. 8). The model with CV had a higher accuracy in the density curve, indicating a lower residual between the prediction and the real values.

The statistics in Tables 5 and 6 show the affinity between the tree models in this application. Both the mean and the variance were similar. Additionally, in the complex, RFR performance was aligned with the width of the residual distribution. We can say that, even if it is short, RFR residuals assumed a high precision compared to Ranger residuals.

The SVR was influenced by the tendency to overestimate the lowest N, both with and without CV. While in the previous models, the CV limited this type of problem, in this case, the opposite was true. From the plots (Fig. 9) we can see an increase in the overestimated values, although the trend observed in the plot showing the relationships with SOM concentration was decreasing.

The residual statistics in Table 7 indicated that the model had a wider bound than the other models. This suggests that the predictions had a higher error. The density, although

Table 7 Residual Statistics SVR

Models	Min	Max	Mean	Variance
SVR	−1.40	1.44	−0.05	0.14
SVR CV	−1.48	1.5	−0.04	0.14

Table 8 Accuracy value by training stage

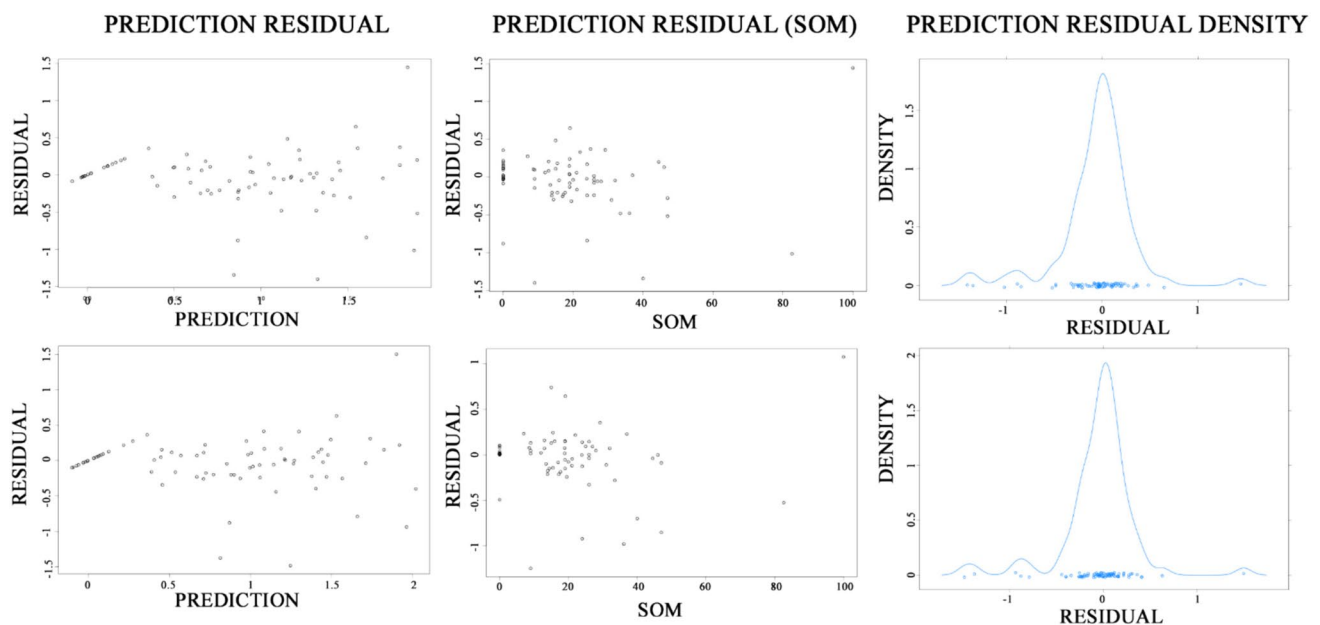
Training models	R ²	RMSE	MAE
RANGER CV	0.85	0.29	0.16
RFR CV	0.86	0.27	0.17
SVR CV	0.91	0.22	0.11

Table 9 Accuracy value by test stage

Test models	R ²	RMSE	MAE
RANGER* CV	0.77	0.34	0.19
RFR* CV	0.79	0.32	0.18
SVR* CV	0.72	0.38	0.23

more balanced, was less concentrated near the value of 0 on the x-axis, indicating an increase in the dispersion of the residuals and, therefore, a general increase in the error.

This analysis shows that CV has a positive significant influence on the model performance, regarding the tree algorithms, reducing some negative systematic trends. This does not happen in the case of the SVR algorithm and there have

**Fig. 9** Plot of residual in SVR model (first row: application without CV *tenfold*; second row: application with CV *tenfold*)

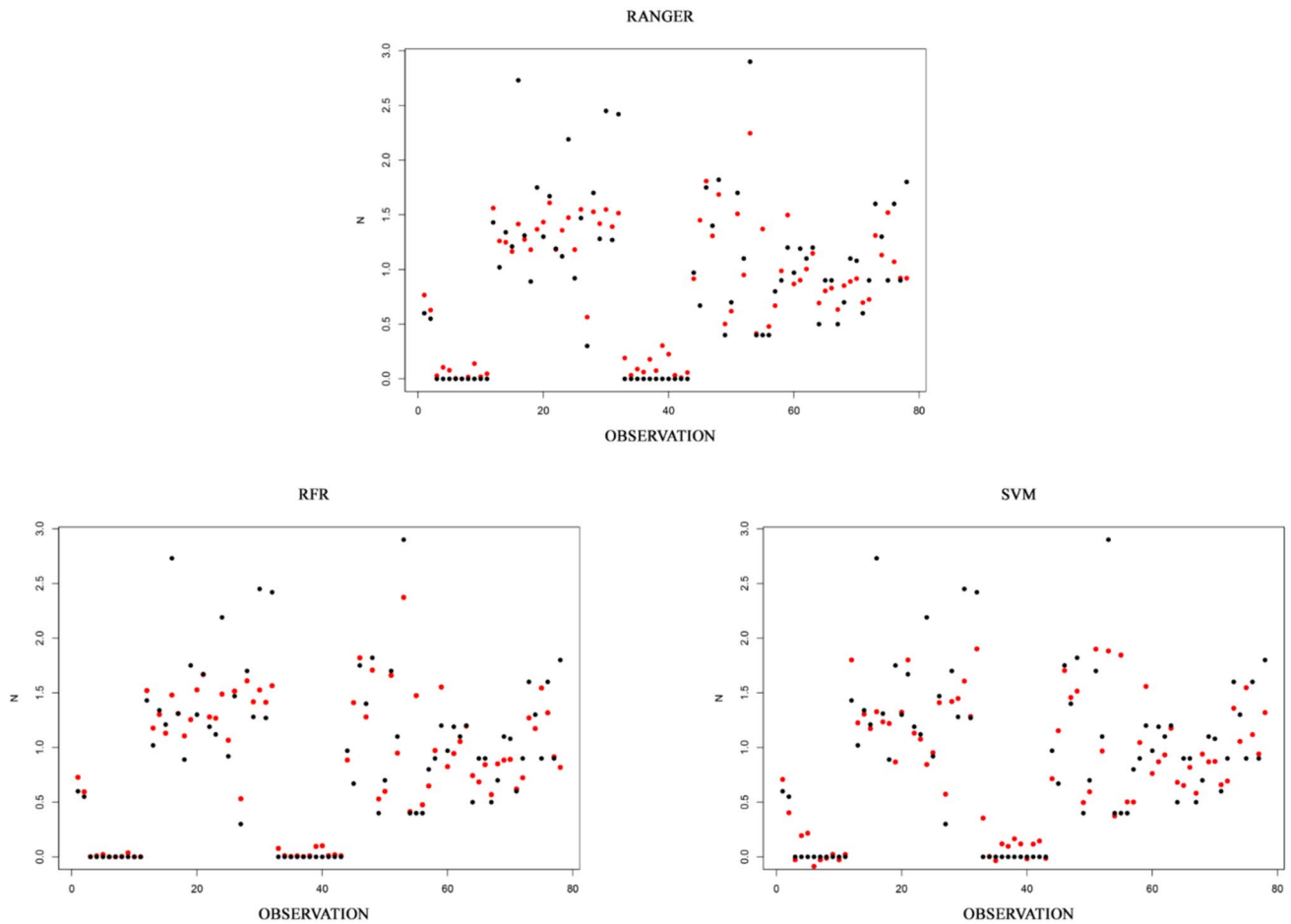


Fig. 10 Graphs of the prediction value

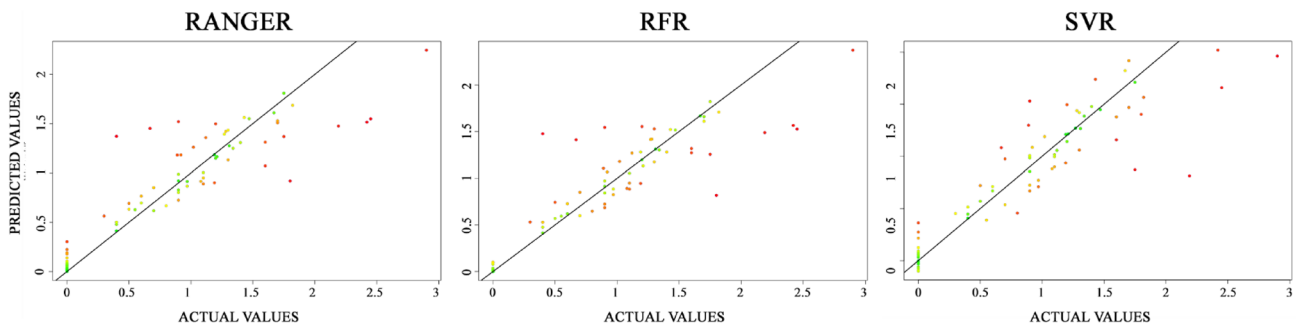


Fig. 11 Graphs of the distance between the prediction values and the actual values

been some difficulties related to the overestimations of the lowest N values.

Accuracy assessment

This analysis demonstrated the reliability of the models in a regression prediction. The results near the real values produced a more solid DSM that was typical of the landscape

characteristics. Part of the potential of these tools lies in providing a measure of the error that underlies the process of producing spatial information.

Table 8 shows the metrics related to the quality of the predictions in the training phase. These metrics are used to assess the quality of the model in predicting the training values.

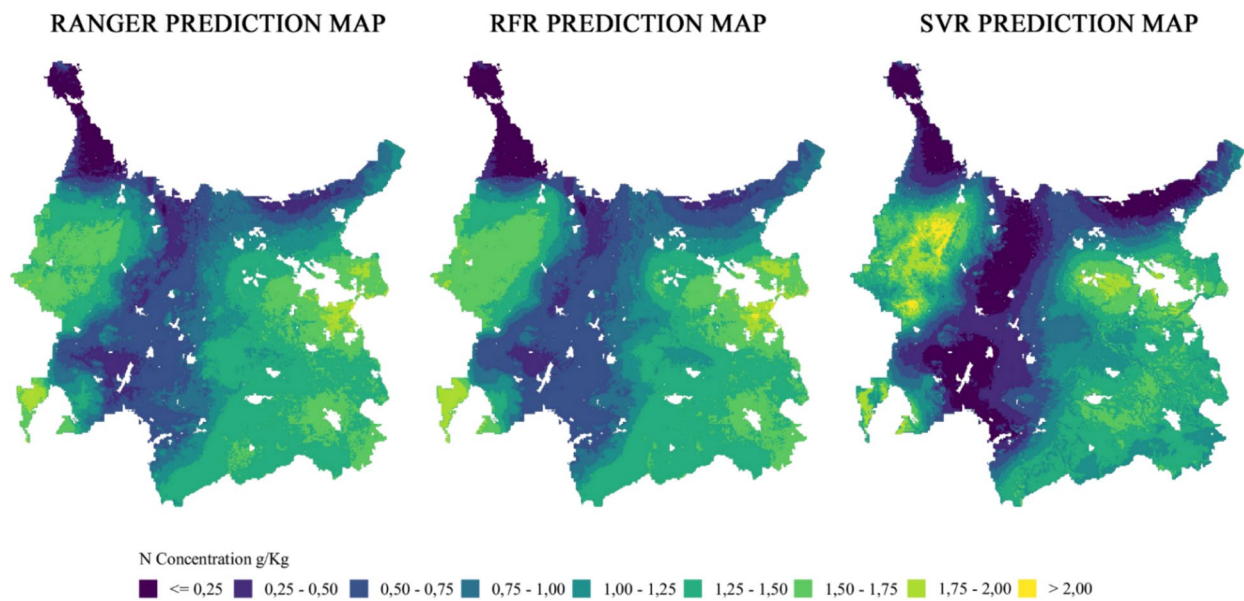


Fig. 12 Predictions Maps

From the values in Table 8, it is possible to state that the better model performance, in the training phase was obtained by SVR since the algorithm had higher R^2 values and the lowest error metrics. RFR had better quality compared to Ranger. RFR had an R^2 of 0.86 while Ranger has 0.85. Additionally, the RMSE value was lower than Ranger which had a 0.29 for the RMSE, while RFR had 0.27. For the MAE, the opposite occurred as RFR and Ranger had 0.17 and 0.16, respectively.

Table 9 shows the metric values that represent the performance quality of the prediction in the test phase.

In the test phase, the situation was reversed as SVR had the lowest performance quality in terms of the selected metrics. RFR had the highest performance quality, with a prediction that approximated the real value. The values were slightly lower than in the training phase, in fact, the highest R^2 value was obtained by RFR at 0.79. Our results align with findings in other similar works. The R^2 of the RFR model predictions was higher than that obtained by Maleki et al. (2023), even if the metric error values were worse in this case. The R^2 of RFR and SVR were comparable to those obtained by other researchers (Lee et al. 2020; Liang et al. 2018), while the RMSE values showed higher precision in respect to those obtained by (Liu et al. 2023; Prado Osco et al. 2019). SVR resulted in RMSE and R^2 values better than those found by (Xiaorui et al. 2023) for the same model application. The MAE values were more moderate than those obtained by Prado Osco et al. (2019).

The graphs in Fig. 10 show the quality of the predictions for each model. In an optimal state, the predictions (red) should agree with the real values (black dots). In this case,

all models had difficulty in predicting the highest values of N. RFR can accurately predict the value of N close to 0. Ranger and SVR cannot accurately predict the value around 0 g Kg^{-1} of N in the soil, in particular SVR which predicts a negative value.

Figure 11 shows the graphs comparing the real N values and the predictions. In an optimal state, the predictions would appear as a perfect diagonal, indicating that the prediction matches the real values. We have used a color scale for the prediction point to show the error: red indicates a high error, orange and yellow indicate a medium error, while the green point indicates a prediction close to the real values. The points in the RFR graph are more aligned along the diagonal, which, when compared to the other graphs, shows the higher quality of its prediction.

As the previous graphs show, SVR and Ranger tended to overestimate N values close to or equal to 0, which did not happen in the case of RFR applications. Finally, it is possible to observe how SVR, in some states, obtains negative values in its prediction, in correspondence with a real value equal to 0.

Prediction maps

The models were used to produce prediction maps (Fig. 12). They showed the distribution of N concentration over the study area and the influence of some critical patterns:

1. In the western part, where there was a wooded vegetation cover (with a predominance of deciduous trees), the N concentration was higher than in the area occupied by agricultural activity, due to the absence of vegetation with a long-life cycle. Even if there was a contribution of N synthesis in the fertilization phase, the N was subject to different types of losses (e.g., denitrification and leaching).
2. The same scenario characterized the arable crops and pastures that occupy the central part, while the opposite was true for the area occupied by shrub and tree vegetation.
3. The hinterland of the city of Sassari (east-central sector) was one of the areas with the higher predicted N values, which was why the area was mostly occupied by olive groves along the city limits.

The presence of a large area cultivated almost exclusively with olive trees ensures, in this condition, an adequate soil N concentration, partly due to the fertilizer applied. The low level was concentrated along the coast, where the highest level of urbanization was found. According to Amicabile (2016), all models showed an increased concentration of N, corresponding to the high levels of SOM. The predictions showed an accumulation of N along the course of the rivers, due to leaching, which manifested itself with a storage towards the lowest part. In the map product of the SVR predictions, this phenomenon was more evident. It was possible to observe high values close to the hydrographic network of the main river (Riu Mannu), localized in the eastern part.

The relationship between N concentrations in the surface horizons clearly shows that in soils of the investigated areas, the N concentrations increased as the ecosystem's conservation status increased. It clearly shows how in areas with a forest cover (with a prevalence of broad-leaved trees), N concentration is higher than in the same areas occupied by agricultural activities, due to the lack of long-cycle, high-coverage vegetation in the latter. Even if there is an input of synthetic N, due to fertilisation actions in the field, it should be remembered that N in soils is subject to various types of loss (mainly through leaching and denitrification (Amicabile 2016)). This is true for agricultural areas affected by arable crops or pastures for sheep breeding, while on soils with tree-type vegetation the opposite phenomenon occurs. Evidence of this can be seen in the fact that the models have, in all three cases, identified the maximum content in the areas bordering the city of Sassari, attributable to the massive presence of olive groves.

Concerning the difference between the model predictions, the main difference between the maps predicted by the tree models and the SVR was the localization of the higher values. In the tree models, the higher values of N

were localized in the boundaries of the city of Sassari, while the SVR predicted higher values along the western coast of the municipality of Sassari. The RFR and Ranger map products showed a high N value on the surface of the municipality of Sorso (northeast of Sassari) compared to the SVR map. This behaviour could be explained by the difference in performance in the presence of low-density sampling points.

Conclusions

This research was conducted to evaluate the effectiveness and performance of some ML models using only open environmental databases. The use of open-source data will be pivotal in the future, especially due to the large datasets acquired by remote sensing or proximity sensors. However, great importance assumes the possibilities of the use of most effective algorithm. The results showed that the RFR performed strongly. The main outcomes also revealed that by using ML algorithms, it was possible to predict N values at a medium scale coupling large open environmental databases to obtain a reliable performance. More specifically, the applied models showed approximately the same performance, with the RFR showing the highest R^2 while the RSME showed the lowest. The spatial visualization of the results demonstrated the distribution of the N value in a middle-scale map, where it was possible to detect potential critical areas that could require specific actions in the environmental policy framework. Our next steps with this research are to improve the models by incorporating additional data sources to improve the spatio-temporal scale, taking into account the quality of the data, assessed on the basis of a deep exploratory data analysis. Indeed, the high spatio-temporal resolution is crucial for the implementation of effective soil management policies in areas of high human activity density.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s40808-024-02127-8>.

Funding Open access funding provided by Università degli Studi di Sassari within the CRUI-CARE Agreement. Partial financial support was received from University of Sassari (FAR 2022, 2023, 2024).

The authors have no relevant financial or non-financial interests to disclose.

Data availability The data used to support this study are available by contacting the corresponding author.

Declarations

Conflict of interest The authors declare no competing interests.

Compliance with ethical standards The authors were compliant with the ethical standards.

Ethical approval Research meets all applicable standards relating to ethics and research integrity.

Informed consent All authors provided informed consent.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Adwad M, Khanna R (2015) Efficient learning machines. Springer, New York
- Aghazadeh M, Orooji A, Kamkar Haghighi M (2019) Developing an intelligent system for prediction of optimal dose of warfarin in Iranian adult patients with artificial heart valve. *Front Health Inform* 8(1):25. <https://doi.org/10.30699/fhi.v8i1.213>
- Amicabile S (2016) Manuale di Agricoltura (Terza). Ulrico Hoepli
- Antognelli S (2018, maggio 28) Indici di vegetazione NDVI e NDMI: Istruzioni per l'uso. *Agricolus*. <https://www.agricolus.com/indici-vegetazione-ndvi-ndmi-istruzioni-luso/>
- Arrouays D, Lagacherie P, Hartemink AE (2017) Digital soil mapping across the globe. *Geoderma Reg* 9:1–4. <https://doi.org/10.1016/j.geodrs.2017.03.002>
- Arru B, Furesi R, Madau FA, Pulina P (2019) Recreational services provision and farm diversification: a technical efficiency analysis on Italian agritourism. *Agriculture* 9(2):42. <https://doi.org/10.3390/agriculture9020042>
- Berrar D (2019) Cross-validation. In: *Encyclopedia of bioinformatics and computational biology*. Elsevier, pp 542–545. <https://doi.org/10.1016/B978-0-12-809633-8.20349-X>
- Brungard C, Nauman T, Duniway M, Veblen K, Nehring K, White D, Salley S, Anchang J (2021) Regional ensemble modeling reduces uncertainty for digital soil mapping. *Geoderma* 397:114998. <https://doi.org/10.1016/j.geoderma.2021.114998>
- Carmignani L, Oggiano G, Funedda A, Conti P, Pasci S (2015) The geological map of Sardinia (Italy) at 1:250,000 scale. *J Maps*. <https://doi.org/10.1080/17445647.2015.1084544>
- Chan JY-L, Leow SMH, Bea KT, Cheng WK, Phoong SW, Hong Z-W, Chen Y-L (2022) Mitigating the multicollinearity problem and its machine learning approach: a review. *Mathematics* 10(8):1283. <https://doi.org/10.3390/math10081283>
- Chen B, Liu E, Tian Q, Yan C, Zhang Y (2014) Soil nitrogen dynamics and crop residues. A review. *Agron Sustain Dev* 34(2):429–442. <https://doi.org/10.1007/s13593-014-0207-8>
- Chen S, Arrouays D, Leatitia Mulder V, Poggio L, Minasny B, Roudier P, Libohova Z, Lagacherie P, Shi Z, Hannam J, Meersmans J, Richerde-Forges AC, Walter C (2022) Digital mapping of GlobalSoilMap soil properties at a broad scale: a review. *Geoderma* 409:115567. <https://doi.org/10.1016/j.geoderma.2021.115567>
- Chlingaryan A, Sukkarieh S, Whelan B (2018) Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: a review. *Comput Electron Agric* 151:61–69. <https://doi.org/10.1016/j.compag.2018.05.012>
- Conrad O, Bechtel B, Bock M, Dietrich H, Fischer E, Gerlitz L, Wehberg J, Wichmann V, Böhner J (2015) System for automated geoscientific analyses (SAGA) v. 2.1.4. *Geosci Model Dev* 8(7):1991–2007. <https://doi.org/10.5194/gmd-8-1991-2015>
- Dai L, Ge J, Wang L, Zhang Q, Liang T, Bolan N, Lischeid G, Rinklebe J (2022) Influence of soil properties, topography, and land cover on soil organic carbon and total nitrogen concentration: a case study in Qinghai-Tibet plateau based on random forest regression and structural equation modeling. *Sci Total Environ* 821:153440. <https://doi.org/10.1016/j.scitotenv.2022.153440>
- Daoud JI (2017) Multicollinearity and regression analysis. *J Phys Conf Ser* 949:012009. <https://doi.org/10.1088/1742-6596/949/1/012009>
- Das PP, Singh KR, Nagpure G, Mansoori A, Singh RP, Ghazi IA, Kumar A, Singh J (2022) Plant-soil-microbes: a tripartite interaction for nutrient acquisition and better plant growth for sustainable agricultural practices. *Environ Res* 214:113821. <https://doi.org/10.1016/j.envres.2022.113821>
- Dharumaran S (2019) The need for digital soil mapping in India. *Geoderma Reg* 16:e00204
- Dimkpa CO, Fugice J, Singh U, Lewis TD (2020) Development of fertilizers for enhanced nitrogen use efficiency—trends and perspectives. *Sci Total Environ* 731:139113. <https://doi.org/10.1016/j.scitotenv.2020.139113>
- Elia M, D'Este M, Ascoli D, Giannico V, Spano G, Ganga A, Colangelo G, Laforzezza R, Sanesi G (2020) Estimating the probability of wildfire occurrence in Mediterranean landscapes using artificial neural networks. *Environ Impact Assess Rev* 85:106474. <https://doi.org/10.1016/j.eiar.2020.106474>
- Ferreira CSS, Seifollahi-Aghmiani S, Destouni G, Ghajarnia N, Kalantari Z (2022) Soil degradation in the European Mediterranean region: processes, status and consequences. *Sci Total Environ* 805:150106. <https://doi.org/10.1016/j.scitotenv.2021.150106>
- Flynn KC, Baath G, Lee TO, Gowda P, Northup B (2023) Hyperspectral reflectance and machine learning to monitor legume biomass and nitrogen accumulation. *Comput Electron Agric* 211:107991. <https://doi.org/10.1016/j.compag.2023.107991>
- Forkuor G, Hounkpatin OKL, Welp G, Thiel M (2017) High resolution mapping of soil properties using remote sensing variables in South-Western Burkina Faso: a comparison of machine learning and multiple linear regression models. *PLoS ONE* 12(1):e0170478. <https://doi.org/10.1371/journal.pone.0170478>
- Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R (2017) Google earth engine: planetary-scale geospatial analysis for everyone. *Remote Sens Environ* 202:18–27. <https://doi.org/10.1016/j.rse.2017.06.031>
- Hengl T, Leenaars JGB, Shepherd KD, Walsh MG, Heuvelink GBM, Mamo T, Tilahun H, Berkhout E, Cooper M, Fegraus E, Wheeler I, Kwabena NA (2017) Soil nutrient maps of Sub-Saharan Africa: assessment of soil nutrient content at 250 m spatial resolution using machine learning. *Nutr Cycl Agroecosyst* 109(1):77–102. <https://doi.org/10.1007/s10705-017-9870-x>
- Heung B, Ho HC, Zhang J, Knudby A, Bulmer CE, Schmidt MG (2016) An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma* 265:62–77. <https://doi.org/10.1016/j.geoderma.2015.11.014>
- Hoffmann J, Zortea M, De Carvalho B, Zadrozny B (2021) Geostatistical learning: challenges and opportunities. *Front Appl Math Stat* 7:689393. <https://doi.org/10.3389/fams.2021.689393>
- Högberg P, Näsholm T, Franklin O, Högberg MN (2017) Tamm review: on the nature of the nitrogen limitation to plant growth in Fennoscandian boreal forests. *For Ecol Manage* 403:161–185. <https://doi.org/10.1016/j.foreco.2017.04.045>

- Houkpatin KOL, Bossa AY, Yira Y, Igue MA, Sinsin BA (2022) Assessment of the soil fertility status in Benin (West Africa)—digital soil mapping using machine learning. *Geoderma Reg* 28:e00444. <https://doi.org/10.1016/j.geodrs.2021.e00444>
- Jain P, Coogan SCP, Subramanian SG, Crowley M, Taylor S, Flannigan MD (2020) A review of machine learning applications in wildfire science and management. *Environ Rev* 28(4):478–505. <https://doi.org/10.1139/er-2020-0019>
- Keskin H, Grunwald S (2018) Regression kriging as a workhorse in the digital soil mapper's toolbox. *Geoderma* 326:22–41. <https://doi.org/10.1016/j.geoderma.2018.04.004>
- Keys to Soil Taxonomy, 13th Edition (2022)
- Khaledian Y, Miller BA (2020) Selecting appropriate machine learning methods for digital soil mapping. *Appl Math Model* 81:401–418. <https://doi.org/10.1016/j.apm.2019.12.016>
- Lee H, Wang J, Leblon B (2020) Using linear regression, random forests, and support vector machine with unmanned aerial vehicle multispectral images to predict canopy nitrogen weight in corn. *Remote Sensing* 12(13):2071. <https://doi.org/10.3390/rs12132071>
- Li C, Li X, Meng X, Xiao Z, Wu X, Wang X, Ren L, Li Y, Zhao C, Yang C (2023a) Hyperspectral estimation of nitrogen content in wheat based on fractional difference and continuous wavelet transform. *Agriculture* 13(5):1017. <https://doi.org/10.3390/agriculture13051017>
- Li J, Zhang T, Shao Y, Ju Z (2023b) Comparing machine learning algorithms for soil salinity mapping using topographic factors and sentinel-1/2 data: a case study in the yellow river delta of China. *Remote Sensing* 15(9):2332. <https://doi.org/10.3390/rs15092332>
- Li R, Xu J, Luo J, Yang P, Hu Y, Ning W (2022) Spatial distribution characteristics, influencing factors, and source distribution of soil cadmium in Shantou City, Guangdong Province. *Ecotoxicol Environ Saf* 244:114064. <https://doi.org/10.1016/j.ecoenv.2022.114064>
- Li X, McCarty GW, Du L, Lee S (2020) Use of topographic models for mapping soil properties and processes. *Soil Systems* 4(2):32. <https://doi.org/10.3390/soilsystems4020032>
- Li Z, Wang J, Tang H, Huang C, Yang F, Chen B, Wang X, Xin X, Ge Y (2016) Predicting grassland leaf area index in the meadow steppes of northern China: a comparative study of regression approaches and hybrid geostatistical methods. *Remote Sensing* 8(8):632. <https://doi.org/10.3390/rs8080632>
- Liang L, Di L, Huang T, Wang J, Lin L, Wang L, Yang M (2018) Estimation of leaf nitrogen content in wheat using new hyperspectral indices and a random forest regression algorithm. *Remote Sensing* 10(12):1940. <https://doi.org/10.3390/rs10121940>
- Lindner T, Puck J, Verbeke A (2022) Beyond addressing multicollinearity: robust quantitative analysis and machine learning in international business research. *J Int Bus Stud* 53(7):1307–1314. <https://doi.org/10.1057/s41267-022-00549-z>
- Liu F, Wu H, Zhao Y, Li D, Yang J-L, Song X, Shi Z, Zhu A-X, Zhang G-L (2022) Mapping high resolution national soil information grids of China. *Sci Bull* 67(3):328–340. <https://doi.org/10.1016/j.scib.2021.10.013>
- Liu J, Yang K, Tariq A, Lu L, Soufan W, El Sabagh A (2023) Interaction of climate, topography and soil properties with cropland and cropping pattern using remote sensing data and machine learning methods. *Egypt J Remote Sens Space Sci* 26(3):415–426. <https://doi.org/10.1016/j.ejrs.2023.05.005>
- Ma Z, Mei G, Piccialli F (2021) Machine learning for landslides prevention: a survey. *Neural Comput Appl* 33(17):10881–10907. <https://doi.org/10.1007/s00521-020-05529-8>
- Maleki S, Karimi A, Mousavi A, Kerry R, Taghizadeh-Mehrjardi R (2023) Delineation of soil management zone maps at the regional scale using machine learning. *Agronomy* 13(2):445. <https://doi.org/10.3390/agronomy13020445>
- Mashaba-Munghemezulu Z, Chirima GJ, Munghemezulu C (2021) Modeling the spatial distribution of soil nitrogen content at smallholder maize farms using machine learning regression and sentinel-2 data. *Sustainability* 13(21):11591. <https://doi.org/10.3390/su132111591>
- Moran PAP (1948) The interpretation of statistical maps. *J Roy Stat Soc Ser B* 10(2):243–251. <https://doi.org/10.1111/j.2517-6161.1948.tb00012.x>
- Nguyen TT, Vu TD (2019) Identification of multivariate geochemical anomalies using spatial autocorrelation analysis and robust statistics. *Ore Geol Rev* 111:102985. <https://doi.org/10.1016/j.oregeorev.2019.102985>
- Nolan BT, Green CT, Juckem PF, Liao L, Reddy JE (2018) Meta-modeling and mapping of nitrate flux in the unsaturated zone and groundwater, Wisconsin, USA. *J Hydrol* 559:428–441. <https://doi.org/10.1016/j.jhydrol.2018.02.029>
- Nussbaum M, Spiess K, Baltensweiler A, Grob U, Keller A, Greiner L, Schaepman ME, Papritz A (2018) Evaluation of digital soil mapping approaches with large sets of environmental covariates. *Soil* 4(1):1–22. <https://doi.org/10.5194/soil-4-1-2018>
- Orgiazzi A, Ballabio C, Panagos P, Jones A, Fernández-Ugalde O (2018) LUCAS soil, the largest expandable soil dataset for Europe: a review. *Eur J Soil Sci* 69(1):140–153. <https://doi.org/10.1111/ejss.12499>
- Padarian J, Minasny B, McBratney AB (2019) Using deep learning for digital soil mapping. *Soil* 5(1):79–89. <https://doi.org/10.5194/soil-5-79-2019>
- Panagos P, Ballabio C, Borrelli P, Meusburger K, Klik A, Rousseva S, Tadić MP, Michaelides S, Hrabalíková M, Olsen P, Aalto J, Lakatos M, Rymaszewicz A, Dumitrescu A, Begueria S, Alewell C (2015a) Rainfall erosivity in Europe. *Sci Total Environ* 511:801–814. <https://doi.org/10.1016/j.scitotenv.2015.01.008>
- Panagos P, Borrelli P, Meusburger K (2015b) A new European slope length and steepness factor (LS-Factor) for modeling soil erosion by water. *Geosciences* 5(2):117–126. <https://doi.org/10.3390/geosciences5020117>
- Panagos P, Borrelli P, Meusburger K, Alewell C, Lugato E, Montanarella L (2015c) Estimating the soil erosion cover-management factor at the European scale. *Land Use Policy* 48:38–50. <https://doi.org/10.1016/j.landusepol.2015.05.021>
- Panagos P, Borrelli P, Meusburger K, van der Zanden EH, Poesen J, Alewell C (2015d) Modelling the effect of support practices (P-factor) on the reduction of soil erosion by water at European scale. *Environ Sci Policy* 51:23–34. <https://doi.org/10.1016/j.envsci.2015.03.012>
- Panagos P, Meusburger K, Ballabio C, Borrelli P, Alewell C (2014) Soil erodibility in Europe: a high-resolution dataset based on LUCAS. *Sci Total Environ* 479–480:189–200. <https://doi.org/10.1016/j.scitotenv.2014.02.010>
- Piunti V (2019) ALGORITMI DI MACHINE LEARNING SUPERVISIONATO: POSSIBILI APPLICAZIONI NEL SETTORE ASSICURATIVOSANITARIO [UNIVERSITÀ POLITECNICA DELLE MARCHE FACOLTÀ DI ECONOMIA “GIORGIO FUÀ”]. <https://tesi.univpm.it/bitstream/20.500.12075/7161/2/TESI%20VALENTINO%20PIUNTI.pdf>
- Poppiel RR, Demattê JAM, Rosin NA, Campos LR, Tayebi M, Bonfatti BR, Ayoubi S, Tajik S, Afshar FA, Jafari A, Hamzehpour N, Taghizadeh-Mehrjardi R, Ostovari Y, Asgari N, Naimi S, Nabiollahi K, Fathizad H, Zeraatpisheh M, Javaheri F, Rahmati M (2021) High resolution middle eastern soil attributes mapping via open data and cloud computing. *Geoderma* 385:114890. <https://doi.org/10.1016/j.geoderma.2020.114890>
- Prado Osco L, Marques Ramos AP, Roberto Pereira D, Akemi Saito Moriya É, Nobuhiro Imai N, Takashi Matsubara E, Estrabis N, De Souza M, Marcato Junior J, Gonçalves WN, Li J, Liesenberg V, Eduardo Creste J (2019) Predicting canopy nitrogen

- content in citrus-trees using random forest algorithm associated to spectral vegetation indices from UAV-imagery. *Remote Sens* 11(24):2925. <https://doi.org/10.3390/rs11242925>
- QGIS Development Team (2023) QGIS [Software]. Open Source Geospatial Foundation Project. <http://qgis.osgeo.org>
- Radočaj D, Gašparović M, Jurišić M (2024) Open remote sensing data in digital soil organic carbon mapping: a review. *Agriculture* 14(7):1005. <https://doi.org/10.3390/agriculture14071005>
- Radočaj D, Jurišić M, Antonić O, Šiljeg A, Cukrov N, Rapčan I, Plaščak I, Gašparović M (2022a) A multiscale cost-benefit analysis of digital soil mapping methods for sustainable land management. *Sustainability* 14(19):12170. <https://doi.org/10.3390/su141912170>
- Radočaj D, Jurišić M, Antonić O, Šiljeg A, Cukrov N, Rapčan I, Plaščak I, Gašparović M (2022b) A multiscale cost-benefit analysis of digital soil mapping methods for sustainable land management. *Sustainability* 14(19):12170. <https://doi.org/10.3390/su141912170>
- Rahman MM, Zhang X, Ahmed I, Iqbal Z, Zeraatpisheh M, Kanzaki M, Xu M (2020) Remote sensing-based mapping of senescent leaf C: N ratio in the sundarbans reserved forest using machine learning techniques. *Remote Sens* 12(9):1375. <https://doi.org/10.3390/rs12091375>
- Ramedani Z, Omid M, Keyhani A, Shamshirband S, Khoshnevisan B (2014) Potential of radial basis function based support vector regression for global solar radiation prediction. *Renew Sustain Energy Rev* 39:1005–1011. <https://doi.org/10.1016/j.rser.2014.07.108>
- Regione Autonoma della Sardegna (2023) Sardegna Geoportale [Webgis]. SardegnaMappe. https://www.sardegnageoportale.it/webgis/s2/sardegnaappe/?map=download_raster
- Ridwan I, Kadir S, Nurlina N (2024) Wetland degradation monitoring using multi-temporal remote sensing data and watershed land degradation index. *Global J Environ Sci Manag* 10(1):83–96. <https://doi.org/10.22034/gjesm.2024.01.07>
- RStudio Team (2011) RStudio: Integrated Development for R [Software]. RStudio Team (2020). <http://www.rstudio.com/>
- Santra P, Kumar M, Panwar N (2017) Digital soil mapping of sand content in arid western India through geostatistical approaches. *Geoderma Reg* 9:56–72. <https://doi.org/10.1016/j.geodrs.2017.03.003>
- Sarica A, Cerasa A, Quattrone A (2017) Random forest algorithm for the classification of neuroimaging data in alzheimer's disease: a systematic review. *Front Aging Neurosci* 9:329. <https://doi.org/10.3389/fnagi.2017.00329>
- Searle R, McBratney A, Grundy M, Kidd D, Malone B, Arrouays D, Stockman U, Zund P, Wilson P, Wilford J, Van Gool D, Triantafyllis J, Thomas M, Stower L, Slater B, Robinson N, Ringrose-Voase A, Padarian J, Payne J, Andrews K (2021) Digital soil mapping and assessment for Australia and beyond: a propitious future. *Geoderma Reg* 24:e00359. <https://doi.org/10.1016/j.geodrs.2021.e00359>
- Sequi P, Ciavatta C, Milano T (2017) *Fondamenti della chimica del Suolo*. Pàtron Editore
- Shrestha N (2020) Detecting Multicollinearity in regression analysis. *Am J Appl Math Stat* 8(2):39–42. <https://doi.org/10.12691/ajams-8-2-1>
- Singh B (2018) Are nitrogen fertilizers deleterious to soil health? *Agronomy* 8(4):48. <https://doi.org/10.3390/agronomy8040048>
- Söderström M, Sohlenius G, Rodhe L, Piikki K (2016) Adaptation of regional digital soil mapping for precision agriculture. *Precision Agric* 17(5):588–607. <https://doi.org/10.1007/s11119-016-9439-8>
- Taghizadeh-Mehrjardi R, Hamzehpour N, Hassanzadeh M, Heung B, Ghebleh Goydaragh M, Schmidt K, Scholten T (2021) Enhancing the accuracy of machine learning models using the super learner technique in digital soil mapping. *Geoderma* 399:115108. <https://doi.org/10.1016/j.geoderma.2021.115108>
- Tybl A (2016) An overview of spatial econometrics. SSRN Electron J. <https://doi.org/10.2139/ssrn.2778679>
- Uddameri V, Silva A, Singaraju S, Mohammadi G, Hernandez E (2020) Tree-based modeling methods to predict nitrate exceedances in the Ogallala aquifer in Texas. *Water* 12(4):1023. <https://doi.org/10.3390/w12041023>
- van der Westhuizen S, Heuvelink GBM, Hofmeyr DP (2023) Multivariate random forest for digital soil mapping. *Geoderma* 431:116365. <https://doi.org/10.1016/j.geoderma.2023.116365>
- Van Der Westhuizen S, Heuvelink GBM, Hofmeyr DP, Poggio L (2022) Measurement error-filtered machine learning in digital soil mapping. *Spat Stat* 47:100572. <https://doi.org/10.1016/j.spasta.2021.100572>
- Wadoux AMJ-C, Minasny B, McBratney AB (2020) Machine learning for digital soil mapping: applications, challenges and suggested solutions. *Earth Sci Rev* 210:103359. <https://doi.org/10.1016/j.earscirev.2020.103359>
- Wang L, Chen S, Li D, Wang C, Jiang H, Zheng Q, Peng Z (2021) Estimation of paddy rice nitrogen content and accumulation both at leaf and plant levels from UAV hyperspectral imagery. *Remote Sens* 13(15):2956. <https://doi.org/10.3390/rs13152956>
- Wang N, Luo Y, Liu Z, Sun Y (2022) Spatial distribution characteristics and evaluation of soil pollution in coal mine areas in Loess Plateau of northern Shaanxi. *Sci Rep* 12(1):16440. <https://doi.org/10.1038/s41598-022-20865-6>
- Wang X, Fan J, Xing Y, Xu G, Wang H, Deng J, Wang Y, Zhang F, Li P, Li Z (2019) The effects of mulch and nitrogen fertilizer on the soil environment of crop plants. *Adv Agron* 153:121–173. <https://doi.org/10.1016/bs.agron.2018.08.003>
- Weintraub SR, Brooks PD, Bowen GJ (2017) Interactive effects of vegetation type and topographic position on nitrogen availability and loss in a temperate montane ecosystem. *Ecosystems* 20(6):1073–1088. <https://doi.org/10.1007/s10021-016-0094-8>
- Worthy B (2015) The impact of open data in the UK: complex, unpredictable, and political. *Public Adm* 93(3):788–805. <https://doi.org/10.1111/padm.12166>
- Wright MN, Ziegler A (2017) Ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J Stat Softw* 77:1–17. <https://doi.org/10.18637/jss.v077.i01>
- Xiaorui L, Jiamin Y, Longji Y (2023) Predicting the high heating value and nitrogen content of torrefied biomass using a support vector machine optimized by a sparrow search algorithm. *RSC Adv* 13(2):802–807. <https://doi.org/10.1039/D2RA06869A>
- Xu R, Nettleton D, Nordman DJ (2016) Case-specific random forests. *J Comput Graph Stat* 25(1):49–65. <https://doi.org/10.1080/10618600.2014.983641>
- Xu S, Wang M, Shi X, Yu Q, Zhang Z (2021) Integrating hyperspectral imaging with machine learning techniques for the high-resolution mapping of soil nitrogen fractions in soil profiles. *Sci Total Environ* 754:142135. <https://doi.org/10.1016/j.scitotenv.2020.142135>
- Zhang G, Liu F, Song X (2017) Recent progress and future prospect of digital soil mapping: a review. *J Integr Agric* 16(12):2871–2885. [https://doi.org/10.1016/S2095-3119\(17\)61762-3](https://doi.org/10.1016/S2095-3119(17)61762-3)
- Zhang P, Yin Z-Y, Jin Y-F (2021) State-of-the-art review of machine learning applications in constitutive modeling of soils. *Archiv Comput Methods Eng* 28(5):3661–3686. <https://doi.org/10.1007/s11831-020-09524-z>
- Zhang Y, Ji W, Saurette DD, Easher TH, Li H, Shi Z, Adamchuk VI, Biswas A (2020) Three-dimensional digital soil mapping of multiple soil properties at a field-scale using regression kriging. *Geoderma* 366:114253. <https://doi.org/10.1016/j.geoderma.2020.114253>
- Zhang Y, Sui B, Shen H, Ouyang L (2019) Mapping stocks of soil total nitrogen using remote sensing data: a comparison of random forest models with different predictors. *Comput Electron Agric* 160:23–30. <https://doi.org/10.1016/j.compag.2019.03.015>

Zhou J, Xu Y, Gu X, Chen T, Sun Q, Zhang S, Pan Y (2023) High-precision mapping of soil organic matter based on UAV imagery using machine learning algorithms. *Drones* 7(5):290. <https://doi.org/10.3390/drones7050290>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.