Check for updates

RESEARCH ARTICLE

REVISED The possibility of spatial mapping of SOC content in olive groves under integrated production using easy-to-obtain ancillary data in a Mediterranean area

[version 2; peer review: 1 approved, 2 approved with reservations, 2 not approved]

Francisco José Blanco Velázquez[1], Mahmoud Shahabi[2], Hossein Rezaei[2], Félix González-Peñaloza[1], Farzin Shahbazi[2], María Anaya-Romero[1]

[1]Evenor-Tech, Seville, Andalusia, 41018, Spain
[2]Soil Science Department, Faculty of Agriculture, University of Tabriz, Tabriz, Iran

v2 **First published:** 20 Sep 2022, **2**:110
https://doi.org/10.12688/openreseurope.14716.1
Latest published: 17 Jan 2024, **2**:110
https://doi.org/10.12688/openreseurope.14716.2

Abstract

Background

Unlike most of Europe, Andalucía in southern Spain as a Mediterranean area still lacks digital maps of SOC content provided by machine learning algorithms. The wide diversity of climate, geology, hydrology, landscape, topography, vegetation, and micro-relief data as easy-to-obtain covariates facilitated the development of digital soil mapping (DSM). The purpose of this research is to model and map the spatial distribution of SOC at three depths, in an area of approximately 10000 km2 located in Seville and Cordoba Provinces, and to use R programming to compare two machine learning techniques (cubist and random forest) for developing SOC maps at multiple depths.

Methods

Environmental covariates used in this research include nine derivatives from digital elevation models (DEM), three climatic variables and finally eighteen remotely-sensed spectral data (band ratios calculated by the acquired Landsat-8 OLI and Sentinel-2A MSI in July 2019). In total, 300 soil samples from 100 points were taken (0-25 cm). The purpose of this research is to model and map the spatial distribution of SOC, in an area with approximately 10000 km2 located in Seville and Cordoba Provinces, and to compare two machine learning techniques (cubist and random forest) by R programming.

Open Peer Review

Approval Status ✓ ✗ ? ? ✗

	1	2	3	4	5
version 2 (revision) 17 Jan 2024			? view		✗ view
version 1 20 Sep 2022	✓ view	✗ view		? view	

1. **Sameh Kotb Abd-Elmabod**, National Research Centre, Cairo, Egypt
2. **Fuat Kaya**, Isparta University of Applied Sciences, Isparta, Turkey
3. **Calogero Schillaci**, JRC European Commission, Ispra, Italy
4. **Prava Kiran Dash**, Odisha University of Agriculture and Technology, Bhubaneswar, India
5. **Odunayo David Adeniyi**, University of Pavia, Pavia, Italy

Any reports and responses or comments on the article can be found at the end of the article.

Results

The findings showed that the novel approach for integrating the indices using Landsat-8 OLI and Sentinel-2A MSI satellite data had a better result.

Conclusions

Finally, we obtained evidence that the resolution of satellite images is more important in modelling and digital mapping.

Keywords
Digital soil mapping, Environmental covariates, organic carbon, uncertainty analysis



This article is included in the Horizon 2020 gateway.



This article is included in the Plant and Soil Health collection.

Corresponding author: Francisco José Blanco Velázquez (fj.blanco@evenor-tech.com)

Author roles: Blanco Velázquez FJ: Data Curation, Formal Analysis, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Shahabi M**: Formal Analysis, Investigation, Methodology, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Rezaei H**: Formal Analysis, Investigation, Methodology, Software, Writing – Original Draft Preparation; **González-Peñaloza F**: Data Curation, Formal Analysis, Investigation, Validation, Writing – Review & Editing; **Shahbazi F**: Investigation, Methodology, Software, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Anaya-Romero M**: Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This research was financially supported by the European Union's Horizon 2020 research and innovation programme under the grant agreement No 817949 (CONtract SOLutions for Effective and lasting delivery of agri-environmental-climate public goods by EU agriculture and forestry [CONSOLE]).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2024 Blanco Velázquez FJ *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Blanco Velázquez FJ, Shahabi M, Rezaei H *et al.* **The possibility of spatial mapping of SOC content in olive groves under integrated production using easy-to-obtain ancillary data in a Mediterranean area [version 2; peer review: 1 approved, 2 approved with reservations, 2 not approved]** Open Research Europe 2024, **2**:110
https://doi.org/10.12688/openreseurope.14716.2

First published: 20 Sep 2022, **2**:110 https://doi.org/10.12688/openreseurope.14716.1

REVISED **Amendments from Version 1**

The revision of the article entitled "The possibility of spatial mapping of SOC content in olive groves under integrated production using easy-to-obtain ancillary data in a Mediterranean area" has focused on the suppression of data and information related to the subsoil. This, following the corrections and suggestions of the reviews received. Under the criteria established by the corrections, we proceeded to modify the sampling data and the analyses with the satellite images, using only those that match with the time of the samplings. The retained data are mostly from the topsoil, modifying the entire article to focus on monitoring carbon in the topsoil.

Any further responses from the reviewers can be found at the end of the article

Introduction

The knowledge of the spatial and vertical distribution of soil organic carbon (SOC) is helpful either in optimizing management for soil fertility or in monitoring the potential of carbon sequestration (Bas van Wesemael *et al.*, 2023). The EU Commission is working on the future EU Soil Strategy whose objectives include protecting soil fertility, reducing soil erosion, increasing soil organic matter and restoring carbon-rich ecosystems (EC, 2021). SOC is one of the soil properties which will determine the vegetation suitable for growing on that soil. In addition, this vegetation will influence the amount and vertical distribution and forms of SOC in the soil (Bailey *et al.*, 2019).

There is a large variation of SOC at field scale due to its dependency on environmental variables. Moreover, it is necessary to develop digital maps in an accurate manner using cost-free indices over the area of interest. A compilation of relevant covariates (*scorpan* factors) can be found in Minasny *et al.*, 2013, related to SOC. McBratney *et al.* (2003) reported that digital elevation model (DEM) and spectral reflectance bands from satellite imageries were commonly used to handle digital soil mapping (DSM). An accurate prediction of SOC using solely easy-to-obtain indices e.g., DEM-derived data can save considerable labour and costs (She *et al.*, 2014). The spatial patterns of SOC were linked topography in a case study conducted in south-eastern Spain (Schwanghart & Jarmer, 2011). In addition, SOC can be successfully predicted by using remotely-sensed data (Odebiri *et al.*, 2020). It was also reported by Lamichhane *et al.* (2019) that climate is an influential factor in predictive mapping of SOC level at regional scales, followed by parent materials, topography and land use. Venter *et al.* (2021) relates elevation, mean annual precipitation, leaf area index (LAI) standard deviation, temperature wettest quarter and topographic diversity index as the variables of most importance for predicting SOC through Random Forest (RF) in South Africa. Furthermore, there are other research works with Landsat-8 OLI, Sentinel-2A and EnMap spectral bands to estimate soil organic matter (Rosero-Vlasova *et al.*, 2019) in Brazil.

In summary, DSM has shifted from a research phase into practical usage (Minasny & McBratney, 2016). One of the most important pathways relevant to the applied DSM is digital soil assessment (DSA) which was suggested by Carré *et al.* (2007). In DSA, the understanding of the spatial prediction of soil attributes can relate to many other parameters e.g., potential distribution of oaks in southern Spain (Pino-Mejías *et al.*, 2010); agriculturally important nutrients (Shahbazi *et al.*, 2019a) and the soil ripening process in Iran (Mousavi *et al.*, 2020); and available water content in Australia (Gooley *et al.*, 2014).

Contrary to conventional soil mapping, DSM uses quantitative inference models to generate predictions of soil properties in a geographic database (Zeraatpisheh *et al.*, 2020). To avoid a lot of soil surveying and sampling for mapping the spatial distribution of each property, it can be successfully conducted using easy-to-obtain indices as an envoy of aforementioned factors (Shahbazi *et al.*, 2019b). Data mining or machine learning techniques are popular for mapping soil properties with the application of numerous spatially explicit covariates (Blanco-Velázquez *et al.*, 2019). In this way, the literacy of global positioning systems (GPS), geographic information systems (GIS), acquisition of remotely-sensed data, terrain-related attributes, inference models, and software for data analysis e.g. R programming language are essential (Malone *et al.*, 2017).

In terms of SOC, the literature revealed that various environmental covariates as well as different data mining techniques have been applied worldwide (Muñoz-Rojas *et al.*, 2013 and Muñoz-Rojas *et al.*, 2012). Taghizadeh-Mehrjardi *et al.* (2016) reported that artificial neural networks (ANN) have the highest performance in prediction of SOC in Iran using ancillary data variables derived from a DEM and Landsat 8 images. A review by Lamichhane *et al.* (2019) showed that RF was the better-fitting model than multiple linear regression and other machine learning techniques in most comparative studies. While, Rentschler *et al.* (2020) highlighted that the spatial prediction of SOC with cubist provided slightly lower error than with RF in the north of Leipzig, Saxony, Germany. It was generally recognised that application of various techniques will result in different importance of predictors in DSM of continuous attributes for a specific study area (Piikki *et al.*, 2021).

Based on the literature, this research evaluates two machine learning techniques, namely cubist and RF, as candidates for prediction of SOC in southern Spain (Seville) using a set of environmental covariates derived by Landsat-8 OLI, Sentinel-2A Multispectral Instrumental (MSI), DEM, and climate factors such as temperature and precipitation. CONSOLE (Contract Solutions for Effective and lasting delivery of agri-environmental-climate public goods by EU agriculture and forestry, GA 817949) is a H2020 project which focuses on promoting the delivery of Agri-Environmental Climate Public Goods by agriculture and forestry through the development of improving contractual solutions. In the framework of this

project, we have selected olive groves under integrated production as a contract solution that promotes climate regulation through increase of soil organic carbon sequestration.

The aims of this study are: 1) Identify and map SOC between 0–25 cm of depth using satellite multispectral images (Landsat-8 OLI, Sentinel-2A MSI) as well as integration and terrain-related attributes (DEM-derived data), and climatic variables (temperature, precipitation and evapotranspiration) in a Mediterranean area (Seville, southern Spain); 2) Compare the capability of the cubist and RF models in mapping the vertical and spatial distribution of SOC and; 3) Describe the ranking of variables importance in prediction of SOC using the parsimonious model across the study area.

Methods

Area description and summary of methods

The study area is located in Seville, southern Spain, which has a Mediterranean climate. It has a dry summer and mild, wet winter and covers an area of 10000 km^2. The altitude varies from 0 to 1165 m above sea level (Figure 1). Also, the mean annual precipitation, temperature and evapotranspiration for the last 10 consecutive years were measured by the regional administration as 248 mm to 716 mm, 15 °C to 20 °C and 624 mm to 829 mm, respectively (REDIAM, 2021). The olive groves under integrated production are one of the main land use types in the entire study area. Integrated production is an agricultural system of production of plants using farming techniques that ensures sustainable agriculture, using methods of integrated pest management compatible with environmental protection, agricultural productivity and the use of natural production mechanisms and resources described in Hinojosa-Rodríguez *et al.* (2014).

Integrated production includes several rules (Romero-Gámez *et al.*, 2017) for crop, soil and water management in order to increase the sustainability production and other agri-environmental public goods, such as carbon sequestration,

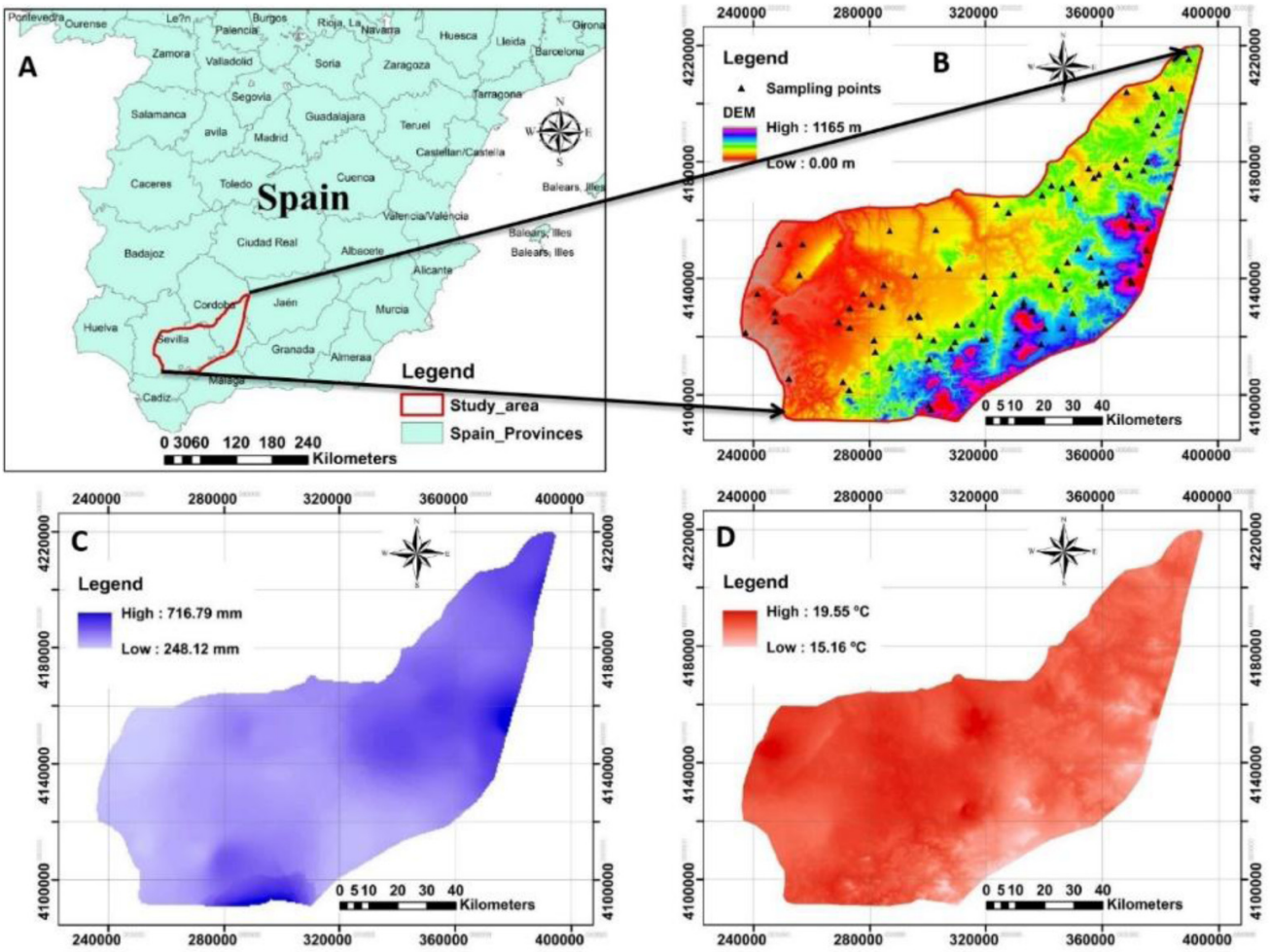

Figure 1. The location and outline of the study area. A: Represents the study area in Spain; **B**: the digital elevation model (DEM) of the study area associated with sampling points; **C** and **D**: represent the spatial maps of precipitation and temperature respectively measured for the 10 last consecutive years across the study area. (Own source).

and promoting the reduction of fertilizers and pesticides. To facilitate the employed procedures across this research, the flowchart is represented in Figure 2.

Field sampling and soil analysis

The first step was compiling all data available for Seville, southern Spain from 1990–2019. For that, three data sources were analysed: 1) REDIAM (Regional Environmental database; 2) SEISNET: Spanish Soil Information System on the Internet and; 3) ASAJA-Sevilla: farmer association from Seville that collect data from farmers and who provided it directly to the authors.

The information compiled (physical and chemical soil variables (SEISNET and ASAJA-Sevilla); site, mean annual precipitation, temperature and evapotranspiration (REDIAM)) were harmonised using R Software (identifying potential duplicates and errors) so that they had the same resolution and extent for analysing the relationship between SOC and other variables in-situ and ex-situ. A total of 22 commercial olive groves included in the data compiled from ASAJA-Sevilla were selected, taking into account their representativeness in terms of soil types and climatic conditions.

SOC was measured in the laboratory in previous research/technical works (this information can be found in the metadata in SEISNET (Information level #3) by wet oxidation with chromic acid and back titration with ferrous ammonium sulphate according to the methods outlined in Nelson and Sommers (1996).

Statistical analysis

Since RF and cubist were used to model and map the spatial distribution of SOC, and the data do not need to be normally distributed (Svetnik *et al.*, 2003), the original data were employed in modelling. However, seven statistical criteria e.g., min, max, mean, standard deviation, coefficient of variation, skewness and kurtosis were calculated for the data set. The aforementioned basic statistics were performed using the package of *"fBasics"* with R language (Wuertz *et al.*, 2017).

Ancillary data

Due to variation on elevation, land uses, parent material and even climate, the spatial distribution of SOC is likely to be predicted as some function of given ancillary data. For this purpose, to gain all required factors identified by *scorpan* model, a DEM, Landsat-8 OLI data and Sentinel-2A MSI imageries spectral data were obtained via USGS-EROS (see *Underlying data*). Climatic variables (precipitation, temperature and evapotranspiration) were also used in this study (obtained from REDIAM, see *Underlying data*) (Table 1). Since all covariates (a total number of 30) have a different resolution,

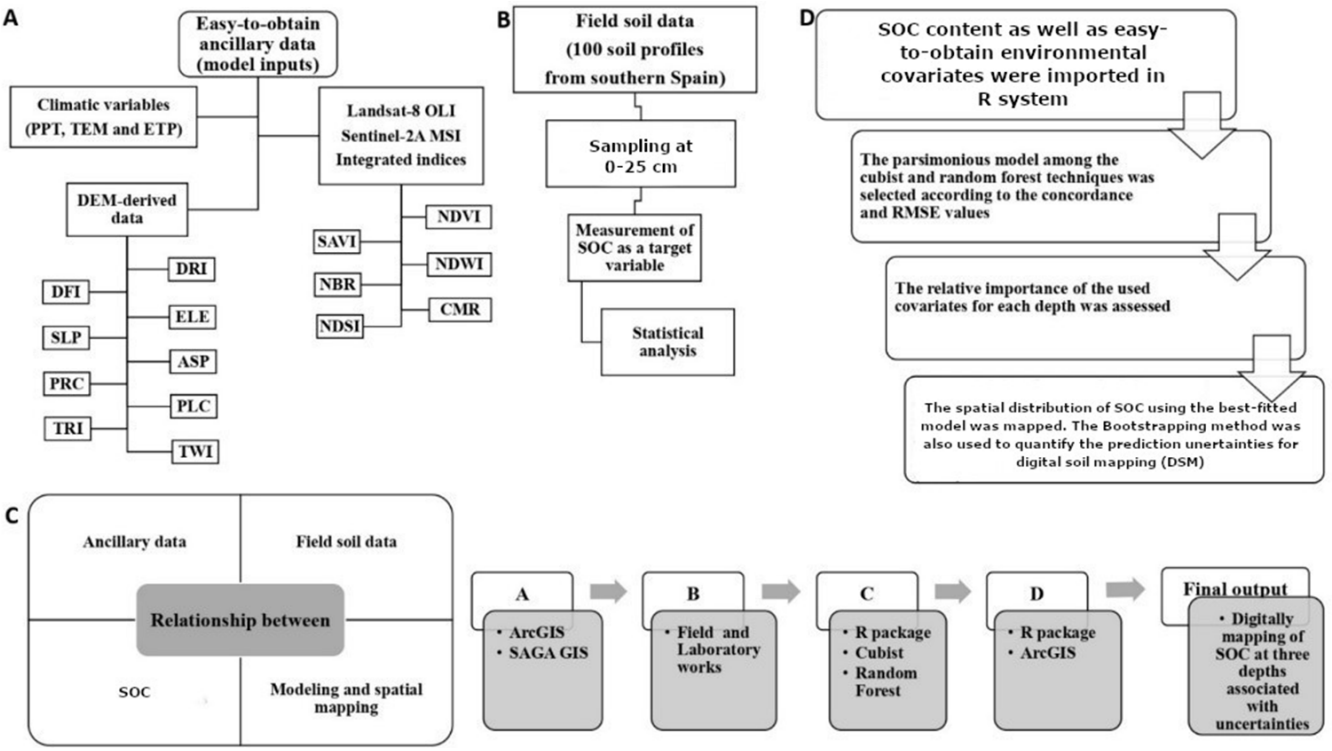

Figure 2. Simplified flowchart of the research in this study. SOC: Soil organic carbon (%); PPT: precipitation; TEM: temperature; ETP: Evapotranspiration; DRI: direct insolation; DFI: diffuse insolation; ELE: elevation; SLP: slope; ASP: aspect; PRC: profile curvature; PLC: plan curvature; TRI: terrain ruggedness index; TWI: topographic wetness index; NDVI: normalized difference vegetation index; SAVI: soil adjusted total vegetation index; NDWI: normalized difference water index; NBR: normalized burn ratio; CMR: clay minerals ratio; NDSI: normalized difference salinity index.

Table 1. The specifications of used indices derived by DEM and remote sensing imageries.

Index	Abbreviation	Description	Derived by	Details and formulation
Direct insolation	DRI	Topography-Potential incoming solar radiation	DEM	Terrain analysis, lightening, visibility
Diffuse insolation	DFI			
Elevation	ELE	Topography-Morphometry	DEM	Terrain analysis
Slope	SLP			
Aspect	ASP			
Profile curvature	PRC			
Plan curvature	PLC			
Terrain ruggedness index	TRI			
Topographic wetness index	TWI	Topography-Hydrology	DEM	Terrain analysis
Normalized difference vegetation index	NDVI_LST	Vegetation	Landsat-8 OLI	$NDVI = \frac{(NIR - Red)}{(NIR + Red)}$
	NDVI_SEN		Sentinel-2 MSI	
	NDVI_ITG		Integrated index	$(\frac{NDVI_{LST} + NDVI_{SEN}}{2}$
Soil Adjusted Total Vegetation Index	SAVI_LST	Vegetation	Landsat-8 OLI	$SAVI = \left(\frac{(NIR - Red)}{NIR + Red + L^{a}}\right) \times (1 + L)$
	SAVI_SEN		Sentinel-2 MSI	
	SAVI_ITG		Integrated index	$(\frac{SAVI_{LST} + SAVI_{SEN}}{2}$
Normalized Difference Water Index	SAVI_LST	Vegetation	Landsat-8 OLI	$NDWI = \frac{(Green - NIR)}{(Green + NIR)}$
	SAVI_SEN		Sentinel-2 MSI	
	SAVI_ITG		Integrated index	$(\frac{NDWI_{LST} + NDWI_{SEN}}{2}$
Normalized Burn Ratio	NBR_LST	Landscape	Landsat-8 OLI	$NBR = \frac{(NIR - SWIR2)}{(NIR + SWIR2)}$
	NBR_SEN		Sentinel-2 MSI	
	NBR_ITG		Integrated index	$(\frac{NBR_{LST} + NBR_{SEN}}{2}$
Clay Minerals Ratio	CMR_LST	Geology	Landsat-8 OLI	$CMR = \frac{SWIR1}{SWIR2}$
	CMR_SEN		Sentinel-2 MSI	
	CMR_ITG		Integrated index	$(\frac{CMR_{LST} + CMR_{SEN}}{2}$
Normalized Difference Salinity Index	NDSI_LST	Geology	Landsat-8 OLI	$NDSI = \frac{(Red - NIR)}{(Red + NIR)}$
	NDSI_SEN		Sentinel-2 MSI	
	NDSI_ITG		Integrated index	$(\frac{NDSI_{LST} + NDSI_{SEN}}{2}$
Precipitation	PPT	Climate	information (climate) layers of Andalusia, Spain https://bit.ly/3CX7zOM https://bit.ly/3zS30Dm	
Temperature	TEM			
Evapotranspiration	ETP			

[a] The L value varies depending on the amount of green vegetative cover (L=0.5)

those were initially aligned to the same grid cell resolution and extent as exemplified by Malone *et al*. (2017). Here, a 30m grid was used. The coordinate reference system used in this study was WGS1984 UTM Zone 30.

***DEM-derived data*.** The terrain-related attributes were derived from the original DEM layer at a resolution of 30 m using various functions (slope gradients, profile, planform curvatures and so on) made available in SAGA GIS version 3.5.1 (Conrad *et al*., 2015). The terrain analysis allows accurate routing of surface and subsurface flow through the landscape (Ranatunga *et al*., 2008) which can be rapidly made with application of SAGA GIS. In this research, in addition to elevation (ELE) which was directly derived by DEM, eight derivatives were also classified under three broad categories. 1) Lighting or visibility: with derivatives including potential incoming solar radiation (direct and diffuse insolation abbreviated by DIR and DFI, respectively). 2) Morphometry: with derivatives including slope (SLP), aspect (ASP), curvature (plan and profile abbreviated by PLC and PRC, respectively), and terrain ruggedness index (TRI). 3) Hydrology: includes the topographic wetness index (TWI).

The DEM-derivatives relevant to morphometry e.g., SLP, ASP, PLC, PRC and TRI were numerously used in digital mapping of SOC (She *et al*., 2014). Kravchenko & Bullock (2000) have defined these indices in a detailed manner.

The third category identifying the terrain-related attributes is hydrology. Topographic wetness index (TWI) was used in this work. Schwanghart and Jarmer (2011) found that there was a positive correlation between TWI and SOC even on steep slopes, and that correlation is not significant on wide pediments. To manually gain the TWI, it is necessary to calculate the specific upslope contributing area and slope while it was easily earned using SAGA GIS across the study area.

***Remotely-sensed data*.** Landsat-8 OLI and Sentinel-2A MSI images acquired in July 2019 (see *Underlying data*) were selected for further analysis in this project due to the high probability of no clouds in the images. Fortunately, the selected scenes comprise minimal cloud coverage and maximum soil surface exposure. On contrary of the previously performed research (such as Shahbazi *et al*., 2019a), the individual bands and Principal Component Analyses (PCA) of bands were not used here. To explain the vegetation, soil and water, landscape as well as geology of the study area, six band ratios were calculated using both images as well as their integrations for each index. The hypothesis of integrating the indices has been raised from the report of Wang *et al*. (2020) which demonstrated that combining Landsat TM and ALOS PALSAR images could predict more accurate SOC content, especially for soils with high vegetation canopy density in Spain. Those were normalized difference vegetation index (NDVI) (Rouse *et al*., 1973), soil adjusted vegetation index (SAVI) (Huete, 1988), normalized difference water index (NDWI) (McFeeters, 1996), normalized burn ratio (NBR) (Lanorte *et al*., 2013), clay minerals ratio (CMR) (Drury, 2016), and normalized difference salinity index (NDSI) (Allbed & Kumar, 2013).

***Climatic variables*.** To predict the spatial distribution of SOC, it was found that despite topographic variables which had higher influence at finer scales, the climatic variables were more important at coarser scales e.g., sub-regional to global scales (Adhikari *et al*., 2020). Among the previously described indices, climate variables contributed the most to explaining SOC content variation (Zhou *et al*., 2020). Literature revealed that climate significantly controls the spatial pattern of SOC and its stock in soil. Therefore, precipitation, temperature and evapotranspiration at a 100-m resolution were used as the main attributes in prediction of SOC across the study area.

Machine learning techniques

Regarding the accelerated adoption of machine learning (ML) techniques in prediction of soil properties, especially in the last 10 years (Padarian *et al*., 2020), two methods, namely cubist and random forest (RF) were tested within R programming (version 3.5.1) for the prediction of SOC at three depths entire study area using our field soil data set and environmental covariates. Compared with the cubist, the difference between lower and upper quartiles in the RF is larger. Furthermore, RF tends to perform better in different iterations (Zhou *et al*., 2019). Khaledian and Miller (2020) reported that the results taken by cubist are more interpretable than RFs' which are semi-interpretable. We therefore performed both models, then one of them was selected as a parsimonious model to be applied in spatial mapping. The bootstrap aggregation or bagging that reduces variance within a noisy data set (Breiman, 1996), was also set on 200 in this work. In this way the stability and accuracy of both models will be improved. The efficiency of tested models was evaluated in a calibration and validation data set (composed by 20% of data no included in the modelling process) using the goodness-of-fit criteria e.g., concordance (CCC) and root mean square error (RMSE). It is clear that R^2 measures the precision of the relationship (between observed and predicted) while concordance as a single statistic evaluates the accuracy and precision of the relationship. Similar work was fully explained by Malone *et al*. (2017) using the "*goof*" function within the "*ithir*" package in R (Malone, 2016).

***Cubist*.** Cubist is a rule-based model that has been numerously applied for prediction of SOC (Akpa *et al*., 2016; Miller *et al*., 2015) with success. The tuning parameters in the cubist model are committees and neighbours. Regarding these parameters, it may improve the prediction performance due to its ability to "mine" non-linear relationships in data. The path through the tree was collapsed to create rules using boosting training. The rules, extrapolation and committee were set on 5, 10, and 5, respectively in the used script with the "*Cubist*" package (Kuhn & Quinlan, 2018).

***Random forest*.** The other most commonly used ML technique in prediction of SOC is RF (Dharumarajan *et al*., 2017; Grimm *et al*., 2008; Wiesmeier *et al*., 2011). Mapping using RF creates the spatial distribution of SOC in a quicker way because fitting a RF model in R is relatively straightforward. The tuning parameter in the RF model is the number of trees (Liaw & Wiener, 2002) which were set on 1000 in this research using the package of "*randomForest*" in R

programming. In this technique, a group of observations from the training dataset (20% of data) was selected and built a decision tree associated with those selected observations. The calibration and validation of the model were derived using in-the-bag and out-of-bag (OOB). A total of 200 bootstrap samples were supplied to model due to setting the bags on 200 iterations. Since the arrangement of 30 used environmental covariates is randomly ordered, all possibilities were therefore considered in selection of covariates.

Acquiring spatial maps

In addition to the rich capacity of modelling in R, it has a user-friendly environment to map spatial data (Malone *et al.*, 2017). Bivand *et al.* (2013) documented some methods for handling spatial data and analysis in R in a detailed manner. The spatial mapping was facilitated using the "*ggplot2*" package in R (Wickham, 2016) as a way to create codes that make sense to the user. It is obvious that both point data and covariates should have the same projection in advance to be easily import, view, and export points and rasters to, in, and from a GIS. The final digital maps (Figure 6, Figure 7) will be as the main output for further visualization as well as interpretation in digital assessments.

Uncertainty analysis

Arrouays *et al.* (2020) reported that uncertainty maps may give very strong arguments in DSM because the identifying missing soil data and excessive field work will be distinguished. On the other hand, the quantified uncertainty analysis is essential to be associated with digital maps because they come with some errors (Wadoux *et al.*, 2020). Various methods depending on the objectives were used in such analysis e.g., Monte Carlo (Tan *et al.*, 2014); Bayesian Network (Stritih *et al.*, 2019); and computing a vector total of RMSE (Weng, 2002).

In our research, the next step was to calculate mean values by taking the average of all the simulated predictions at each pixel using the base function "*mean*" to a given stack of rasters. Since there is no simple function to estimate the variance, the sum of square differences was firstly calculated at each pixel from the prediction maps. The maps of prediction interval (PI) range for SOC were also provided by calculating the difference between upper and lower 95% limit of predictions (Xiong *et al.*, 2015). The high value of PI at any pixel demonstrates the low level of confidence. For further visualisation, the standard deviation and standard error of predictions at any depth were also associated with each digital map.

Results and discussion

Brief soil data

The soil data compiled across the study area illustrates OC which may be related to natural processes such as alluvial deposition. Caro Gomez *et al.* (2011) reported that alluvial episodes with associated lithic techno-complexes were identified across the Guadalquivir River valley (Seville, Spain). Variations may have occurred due to different soil types e.g., Vertisols, Nitisols, Luvisols, etc in the entire study area. In fact, the impact of Integrated Production was analysed comparing SOC data previous Integrated Production implementation (2005) and current SOC data. The results showed the increase in SOC due to the implementation of Integrated Production in terms of crop and soil management. (Figure 3).

Terrain-related attributes

A total of nine DEM-derived data were supplied as predictors relevant to the lighting, morphometry and hydrology. Four examples of spatial distribution of DRI, DFI, TRI and TWI provided by SAGA GIS were presented in Figure 4 to facilitate in interpreting the study's area condition. These covariates allow the analysis of SOC across case study areas and their relation with terrain attributes. In general, there was a different gradient of aforementioned covariates. DRI and DFI mean values varied from zero to 3.96 (1.66, on average) and 0.46 to 0.72 (0.70, on average). Also, the values of SD for both covariates were 0.29 and 0.01, respectively. In terms of TRI,

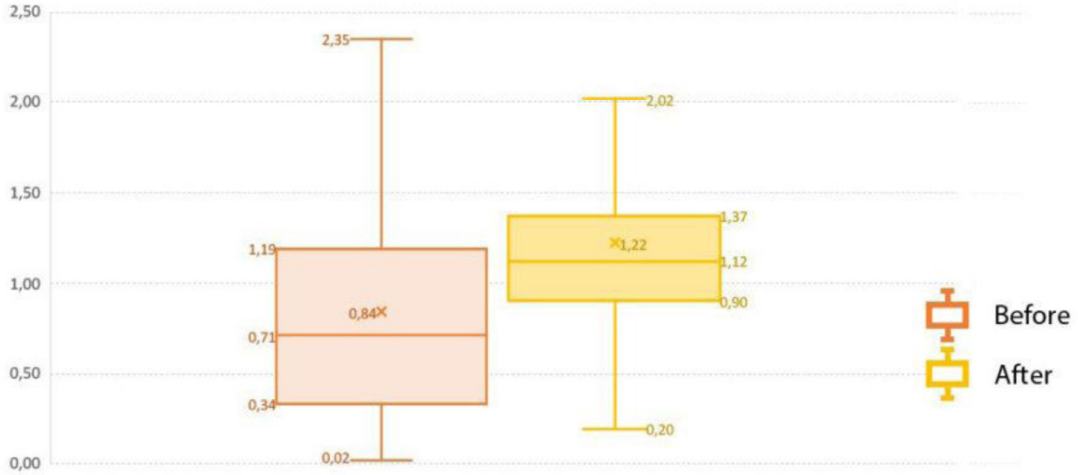

Figure 3. Comparison of soil organic carbon before (1990–2000) from REDIAM and after (2010–2020) from ASAJA-Sevilla applying the Integrated Production to representative Mediterranean benchmark soils under olive crop.

it has been developed by Riley *et al*. (1999) as an index to quantify the topographic heterogeneity from level to extremely rugged by seven classes. It is interesting that TRI varied from zero to 55.69 indicating the level terrain surface across the study area. The mean value and SD were 1.82 and 1.55, respectively. The fourth studied covariate was TWI. It varies from 1.01 to 11.06 with a mean value of 5.43 across the study area. The ability of TWI in prediction of soil moisture and plant assemblages was recently reported by Kopecký *et al*. (2021). Similar observations were performed for the rest of the terrain-related attributes.

Remotely-sensed data

The next step was to acquire both Landsat-8 OLI and Sentinel-2A MSI imageries. To monitor the spatial distribution of each index calculated by each image, NBR and NDVI were illustrated in Figure 5.

The original raster files in QGIS showed that the calculated mean values for NBR derived from Landsat-8 and Sentinel-2A were no longer different (approximately 0.12, on average). However, Landsat-based NDVI did differ significantly from Sentinel-based NDVI. Further analysis by converting the NDVI maps from stretch to classified mode has provided an opportunity to identify bare soils, water and very low to high dense plant covers entire area. This idea was previously reported by Julien *et al*. (2011) in observation of land surface temperature for detecting changes in the Iberian land covers, Spain. The map of NDVI_LST strongly implies that about 0.2% of the study area has been categorized as the bare soil or water. The highest area (70%) was classified as very low cover (0<NDVI<0.2), followed by low cover (29%, 0.2<NDVI<0.4), moderately cover (0.7%, 0.4<NDVI<0.6) and moderately high cover (0.1%, 0.6<NDVI<0.8). Similar classification on the NDVI_SEN showed that 46%, 44%, 8% and 1% of the study area have very low, low, moderately low and moderately high cover, respectively. The results showed a significant increase in overall classification when the NDVI derived from Sentinel-2A compared to Landsat-8. One interpretation of this finding is that the launch of Sentinel-2A

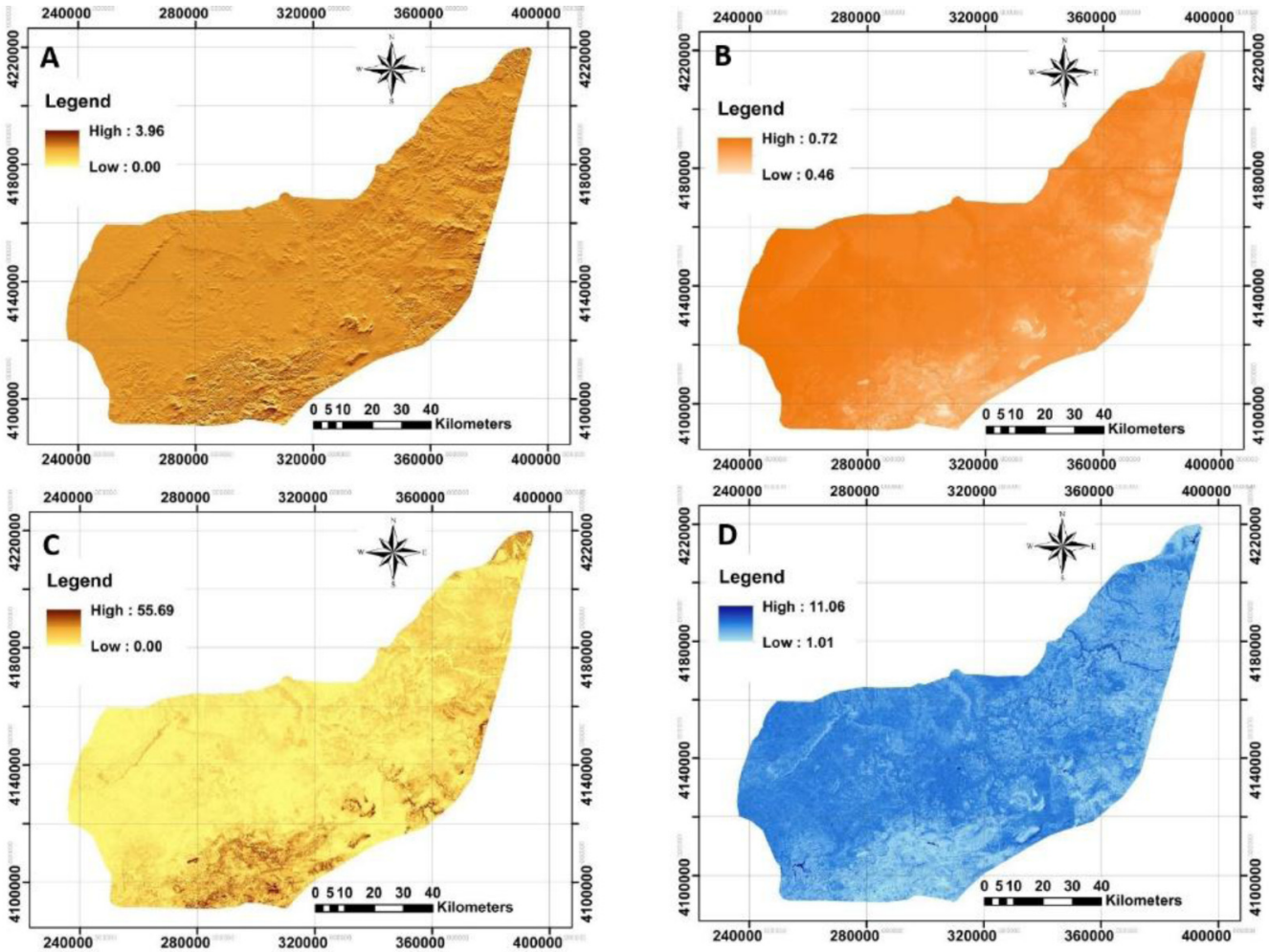

Figure 4. Four examples of applied terrain-related attributes to model across the study area. A: Direct insolation (DRI); **B**: Diffuse insolation (DFI); **C**: terrain ruggedness index (TRI); **D**: topographic wetness index (TWI). (Own source).

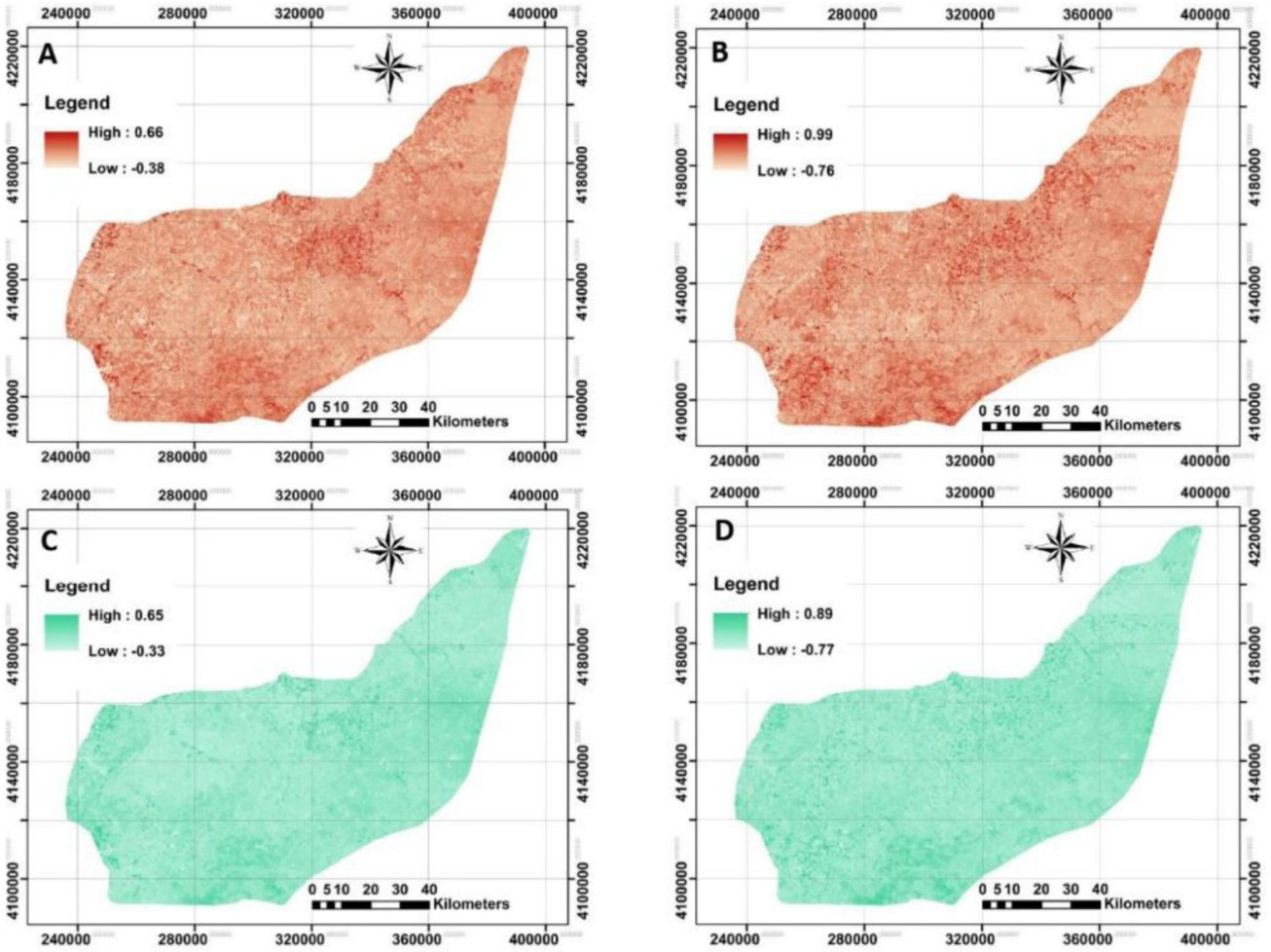

Figure 5. Four examples of applied Remote Sensing data using Landsat-8 OLI and Sentinel-2A MSI data to model across the study area. A and **B**: represent Normalized burn ratio (NBR) derived from Landsat-8 and Sentinel-2A, respectively; **C** and **D**: represent normalized difference vegetation index (NDVI) derived from Landsat-8 and Sentinel-2A, respectively. (Own source).

satellites has boosted the development of many applications that could benefit from the fine resolution of the supplied information (Chakhar *et al*., 2020). These kinds of observations were performed for the remaining covariates, but are not fully presented here.

Final machine learning technique

Machine learning allowed us to analysis the huge dataset and their relevance to SOC in the different circumstances. For that, the use of open software like as R facilitated modeling and then mapping. For this study, the statistical criteria e.g., CCC, RMSE and bias for calibration and validation dataset using two machine leaning techniques were summarized in Table 2 and Table 3.

Finally, we deduced that RF outperformed cubist in prediction accuracy. Since RF was selected as a good model in predicting, therefore, the importance of covariates was ranked for each layer (Figure 6). The results generally revealed that Landsat -based indices are not better predictors. In addition, the climate data were not as important as other supplied covariates. A possible reason is that the topography and plant cover were more variable than climate over the study area. The integrated indices e.g., NBR_ITG, NDWI_ITG and CMR_ITG were identified as the most important covariates in prediction of $SOC_{0–25cm}$. These findings showed that the idea of combining indicators was an interesting idea and presented a good result. Similar work has been successfully conducted by Quintano *et al*. (2018) in Sierra de Gata (central-western Spain) for initial assessment of burn severity. The novelty of the approach was also reported by Singh *et al*. (2020) in water-use mapping. As well as, SAVI_SEN, NDWI_SEN and CMR_SEN showed good results in the depth analysed. This remark may suggest that Sentinel data is the most influential data in the

integrated indexes. In our view, the most compelling explanation for the present set of findings is that integrated-based indices were in the top ranks for predit $SOC_{0–25cm}$. This finding may be explained by the idea that SOC content near the surface is linked with water availability or other dynamic properties, related directly with the spectral information of the surfaces. Furthermore, variables as land use, physical site factors or temperature are linked with bottom soil part dynamic (Hobley & Wilson, 2016).

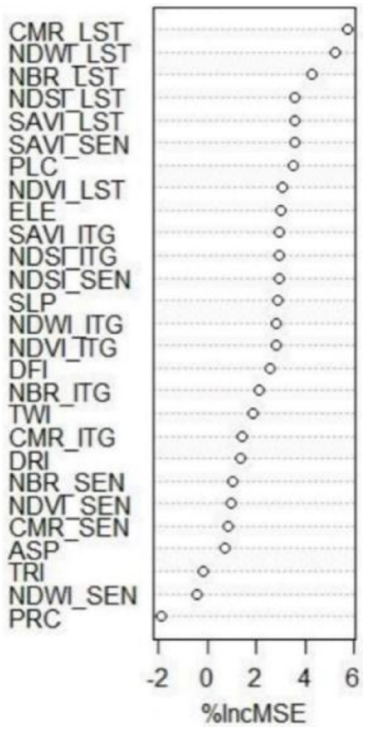

Figure 6. The importance ranking of used environmental covariates (n=30) for predicting SOC across the study area for any depth SOC0-25cm. The importance ranking of used environmental covariates (n=30) for predicting SOC across the study area for SOC0–25cm._LST: Indices derived from Landsat-8 OLI; _SEN: indices derived from Sentinel-2A MSI; _ITG: An index illustrating the average of Landsat-8 OLI and Sentinel-2A MSI relevant to each index; PPT: precipitation; TEM: temperature; ETP: Evapotranspiration; DRI: direct insolation; DFI: diffuse insolation; ELE: elevation; SLP: slope; ASP: aspect; PRC: profile curvature; PLC: plan curvature; TRI: terrain ruggedness index; TWI: topographic wetness index; NDVI: normalized difference vegetation index; SAVI: soil adjusted total vegetation index; NDWI: normalized difference water index; NBR: normalized burn ratio; CMR: clay minerals ratio; NDSI: normalized difference salinity index.

Digital maps and quantified uncertainties

The next step was to provide digital maps of SOC for selected depth associated with its uncertainty using bootstrapping method (Figure 7). The standard error and variance maps of predictions were also illustrated.

The next step was to quantify the map's accuracy using the bootstrapping method. The pixel-by-pixel values of standard error and variance at any depth were illustrated before calculating the prediction interval ranges (Figure 7). These results support that the SOC values in upper layers are affected by crop or farm activities (Montes-Pulido *et al.*, 2016; Zhu *et al.*, 2017) and, for that, crop monitoring through satellites is feasible for monitoring SOC in topsoil. In general, the high values of uncertainties were found in the south of the study area.

Conclusions

This study aims to reveal the usage of DSM and its importance in predicting and spatial distribution of SOC using machine learning techniques. In general, the results showed that random forest outperformed cubist in all predictions. Furthermore, the spatial distribution SOC was successfully mapped using integrated indices as a novel idea.

Although this research focuses on the mapping of SOC at the top soil. These lack of harmonised data is a barrier to develop comparative analysis between data sources so harmonisation and standardisation processes are required. A second potential limitation is that because there was no data on bulk density, the soil carbon storage was not calculated to monitor the climate change impact. Finally, the last potential limitation may be the influence of meteorological conditions on the satellite imageries used. We suggest using a time series of finer integrated remotely-sensed data in the future.

The analysis of SOC on different soil types allows enhancing of results. The terrain attributes analysis was a key step in order to obtain SOC data. Taken together, our findings indicate the successful usage of machine learning with application of user-friendly software e.g., R programming.

The use of satellites for monitoring SOC trends is feasible for soil top layers due to crop impact on it. In addition, several agricultural policies and/or contract solutions such as new CAP

Table 2. The statistical descriptive of used field observations across the study area (n=100).

	min	max	mean	SD	CV (%)	skewness	kurtosis
$SOC_{0–25cm}$	0.02	2.63	0.81	0.53	65.08	-0.29	0.99

SOC: soil organic carbon (%); SD: standard deviation; CV: coefficient of variation.

Table 3. The performance of candidate models for estimating SOC across the study area (n=100).

Calibration						
	Concordance		Root mean square error		Bias	
	Random Forest	Cubist	Random Forest	Cubist	Random Forest	Cubist
$SOC_{0\text{-}25cm}$	0.86	0.32	0.22	0.46	0.00	-0.06
Validation						
$SOC_{0\text{-}25cm}$	0.15	0.07	0.54	0.55	0.02	-0.08

SOC: soil organic carbon.

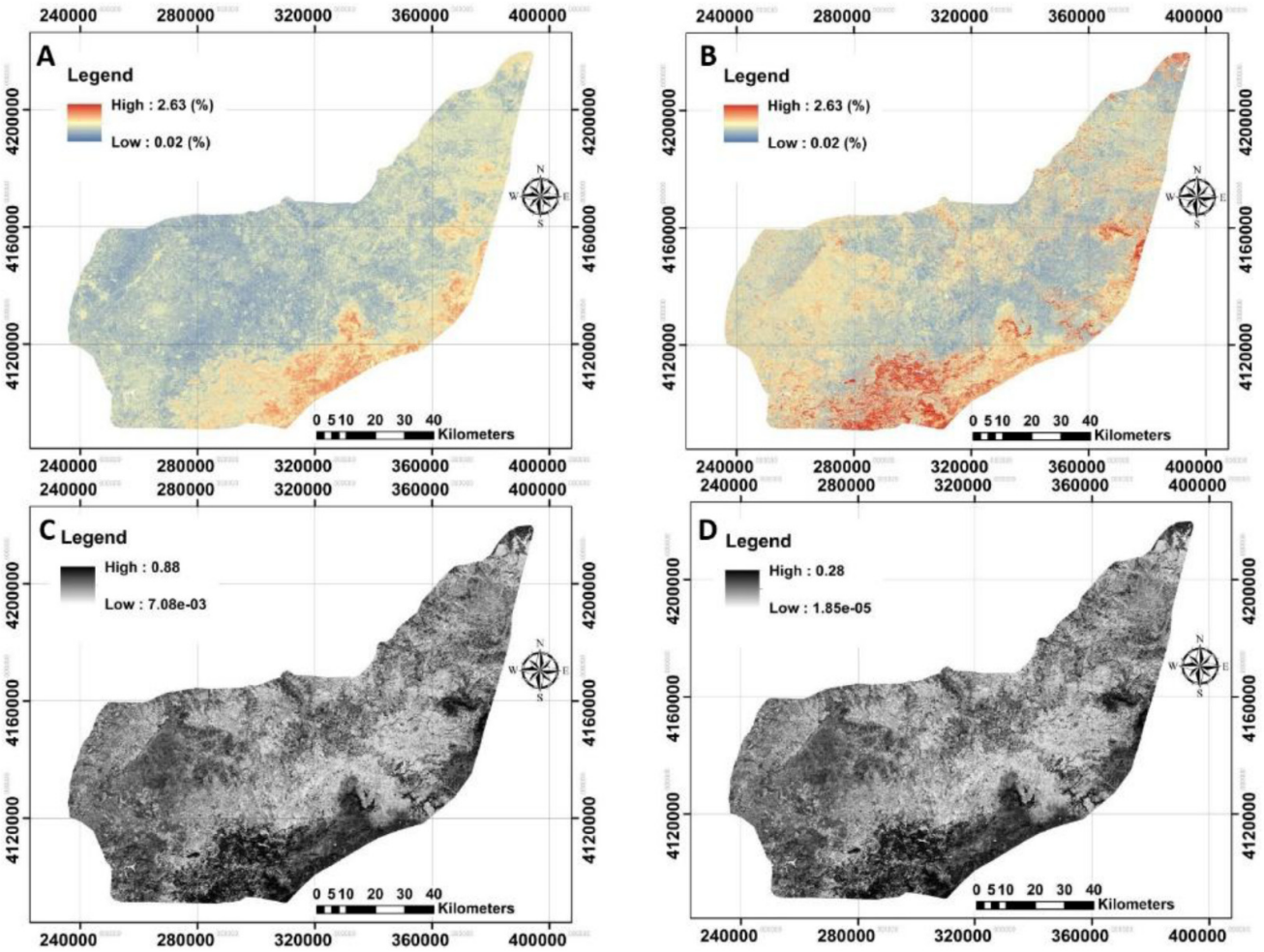

Figure 7. The provided digital maps of SOC0–25cm across the study area using RF model. A: Mean prediction (%); **B**: prediction of interval range (%); **C**: standard error map of the prediction; **D**: variance map of the prediction. (Own source).

(Common Agricultural Policies), Integrated production, Ecologic production and so on, may be beneficial for this methodology in order to reduce costs in hot-spot checks. Furthermore, the technology proposed may support the development of new contract solutions based on results within a payment per carbon sequestered.

Ethics and consent

Ethical approval and consent were not required.

Data availability

Source data

- The Landsat-8 OLI images used in the study are publicly available for download here: https://www.usgs.gov/landsat-missions/landsat-data-access. All pictures from July 2019 coincident to the study area were used.

- The Sentinel-2A MSI images used in the study are publicly available for download here: https://scihub.copernicus.eu/. All pictures from July 2019 coincident to the study area were used.

- The DEM, soil maps and climate variables used in this study are available on REDIAM: https://descargasrediam.cica.es/repo/s/RUR. The data can be found categorized by topic in 01_CARACTERIZACION_TERRITORIO/07_BASES_REF_ELEV (DEM data) and 16_INDICADORES_ESTADISTICAS/05_DATOS_BASICOS (climate data). The reader can download the data in several formats (Access, shapes, WMS, *etc*).

- SEISNET: https://microleis.evenor-tech.com/banco/seisnet/seisnet.htm In SEISNET platform, soil data can be downloaded at different details scale. T data used in this research work can be downloaded in the subsection 'Nivel de información #3' (registration is required).

Underlying data

Open Science Framework: CONSOLE. https://doi.org/10.17605/OSF.IO/B7S4Z (Velázquez, 2022).

This project contains the following underlying data in the file *source_data.rar*:

- Rasters (indexes obtained from Sentinel and Landsat satellites).

- Field_data.xls (physical and chemical soil variables, climate variables and DEM obtained from REDIAM, SEISNET and ASAJA-SEVILLA used in this work).

Data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

Acknowledgments

The authors would like to thank both the University of Tabriz and Evenot-Tech (Spain) for their collaboration.

References

Adhikari K, Mishra U, Owens PR, *et al.*: **Importance and strength of environmental controllers of soil organic carbon changes with scale.** *Geoderma.* 2020; **375**: e114472.
Publisher Full Text

Akpa SI, Odeh IO, Bishop TF, *et al.*: **Total soil organic carbon and carbon sequestration potential in Nigeria.** *Geoderma.* 2016; **271**: 202–215.
Publisher Full Text

Allbed A, Kumar L: **Soil salinity mapping and monitoring in arid and semi-arid regions using remote sensing technology: A review.** *Advances in Remote Sensing.* 2013; **2**(4): 373–385.
Publisher Full Text

Arrouays D, McBratney A, Bouma J, *et al.*: **Impressions of digital soil maps: The good, the not so good, and making them ever better.** *Geoderma Regional.* 2020; **20**: e00255.
Publisher Full Text

Bailey VL, Pries CH, Lajtha K: **What do we know about soil carbon destabilization?** *Environ Res Lett.* 2019; **14**(8): e083004.
Publisher Full Text

Bivand RS, Pebesma E, Gómez-Rubio V: **Further methods for handling spatial data.** In: *Applied Spatial Data Analysis with R.* Use R!, Springer, New York, NY, 2013; **10**: 127–150.
Publisher Full Text

Blanco-Velázquez FJ, Muñoz-Vallés S, Anaya-Romero M: **Assessment of sugar beet lime measure efficiency for soil contamination in a Mediterranean Ecosystem. The case study of Guadiamar Green Corridor (SW Spain).** *Catena.* 2019; **178**: 163–171.
Publisher Full Text

Breiman L: **Bagging predictors.** *Mach Learn.* 1996; **24**: 123–140.
Publisher Full Text

Caro Gómez JA, Díaz Del Olmo F, Artigas RC, *et al.*: **Geoarchaeological alluvial terrace system in Tarazona: Chronostratigraphical transition of Mode 2 to Mode 3 during the middle-upper pleistocene in the Guadalquivir River valley (Seville, Spain).** *Quat Int.* 2011; **243**(1): 143–160.
Publisher Full Text

Carré F, McBratney AB, Mayr T, *et al.*: **Digital soil assessments: beyond DSM.** *Geoderma.* 2007; **142**(1–2): 69–79.
Publisher Full Text

Chakhar A, Ortega-Terol D, Hernández-López D, *et al.*: **Assessing the accuracy of multiple classification algorithms for crop classification using Landsat-8 and Sentinel-2 data.** *Remote Sens.* 2020; **12**(11): e1735.
Publisher Full Text

Conrad O, Bechtel B, Bock M, *et al.*: **System for Automated Geoscientific Analyses (SAGA) v. 2.1.4.** *Geosci Model Dev.* 2015; **8**(7): 1991–2007.
Publisher Full Text

Dharumarajan S, Hegde R, Singh SK: **Spatial prediction of major soil properties using Random Forest techniques - A case study in semi-arid tropics of South India.** *Geoderma Regional.* 2017; **10**: 154–162.
Publisher Full Text

Drury SA: **Image Interpretation in Geology, second ed.** Nelson Thornes, London, 2016.

European Commission: **EU Soil Strategy for 2030.** Reaping the benefits of healthy soils for people, food, nature and climate. 2021.
Reference Source

Gooley L, Huang J, Pagé D, *et al.*: **Digital soil mapping of available water content using proximal and remotely sensed data.** *Soil Use Manag.* 2014; **30**(1): 139–151.
Publisher Full Text

Grimm R, Behrens T, Märker M, *et al.*: **Soil organic carbon concentrations and stocks on Barro Colorado Island - Digital soil mapping using Random Forests analysis.** *Geoderma.* 2008; **146**(1–2): 102–113.
Publisher Full Text

Hinojosa-Rodríguez A, Parra-López C, Carmona-Torres C, *et al.*: **Certified quality systems and farming practices in olive growing: The case of integrated production in Andalusia.** *Renewable Agriculture and Food Systems.* 2014; **29**(4): 291–309.
Publisher Full Text

Hobley EU, Wilson B: **The depth distribution of organic carbon in the soils of eastern Australia.** *Ecosphere.* 2016; **7**(1): e01214.
Publisher Full Text

Huete AR: **A soil-adjusted vegetation index (SAVI).** *Remote Sens Environ.* 1988; **25**(3): 295–309.
Publisher Full Text

Julien Y, Sobrino JA, Mattar C, *et al.*: **Temporal analysis of normalized difference vegetation index (NDVI) and land surface temperature (LST)**

parameters to detect changes in the Iberian land cover between 1981 and 2001. *Int J Remote Sens.* 2011; **32**(7): 2057–2068.
Publisher Full Text

Khaledian Y, Miller BA: **Selecting appropriate machine learning methods for digital soil mapping.** *Appl Math Model.* 2020; **81**: 401–418.
Publisher Full Text

Kopecký M, Macek M, Wild J: **Topographic Wetness Index calculation guidelines based on measured soil moisture and plant species composition.** *Sci Total Environ.* 2021; **757**: e143785.
PubMed Abstract | Publisher Full Text

Kravchenko AN, Bullock DG: **Correlation of corn and soybean grain yield with topography and soil properties.** *Agron J.* 2000; **92**(1): 75–83.
Publisher Full Text

Kuhn M, Quinlan R: **Cubist: Rule- And Instance-Based Regression Modeling.** R package version 0.2.2. 2018.
Reference Source

Lamichhane S, Kumar L, Wilson B: **Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review.** *Geoderma.* 2019; **352**: 395–413.
Publisher Full Text

Lanorte A, Danese M, Lasaponara R, *et al.*: **Multiscale mapping of burn area and severity using multisensor satellite data and spatial autocorrelation analysis.** *Int J Appl Earth Obs Geoinf.* 2013; **20**: 42–51.
Publisher Full Text

Liaw A, Wiener M: **Classification and regression by randomForest.** *R News.* 2002; **2**(3): 18–22.
Reference Source

Malone B: **ithir: Functions and algorithms specific to pedometrics.** R package version 1.0/r126. 2016.
Reference Source

Malone BP, Styc Q, Minasny B, *et al.*: **Digital soil mapping of soil carbon at the farm scale: A spatial downscaling approach in consideration of measured and uncertain data.** *Geoderma.* 2017; **290**: 91–99.
Publisher Full Text

McBratney AB, Mendoça Santos ML, Minasny B: **On digital soil mapping.** *Geoderma.* 2003; **117**(1–2): 3–52.
Publisher Full Text

McFeeters SK: **The use of the normalized difference water index (NDWI) in the delineation of open water features.** *Int J Remote Sens.* 1996; **17**(7): 1425–1432.
Publisher Full Text

Miller BA, Koszinski S, Wehrhan M, *et al.*: **Comparison of spatial association approaches for landscape mapping of soil organic carbon stocks.** *Soil.* 2015; **1**(1): 217–233.
Publisher Full Text

Minasny B, McBratney AB: **Digital soil mapping: a brief history and some lessons.** *Geoderma.* 2016; **264**(Part B): 301–311.
Publisher Full Text

Minasny B, McBratney AB, Malone BP, *et al.*: **Chapter One - Digital mapping of soil carbon.** *Advances in Agronomy.* 2013; **118**: 1–47.
Publisher Full Text

Montes-Pulido CR, Ramos-Miras JJ, San José-Wery A: **Estimation of soil organic carbon (SOC) at different soil depths and soil use in the Sumapaz paramo, Cundinamarca - Colombia.** *Acta Agronómica.* 2016; **66**(1): 95–101.
Publisher Full Text

Mousavi A, Shahbazi F, Oustan S, *et al.:* **Spatial distribution of iron forms and features in the dried lake bed of Urmia Lake of Iran.** *Geoderma Regional.* 2020; **21**: e00275.
Publisher Full Text

Muñoz-Rojas M, Jordán A, Zavala LM, *et al.*: **Organic carbon stocks in Mediterranean soil types under different land uses (Southern Spain).** *Solid Earth.* 2012; **3**(2): 375–386.
Publisher Full Text

Muñoz-Rojas M, Jordán A, Zavala LM, *et al.*: **Modelling soil organic carbon stocks in global change scenarios: a CarboSOIL application.** *Biogeosciences.* 2013; **10**(12): 8253–8268.
Publisher Full Text

Nelson DW, Sommers LE: **Total carbon, organic carbon, and organic matter.** In: Sparks D.L. (Ed.), *Methods of Soil Analysis. Chemical Methods.* Part 3. ASA, CSSA, and SSSA, Madison, WI. 1996; 961–1010.
Publisher Full Text

Odebiri O, Mutanga O, Odindi J, *et al.*: **Predicting soil organic carbon stocks under commercial forest plantations in KwaZulu-Natal province, South Africa using remotely sensed data.** *GIsci Remote Sens.* 2020; **57**(4): 450–463.
Publisher Full Text

Padarian J, Minasny B, McBratney AB: **Machine learning and soil sciences: a review aided by machine learning tools.** *Soil.* 2020; **6**(1): 35–52.
Publisher Full Text

Piikki K, Wetterlind J, Söderström M, *et al.*: **Perspectives on validation in digital soil mapping of continuous attributes—A review.** *Soil Use Manag.* 2021; **37**(1): 7–21.
Publisher Full Text

Pino-Mejías R, Cubiles-de-la-Vega MD, Anaya-Romero M, *et al.*: **Predicting the potential habitat of oaks with data mining models and the R system.** *Environ Model Softw.* 2010; **25**(7): 826–836.
Publisher Full Text

Quintano C, Fernández-Manso A, Fernández-Manso O: **Combination of Landsat and Sentinel-2 MSI data for initial assessing of burn severity.** *Int J Appl Earth Obs Geoinf.* 2018; **64**: 221–225.
Publisher Full Text

Ranatunga K, Nation ER, Barratt DG: **Review of soil water models and their applications in Australia.** *Environ Model Softw.* 2008; **23**(9): 1182–1206.
Publisher Full Text

REDIAM: **Environmental Information Network of Andalusia**. 2021.
Reference Source

Rentschler T, Werban U, Ahner M, *et al.*: **3D mapping of soil organic carbon content and soil moisture with multiple geophysical sensors and machine learning.** *Vadose Zone J.* 2020; **19**(1): e20062.
Publisher Full Text

Riley SJ, DeGloria SD, Elliot R: **A terrain ruggedness index that quantifies topographic heterogeneity.** *Intermt J Sci.* 1999; **5**(1–4): 23–27.
Reference Source

Romero-Gámez M, Castro-Rodríguez J, Suárez-Rey EM: **Optimization of olive growing practices in Spain from a life cycle assessment perspective.** *J Clean Prod.* 2017; **149**: 25–37.
Publisher Full Text

Rosero-Vlasova OA, Vlassova L, Pérez-Cabello F, *et al.*: **Soil organic matter and texture estimation from visible-near infrared-shortwave infrared spectra in areas of land cover changes using correlated component regression.** *Land Degrad Dev.* 2019; **30**(5): 544–560.
Publisher Full Text

Rouse JW, Haas RH, Schell JA, *et al.*: **Monitoring vegetation systems in the Great Plains with ERTS.** 3rd ERTS Symposium, NASA SP-351, Washington DC 10– 14 December. 1973; 309–317.
Reference Source

Schwanghart W, Jarmer T: **Linking spatial patterns of soil organic carbon to topography a case study from south-eastern Spain.** *Geomorphology.* 2011; **126**(1–2): 252–263.
Publisher Full Text

Shahbazi F, Hughes P, McBratney AB, *et al.*: **Evaluating the spatial and vertical distribution of agriculturally important nutrients — nitrogen, phosphorous and boron — in North West Iran.** *Catena.* 2019a; **173**: 71–82.
Publisher Full Text

Shahbazi F, McBratney A, Malone B, *et al.*: **Retrospective monitoring of the spatial variability of crystalline iron in soils of the east shore of Urmia Lake, Iran using remotely sensed data and digital maps.** *Geoderma.* 2019b; **337**: 1196–1207.
Publisher Full Text

She D, Xuemei G, Jingru S, *et al.*: **Soil organic carbon estimation with topographic properties in artificial grassland using a state-space modeling approach.** *Can J Soil Sci.* 2014; **94**(4): 503–514.
Publisher Full Text

Singh RK, Khand K, Kagone S, *et al.*: **A novel approach for next generation water-use mapping using Landsat and Sentinel-2 satellite data.** *Hydrol Sci J.* 2020; **65**(14): 2508–2519.
Publisher Full Text

Stritih A, Bebi P, Grêt-Regamey A: **Quantifying uncertainties in earth observation-based ecosystem service assessments.** *Environ Model Softw.* 2019; **111**: 300–310.
Publisher Full Text

Svetnik V, Liaw A, Tong C, *et al.*: **Random forest: a classification and regression tool for compound classification and QSAR modeling.** *J Chem Inf Comput Sci.* 2003; **43**(6): 1947–1958.
Publisher Full Text

Taghizadeh-Mehrjardi R, Toomanian N, Khavaninzadeh AR, *et al.:* **Predicting and mapping of soil particle-size fractions with adaptive neuro-fuzzy inference and ant colony optimization in central Iran.** *Eur J Soil Sci.* 2016; **67**(6): 707–725.
Publisher Full Text

Tan Y, Robinson AL, Presto AA: **Quantifying uncertainties in pollutant mapping studies using the Monte Carlo method.** *Atmos Environ.* 2014; **99**: 333–340.
Publisher Full Text

Van Wesemael B, Chabrillat S, Sanz Dias A, *et al.*: **Remote Sensing for Soil Organic Carbon Mapping and Monitoring.** *Remote Sens.* 2023; **15**: 3464.
Publisher Full Text

Velázquez FJB: **"CONSOLE."** Dataset. OSF. 2022.
http://www.doi.org/10.17605/OSF.IO/B7S4Z

Venter ZS, Hawkins HJ, Cramer MD, *et al.*: **Mapping soil organic carbon stocks and trends with satellite-driven high resolution maps over South Africa.** *Sci Total Environ.* 2021; **771**: 145384.
PubMed Abstract | Publisher Full Text

Wadoux AMJC, Minasny B, McBratney AB: **Machine learning for digital soil mapping: Applications, challenges and suggested solutions.** *Earth-Sci Rev.* 2020; **210**: e103359.
Publisher Full Text

Wang X, Zhang Y, Atkinson PM, *et al.*: **Predicting soil organic carbon content in Spain by combining Landsat TM and ALOS PALSAR images.** *Int J Appl Earth Obs Geoinf.* 2020; **92**: e102182.
Publisher Full Text

Weng Q: **Quantifying uncertainty of digital elevation models derived from topographic maps.** In: Richardson, D.E., van Oosterom P., (Eds.). *Advances in Spatial Data Handling.* Springer, Berlin, Heidelberg, 2002; 403–418.
Publisher Full Text

Wickham H: **ggplot2: Elegant Graphics for Data Analysis**. Springer-Verlag New York, 2016.
Publisher Full Text

Wiesmeier M, Barthold F, Blank B, *et al.*: **Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem.** *Plant and Soil.* 2011; **340**: 7–24.
Publisher Full Text

Wuertz D, Setz T, Chalabi Y: **fNonlinear: Rmetrics - Nonlinear and Chaotic Time Series Modelling.** R package version 3042.792017.

Xiong X, Grunwald S, Myers DB, *et al.*: **Assessing uncertainty in soil organic carbon modeling across a highly heterogeneous landscape.** *Geoderma.* 2015; **251–252**: 105–116.
Publisher Full Text

Zeraatpisheh M, Jafari A, Bodaghabadi MB, *et al.*: **Conventional and digital soil mapping in Iran: Past, present, and future.** *Catena.* 2020; **188**: e104424.
Publisher Full Text

Zhou T, Geng Y, Chen J, *et al.*: **Mapping soil organic carbon content using multi-source remote sensing variables in the Heihe River Basin in China.** *Ecol Indic.* 2020; **114**: 106288.
Publisher Full Text

Zhou J, Li E, Wei H, *et al.*: **Random forests and cubist algorithms for predicting shear strengths of rockfill materials.** *Appl Sci.* 2019; **9**(8): 1621.
Publisher Full Text

Zhu M, Feng Q, Qin Y, *et al.*: **Soil organic carbon as functions of slope aspects and soil depths in a semiarid alpine region of Northwest China.** *Catena.* 2017; **152**: 94–102.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ✘ ? ? ✘

Version 2

Reviewer Report 03 May 2024

https://doi.org/10.21956/openreseurope.18038.r39750

© **2024 David Adeniyi O.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✘ **Odunayo David Adeniyi**
University of Pavia, Pavia, Italy

1. The R software version used should be cited.
2. In the Field sampling and soil analysis section, the total number of soil datasets used wasn't stated.
3. Still under the Field sampling and soil analysis section, "The information compiled (physical and chemical soil variables (SEISNET and ASAJA-Sevilla); site, mean annual precipitation, temperature and evapotranspiration (REDIAM)) were harmonised using R Software (identifying potential duplicates and errors) so that they had the same resolution and extent for analysing the relationship between SOC and other variables in-situ and ex-situ. A total of 22 commercial olive groves included in the data compiled from ASAJA-Sevilla were selected, taking into account their representativeness in terms of soil types and climatic conditions." Kindly rewrite this paragraph so that it can be well understood.
4. In Statistical analysis section, kindly rewrite this paragraph. Maybe you shouldn't start with "Since...". Or maybe the section is not even needed. Kindly delete.
5. What is the name of the DEM (where did you get it from and Why such product?)
6. The remote sensing data were gotten from July 2019, what is the percentage cloud cover condition on that day? Why this date? Does it coincide with the soil sampling date? Which day in July 2019?
7. The Climate variables was gotten from where? How did you downscale this climatic variables from 100-m to 30-m resolution?
8. ML techniques section: R programming (version 3.5.1), Good! kindly cite it.
9. " Compared with the cubist, the difference between lower and upper quartiles in the RF is larger. Furthermore, RF tends to perform better in different iterations (Zhou *et al*., 2019). Khaledian and Miller (2020) reported that the results taken by cubist are more interpretable than RFs' which are semi-interpretable. We therefore performed both models, then one of them was selected as a parsimonious model to be applied in spatial mapping." I don't think this is needed here.
10. How did you arrive at the hyperparameters you used for the RF and Cubist? did you perform hyperparameter tuning?
11. In "Acquiring spatial maps" section: "In addition to the rich capacity of modelling in R, it has a

user-friendly environment to map spatial data" There are other programming languages too that does same thing, such as Python, etc. I think this statement and others like this isn't needed. R is just a tool like Python and others...
12. Interesting, to see that ggplot2 package was used for spatial mapping not 'raster' or 'terra', etc. Are you sure of this? Is the code attached to this article?
13. The Results section. The Brief soil data didn't give us the explanatory/description of the soil data used for the modelling of the SOC.
14. This Terrain-related attributes section is not part of your result nor is it an objective of this article. It shouldn't be here. Maybe in the Materials and Method section. Same as the Remote-sensed data section.
15. In the Final ML technique section: You need to discuss why RF outperformed Cubist model.
16. 'The results generally revealed that Landsat -based indices are not better predictors', How did you come to this conclusion? Maybe it would have been better if you have compare Landsat -based indices + other ancillary variables versus Sentinel-based indices + other ancillary variables.
17. "In addition, the climate data were not as important as other supplied covariates", of course, they were originally in 100-m resolution. Such a coarse resolution might not fit to explain SOC variation at that high resolution. This doesn't mean climate couldn't be an important variable for SOC variability in this study area. Maybe you shouldn't have use a coarser climatic variable raster, or, instead, find a higher climatic raster.
19. You didn't describe the digital maps. Is it coherence with the Input variables? How can this maps be justify?

Is the work clearly and accurately presented and does it cite the current literature?
No

Is the study design appropriate and does the work have academic merit?
No

Are sufficient details of methods and analysis provided to allow replication by others?
No

If applicable, is the statistical analysis and its interpretation appropriate?
No

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Machine Learning, digital soil mapping, remote sensing

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 29 April 2024

https://doi.org/10.21956/openreseurope.18038.r39023

© **2024 Schillaci C.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Calogero Schillaci
JRC European Commission, Ispra, Italy

Dear Authors,

The manuscript "The possibility of spatial mapping of SOC content in olive groves under integrated production using easy-to-obtain ancillary data in a Mediterranean area" is relevant. However, some issues need to be addressed.
I appreciate that you used Open Research EU as the outcome of your research. I suggest to modify the title slightly, the manuscript is a case study, title seems long and not so appealing. Sure, you can apply spatial mapping; the point is, what is the outcome of the case study, and why should Olive Grove have been different from another land cover (e.g. Cropland)? Which aspect can be highlighted from this experience (e.g., particular precaution in the sampling design, difficulties in sampling close to the plant stem, root system making the task challenging, or influence of the soil sample)? All these things deserve the same importance: the algorithm used and the data treatment. Algorithm hyperparameter tuning is fundamental, you can check several good papers report the algorithms and their hyperparameter search for several reasons, to prove that the algorithm learn well and that it does not overfit (see suggested reads, Schillaci C,et.al.,2021 [Ref 1] ,Veronesi F,et.al.,2019 [Ref 2] , Schillaci C,et.al.,2017 [Ref 3], Schillaci C,et.al., 2018 [Ref 4]). Please try to add all these missing parts. The decision to use the cubist and RF should be motivated (e.g. mostly used in SOC mapping, but to state this, you need to perform a literature analysis (see suggested read Schillaci C,et.al., 2018 [Ref 4]) and collect evidence about the algorithm used in similar contexts, Mediterranean, olive grove or fruit trees...at least in croplands).
The statement about the RS indices that are more relevant than other predictors is not so new (See suggested read Schillaci C,et.al.,2017 [Ref 3]).
Please try to update this parts and provide the limitations of the proposed methodologies as well as future perspectives and lesson learned from this experiment.
I provided the reference as suggested reads, no need to add this specific references, but look at them and references therein.
Kind regards

References

1. Schillaci C, Perego A, Valkama E, Märker M, et al.: New pedotransfer approaches to predict soil bulk density using WoSIS soil data and environmental covariates in Mediterranean agro-ecosystems.*Sci Total Environ*. 2021; **780**: 146609 PubMed Abstract | Publisher Full Text
2. Veronesi F, Schillaci C: Comparison between geostatistical and machine learning models as predictors of topsoil organic carbon with a focus on local uncertainty estimation. *Ecological Indicators*. 2019; **101**: 1032-1044 Publisher Full Text
3. Schillaci C, Acutis M, Lombardo L, Lipani A, et al.: Spatio-temporal topsoil organic carbon

mapping of a semi-arid Mediterranean region: The role of land use, soil texture, topographic indices and the influence of remote sensing data to modelling.*Sci Total Environ*. 2017; **601-602**: 821-832 PubMed Abstract | Publisher Full Text
4. Schillaci C, Saia S, Acutis M: Modelling of Soil Organic Carbon in the Mediterranean area: a systematic map. *Rendiconti Online della Società Geologica Italiana*. 2018; **46**: 161-166 Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and does the work have academic merit?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Agricultural and Environmental Sciences

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Version 1

Reviewer Report 03 May 2024

https://doi.org/10.21956/openreseurope.15895.r32365

© 2024 Dash P. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Prava Kiran Dash
1

Department of Soil Science, College of Agriculture, Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India
[2] Department of Soil Science, College of Agriculture, Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

This article is a good attempt to make SOC maps for the said study area. Overall a good attempt, but needs improvements. Although not too novel in terms of methods, but still this kind of regional paper also needs to be published to encourage mapping attempts at a regional scale and also for the agricultural management of the concerned and similar areas. However, it needs some rectifications, deletions, and additions before indexing. I have observed a number of mistakes. I am mentioning my detailed observations below. The authors are requested to improve the manuscript before final indexing.

Best wishes to the authors.

Comments:

Title:

- The title of the article can be modified: “the possibility” can be deleted.
- This study has used legacy soil data, so that should be mentioned in the title as well. (for example, spatial mapping of soil organic carbon content at three depths using legacy soil data and easy-to-obtain ancillary data in a Mediterranean area).
- Exact place name should be mentioned instead of Mediterranean area (for example, Andalucía in southern Spain).

Abstract:

- Unlike most of Europe, Andalucía in southern Spain as a Mediterranean area still lacks digital maps of soil organic carbon (SOC)--> Mediterranean area that still lacks
- Operational Land Imager ‘OLI’ and… -> Operational Land Imager (OLI).
- Multispectral Instrument ‘MSI’… -> Multispectral Instrument (MSI).
- goodness-of-fit criteria: please mention the exact parameters.
- “We obtained evidence that the resolution of satellite images is a key parameter in modelling and digital mapping” - was this the objective of the study to determine the importance of resolution? If yes, mention it in the title and elsewhere and frame the research question accordingly. Otherwise, I just suggest restructuring the conclusion part of the abstract.
- Based on the objectives mentioned in the last paragraph of introduction, the conclusion part should just answer those questions: 1) at what accuracy the maps were prepared. 2) what was the best algorithm/model 3) ranking of the predictors/covariates.

Introduction:

- The first paragraph of the introduction looks unnecessary.
- 2nd paragraph: "DEM-derived data can save considerable labour and costs…" - explain how?
- "The spatial patterns of SOC were linked topography in a case … "-> linked to topography.
- "Venter et al. (2021) relates elevation …" -> related.
- "There are other research works with Landsat-8 OLI, Sentinel-2A and EnMap spectral bands to estimate soil organic matter (Rosero-Vlasova et al, 2019) in Brazil" - it says research works, which means multiple research should have been cited. Please reframe the

sentence.
- "agriculturally important nutrients (Shahbazi et al., 2019a) and the soil ripening process in Iran (Mousavi et al., 2020); and available water content in Australia (Gooley et al., 2014)..." - keep only one 'and'.
- Introduction and elsewhere: Abstract says both Seville and Cordoba provenience. But, in the introduction section, only Seville is mentioned. Please clarify/rectify.
- Last paragraph of introduction: "(Landsat-8 OLI, Sentinel-2A MSI) as well as integration and terrain-related attributes (DEM-derived data)..." -> (Landsat-8 OLI, Sentinel-2A MSI), as well as integration and terrain-related attributes (DEM-derived data)...
- "The information compiled (physical and chemical soil variables (SEISNET and ASAJA-Sevilla); site, mean annual precipitation, temperature and evapotranspiration (REDIAM)) were harmonized using R Software (identifying potential duplicates and errors) so that they had the same resolution and extent for analysing..." - Not clear, please reframe the sentences and use the punctuation marks properly for better understanding.

Materials and methods:

- Figure 2 is unnecessarily complicated. The second line of contents (C) are not at all required. From (A) and (B) remove the example of indices as they are mentioned in details in the tables. In D, shorten the sentences. No need to write full sentences. Remember this is just a flow diagram. The details mentioned in the paragraph format are enough for understanding. Simplify the arrows in Figure 2. It should look simple and less complicated.
- The softwares used for deriving the covariates can be added in an additional column of table 1.
- "Landsat-8 OLI and Sentinel-2A MSI images acquired in July 2019 (see Underlying data) were selected for further analysis in this project due to the high probability of no clouds in the images." Was this the reason for selecting these satellite images? Or, is there any scientific reason?
- "Fortunately, the selected scenes comprise minimal cloud coverage and maximum soil surface exposure." The correct way to write this is...XYZ satellite imageries were selected based on minimum cloud coverage percentage.
- "To explain the vegetation, soil and water, landscape as well as geology of the study area, six band ratios were calculated using both images as well as their integrations for each index. The hypothesis of integrating the indices has been raised from the report of Wang et al. (2020) which demonstrated that combining Landsat TM and ALOS PALSAR images could predict more accurate SOC content, especially for soils with high vegetation canopy density in Spain." What is the need of mentioning this sentence in between methodology. Please delete.
- "To predict the spatial distribution of SOC, it was found that despite topographic variables which had higher influence at finer scales, the climatic variables were more important at coarser scales e.g., sub-regional to global scale (Adhikari et al., 2020). Among the previously described indices, climate variables contributed the most to explaining SOC content variation (Zhou et al., 2020)." No need to repeat the sentences with the same meaning repeatedly. Please delete this and other unnecessary sentences.
- "machine learning (ML)": use the abbreviation only once at the beginning of the paper. No need to mention it again and again.
- "Regarding the accelerated adoption of machine learning (ML) techniques in prediction of soil properties, especially in the last 10 years (Padarian et al., 2020)": Unnecessary, please delete.
- "...(composed by 20% of data no included in the modelling process)..." -> composed of; data

not included.

- "e.g., concordance (CCC)...": Write the full form of CCC.
- "It is clear that R^2 measures the precision of the relationship (between observed and predicted) while concordance as a single statistic evaluates the accuracy and precision of the relationship." Delete 'it is clear that'; single statistical parameter that evaluates; add a reference to the sentence.
- "Similar work was fully explained by Malone et al. (2017) using the "goof" function within the "ithir" package in R (Malone, 2016)." - similar work or the details regarding the parameters? Please clarify.
- "In addition to the rich capacity of modelling in R, it has a user-friendly environment to map spatial data (Malone et al., 2017). Bivand et al. (2013) documented some methods for handling spatial data and analysis in R in a detailed manner." - Not required (may delete).
- "same projection in advance to be easily import..." -> to easily import.
- "Arrouays et al. (2020) reported that uncertainty maps may give very strong arguments in DSM because the identifying missing soil data and excessive field work will be distinguished." - Not clear, please rectify or delete.

Results:

- alluvial episodes?
- Etc -> etc.
- "In fact, the impact of Integrated Production was analysed comparing SOC data previous Integrated Production implementation (2005) and current SOC data. The results showed the increase in SOC due to the implementation of Integrated Production in terms of crop and soil management. (Figure 3)." What is the purpose of giving this information and this figure? It is not the research question. Please stick to the objectives and don't put information unnecessarily.
- The sub sections Terrain-related attributes and Remotely-sensed data in the results section is irrelevant. It could have been placed in the methods section, if it was supposed to be kept in the paper. Please move these contents to methods, else simply delete. This is not the results of this paper based on the objectives set (as mentioned in the introduction).
- "Final machine learning technique" - Please change this sub-section title.
- "Machine learning allowed us to analysis the huge dataset and their relevance to SOC at multiple depths. For that, the use of open software like as R facilitated modeling and then mapping." - Not needed, please delete (already discussed in methods).
- Where is the actual results paragraph? Please explain table 3 in details. This is the actual result of this study. At what accuracy the maps could be prepared? Discuss in detail CCC, RMSE for each depth, each model.
- "The results generally revealed..." -> Results revealed that...
- "Landsat -based indices" - delete the extra space.
- "The integrated indices e.g., NBR_ITG, NDWI_ITG and CMR_ITG were identified as the most important covariates in prediction of SOC0–25cm, SOC25–50cm and SOC50–75cm, respectively." - Figure 6 says something different. Please explain the 3 best predictors for each depth as presented in Fig. 6. Modify the paragraph. Read and match with the figure carefully.
- "In general, the high values of uncertainties were found in the south of the study area." - What could be a possible reason?

Conclusion:

- Limitations can be improved? Is only lack of data availability the only limitation? Try to minimize the limitations in the conclusion section. They can be added in the discussion part after improvisation.
- "The analysis of SOC on different soil types allows enhancing of results. The terrain attributes analysis was a key step in order to obtain SOC data in middle depth. Taken together, our findings indicate the successful usage of machine learning with application of user-friendly software e.g., R programming." - Not necessary, please delete.
- "The use of satellites for monitoring SOC..." -> The use of satellite imageries/satellite derived information.
- "In addition, several agricultural policies and/or contract solutions such as new CAP (Common Agricultural Policies), Integrated production, Ecologic production and so on, may be beneficial for this methodology in order to reduce costs in hot-spot checks. Furthermore, the technology proposed may support the development of new contract solutions based on results within a payment per carbon sequestered." - delete, not necessary. This is not the research question.
- In conclusion the authors should simply say XYZ was the research question and XYZ is the answer. In this paper, if the authors can simply say we found...XYZ model performed better, XYZ was the CCC, XYZ was the RMSE, this was the best predictor (for each depth), that would serve the purpose.

All the best.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and does the work have academic merit?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Digital soil mapping

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 13 June 2023

https://doi.org/10.21956/openreseurope.15895.r32370

© **2023 Kaya F.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Fuat Kaya
[1] Isparta University of Applied Sciences, Isparta, Turkey
[2] Isparta University of Applied Sciences, Isparta, Turkey

I have carefully examined the mentioned article that aims to efficiently model the soil organic carbon content across three different soil depths within a watershed using the DSM (Digital Soil Mapping) framework.
The study does appear to be built on an unsound scientific foundation. Moreover, there are significant shortcomings in the title, abstract, and overall writing of the article.
As can be understood, the manuscript compared the performance of machine learning (ML) algorithms with 2 different mathematical foundations (RF and Cubist). However, what were the specific reasons for choosing these two algorithms among the numerous alternative MLs available? While a comprehensive comparison of all algorithms may prove impractical, the rationale behind our particular choices must be explained.

First of all, the aims of this study are not well defined to fill the gap in the research question and literature. There appears to be an enormous lack of hypotheses in the study. What did you assume? As you emphasized in the conclusion of the abstract, does the spatial resolution of satellite images make a significant difference in the estimation of soil SOC content at 3 depths? Such a hypothesis was studied in different regions and its results were presented. For example;
--Do more detailed environmental covariates deliver more accurate soil maps? (Samuel-Rosa *et al.*, 2015[1]).
--Effects of different sources and spatial resolutions of environmental covariates on predicting soil organic carbon using machine learning in a semi-arid region of Iran (Garosi *et al.*, 2022[2]).
It will be very difficult to develop a research question or hypothesis that can advance scientific knowledge by bypassing this literature. What I understand from the article is an approach such as averaging two different satellite-based indices. An assumption about this would be innovative. However, it is necessary to emphasize both the spectral and spatial resolutions of the relevant sensors (MSI and OLI).

Based on my thorough examination of the article, I will provide detailed explanations below. The authors will have put in considerable effort to salvage the article from being rejected. Consequently, I recommend reading this article to gain insights into potential areas of focus within DSM that could contribute to advancing scientific knowledge.
--Ten challenges for the future of pedometrics (Wadoux *et al.*, 2021[3]).

Furthermore, I suggest restructuring the article by formulating research questions that could lead to innovative solutions for addressing various challenges. As it stands, the current form of the article fails to offer anything beyond the application of existing DSM techniques commonly

employed in the Mediterranean region.

Title: Consider replacing the title: "possibility" with something else.

Abstract: Your manuscripts focus on regression problems, which necessitates structuring the abstract to incorporate quantitative model evaluation measures such as model performance criteria (e.g., "RMSE," "MAE," or "R2") used in your study. It is important to acknowledge that these predictive maps inherently carry a degree of error.
<<<< 'MSI' <<<<<<<<<<< (MSI)
<<<< 'OLI <<<<<<<<<<< (OLI)
use parentheses to show abbreviations.

Introduction:
The introduction is very technically written. DSM literature in these areas is overlooked, as is the SOC cycle in the Mediterranean and the challenges in SOC management.

Material-Methods:
"Integrated production includes several rules (Romero-Gámez et al., 2017) for crop, soil, and water management to increase the sustainable production and other agri-environmental public goods, such as carbon sequestration, and promoting the reduction of fertilizers and pesticides".
The sentences above do not seem appropriate to the section they are in.

Field sampling and soil analysis:
What is the exact date of taken of the samples in 22 commercial olive orchards?

Ancillary data:

- It is important to include multiple digital environmental variables that may be associated with soil formation to explain the variation in the area. However, raw grid resolutions of each must be provided. Next, you need to specify that you are aggregating or disaggregating the spatial modeling framework. For example, when the raw resolutions of your variables vary between 10 and 100 meters, at what meter resolution would you put them in the same frame?

- Provide a proper reference to the R program (citation" ").

- Explanatory and transparent information about the hyperparameter optimization of either algorithm is not presented in manuscript as a supplement.

- I downloaded up to 3GB of your data. but I couldn't find any reproducible code. There are only "Rasters" and "Field_data". Turns out that you're not actually sharing reproducible results.

Results and discussion:

- When I come to the findings of the study, I am completely disappointed. Using very old data, how can we generate a digital map of SOC using RS data from 2019? I haven't seen this approach much in DSM. Of course, similar ones may have been published, but this is completely incomplete. Use the index values you get from the average band values of the satellite images from 1990-2000 (of course, no Sentinel 2 here) and the satellite images

from 2000-2010 because you have a different data set. Otherwise, what you're doing is not guessing, it's giving the model a number and getting a number.

- Terrain-related attributes and Remotely-sensed data sections are placed to fill the page. There's no need for them.

- What is IncMSE in Figure 6? Don't leave everything to the reader. You need to provide information about this in the material method.

- What was the ability of the 22 commercial olive groves to represent the entire area? There is tremendous noise on forecast maps. How well do the environmental variables in the dataset match the whole study area?

- Is your work flawless, or are there any limitations in this study? Additionally, I would appreciate hearing your insights and recommendations for future research. It might be helpful to address these points as a separate paragraph within Section 3 of your work.

References

1. Samuel-Rosa A, Heuvelink G, Vasques G, Anjos L: Do more detailed environmental covariates deliver more accurate soil maps?. *Geoderma*. 2015; **243-244**: 214-227 Publisher Full Text
2. Garosi Y, Ayoubi S, Nussbaum M, Sheklabadi M: Effects of different sources and spatial resolutions of environmental covariates on predicting soil organic carbon using machine learning in a semi-arid region of Iran. *Geoderma Regional*. 2022; **29**. Publisher Full Text
3. Wadoux A, Heuvelink G, Lark R, Lagacherie P, et al.: Ten challenges for the future of pedometrics. *Geoderma*. 2021; **401**. Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

No

Is the study design appropriate and does the work have academic merit?

No

Are sufficient details of methods and analysis provided to allow replication by others?
No

If applicable, is the statistical analysis and its interpretation appropriate?
No

Are all the source data underlying the results available to ensure full reproducibility?
No

Are the conclusions drawn adequately supported by the results?
No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Digitai soil mapping, machine learning, remote sensing

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 05 October 2022

https://doi.org/10.21956/openreseurope.15895.r30142

© 2022 Abd-Elmabod S. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Sameh Kotb Abd-Elmabod
[1] Soils & Water Use Department, Agricultural and Biological Research Division, National Research Centre, Cairo, Egypt
[2] Soils & Water Use Department, Agricultural and Biological Research Division, National Research Centre, Cairo, Egypt

This work is good and interesting because it is based on using different datasets (soil database and Landsat-8 and Sentinel-2A images), and to run two models (random forest and Cubist) for the prediction of SOC at three soil depths. However, there are a few comments that are inserted below:

- The study area is located in Seville and Cordoba but you referred various times in the different manuscript sections only to prediction of SOC in southern Spain (Seville), so please go through the text and add the Sevilla and Cordoba provinces.
- In figure 1, the subfigure A should be improved and add the coordination, and subfigure 1D for which temperature, so please add mean temperature.
- In the flow chart you can use different colours for each step.
- In table 1, in the description of Normalized Difference Water Index, please put water instead

of vegetation, and in the description of Clay Minerals Ratio and Normalized Difference Salinity Index, please put Soil instead of geology.

- In the section Terrain-related attributes, case study areas - change by study area.
- In my opinion you should separate the limitations of this study from the conclusion section. Maybe you can add a new section for the study limitations after the results and discussion. Also, one of the limitations is that the date of satellite images were recent (2019) but the dates of soil sampling were 1990–2019, why did you not use only the recent soil data (2019), as the soil organic carbon can be changed through time.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and does the work have academic merit?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Land evaluation, soil carbon, climate change

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.