UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE AGRARIE E ALIMENTARI

Scuola di dottorato Agricoltura Ambiente e Bioenergia

Mapping Soil Organic Carbon dynamics over the last decades in Mediterranean agro-ecosystems with legacy data

Ph.D. Thesis

Calogero Schillaci

N° R11303

Supervisor	**Academic Year**	*Coordinator*
Prof. Marco Acutis	**2017-2018**	**Prof. Daniele Bassi**
Co-supervisor		
Dr. Sergio Saia		

Title: Mapping Soil Organic Carbon dynamics over the last decades in Mediterranean agro-ecosystems with legacy data

Ph.D. in Agriculture, Environment and Bioenergy - XXXI Cycle

Ph.D candidate: Calogero Schillaci

Supervisor Prof. Marco Acutis

Co-Supervisor Dr. Sergio Saia

Schillaci C., 2018. Mapping Soil Organic Carbon dynamics in Mediterranean agro-ecosystems. Ph.D. Thesis, University of Milano, 172 pp.

Summary

Soil organic carbon (SOC) represents the biggest carbon pool of the biosphere, bigger than the living plant pool. In agriculture, SOC is of pivotal importance for sustainable soil management and is a main soil fertility indicator. As soils are responsible for food production and the provision of various ecosystem services, there is a sturdy interest in understanding how land use and management affect natural plant and crop growth, and ecosystem resilience and functioning. These processes require time and soil sustainability is to be evaluated in a long-term economic perspective by policy makers with the aim of maintaining adequate, and likely improved, conditions of the soil and the whole farm for the future. Thus, long-term actions for crop sustainability could also admit little short-time yield reduction if yield potential, stability and environmental health are maintained at the long-time.

Food production and ecosystem services provision depend on the maintenance, or increase, of SOC in agricultural soil, since SOC act as a short-term nutrient reservoir, increase water holding capacity and soil infiltration rate, reduce soil compaction, and favour soil resilience against pollutants. These effects should be taken into account at both a narrow and broad geographical breadth.

When aiming to manage SOC at broad geographical extent, a detailed knowledge of SOC distribution and likely change in time is required. However, such a knowledge relies on correct sampling method and modelling procedures that in turn depend on the environmental variability of the area under study. Mediterranean areas are frequently variable as an harbour, the area has been subjected to a high share of soil and above-ground biodiversity and experienced long cultivation history and intensification since the last century, which increased their fragility. In this environment, the acquisition of reliable information on SOC can require a highly dense sampling, which can also negatively affect some relict environment. In addition, sampling can imply a high cost for field work and laboratory analyses.

The aim of my Ph.D. work was thus to investigate the main factors related to SOC spatial distribution in agricultural land under various pedoclimatic conditions in semiarid Mediterranean areas, using a

legacy soil database (1968-2008) of SOC and soil bulk density. The dissertation is structured in six chapters: the first one is a general introduction where the rationale of the dissertation is explained, and the research questions are stated. The second chapter is a novel approach to systematically collecting literature from international peer-review issues, namely systematic map. The third one is an analysis of the legacy soil database, which intends to make the database ready to be used for the SOC assessment and for the digital soil mapping. The fourth chapter touches an issue dealing with SOC stock mapping with the boosted regression tree and a set of covariates to produce local SOC benchmarks to be compared with European and Global SOC maps. The fifth chapter fits in the same modelling frame and it is addressed at the SOC dynamics using the most widespread legacy sampling campaign. A high number of available spatial data were collected and computed and used to calibrate the SOC models. At this stage, due to the ungridded structure of the data, a machine learning based model has been used (Boosted Regression Trees). The last chapter is a comparison of models (geostatistical, machine learning and linear), and shows useful information about the way that the error is reported by each algorithm. Soil maps are not just produced for the sake of creating attractive geographical visualizations: they have a very precise task to fulfil, i.e. provide accurate and reliable information on soil properties that decision makers can use to plan interventions of any kind.

The use of the Regression Kriging and Boosted Regression Trees models, which resulted in the best prediction performance in terms of R2 and RMSE, highlighted the SOC dependence on environmental factors, and the prediction of the agricultural land covers. All land cover groups were studied in the preliminary stage of this study (chapter 2), while only the cropland identified with the legacy data was the candidate for the development of the final models which lead to the detection of a positive SOC trend. The last chapter aimed at the comparison between geostatistical, machine learning and linear models to predict SOC in agricultural lands, and an improvement in local uncertainty estimation. The outstanding result was that SOC at the monitoring sites were accurately simulated, being in full agreement with observed data. Once more, actual data will be available and the model will be calibrated and validated, a model of SOC potential sequestration regional scale can be produced. The results of this dissertation has led to a clear and shared vision in the community regarding the selection of the estimation methods for SOC prediction needs to be based on careful considerations. It is good practice to test algorithms already used in literature for similar purposes, but it may be counterproductive to only look at an algorithm because it is new and never used before in a particular field. This sometimes happens in science where methods are selected only because fashionable and not based on real and tested experiments. In the dissertation the origin of the data was sometimes know and sometimes it has been data driven based. In particular, sampling design was based on geostatistics only in the 2008 campaign and it may well be that looking at very advanced

methods like deep-learning could be interesting, but still less accurate than the geostatistical kriging based algorithms, which can also provide robust and well tested uncertainty estimations. In summary, even though we have now access to advanced algorithms it does not mean that we need to use them blindly without fully considering what we are trying to achieve with our working hypothesis and research question.

After the last chapter, a brief paragraph outlines the general conclusions and recommendations.

Foreword

The dissertation aimed at the optimization of an underused legacy database, which can be integrated with other soil database (e.g. at European level), and at the development of a monitoring and mapping solutions for legacy soil data-based analysis for the agro-ecosystem of a semi-arid Mediterranean island (Sicily). With this work, a bunch of regional data takes shape along with the ability to generate information derived mainly from historical data. Processes of these kind are hard to discover and mostly studied, for some environment, only through a wide use of intercontinental benchmarks. These make possible to achieve satisfactory results when evaluating broad changes in wide regions, but are scarcely applicable to downscaling processes for the evaluation in small regions.

The PhD has been driving me through the academic and publishing pipelines, that include ethics. In the papers presented in this dissertation, I have conducted and supervised the very most of the research steps, from data collection, to analysis, to presentation and interpretation of the results. However, I am in debt to some of my colleagues, which made an insightful intellectual contribution to the data analysis and interpretation: firstly with my supervisor Prof. Marco Acutis and my co-supervisor Dr. Sergio Saia.

Credits evaluation

Courses:

- ESA EO Summer School ESRIN, Frascati - Italy, 30 July - 10 August 2018.
- Hands-on Global Soil Information Facilities (GSIF), Wageningen–Nederlands 12-19 June 2017
- Statistics Applied To Environmental Engineering (Milano) March 2017
- Metodologia statistica per le scienze agrarie, i modelli lineari generali e generalizzati, Perugia June 2016.

Papers published in International journals

1. **Schillaci C.**, Acutis M., Vesely F., Saia S., 2018. A simple pipeline for the optimization of legacy soil datasets: an example and test with soil organic carbon from a highly variable area. CATENA doi.org/10.1016/j.catena2018.12.015.
2. **Schillaci** C., Saia S., Acutis M. 2018. Modelling of Soil Organic Carbon in the Mediterranean area: a systematic map. Vol. 46, pp. 161-166, doi.org/10.3301/ROL.2018.68.
3. Lombardo, L., Saia, S., **Schillaci**, C., Mai, P., Huser, R., 2018. Modeling soil organic carbon with Quantile Regression: Dissecting predictors' effects on carbon stocks. Geoderma. Geoderma 318, in press. doi:10.1016/j.geoderma.2017.12.011
4. **Schillaci**, C., Acutis M., Lombardo L., Lipani A., L., Fantappiè M., Märker M., Saia S., 2017. 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. Science of the Total Environment 10.1016/j.scitotenv.2017.05.239
5. **Schillaci** C., Lombardo L., Saia, S., Fantappiè M., Märker M., Acutis M., 2017. Modelling the topsoil carbon stock of agricultural lands with the Stochastic Gradient Treeboost in a semi-arid Mediterranean region. Geoderma 286, 35–45. doi:10.1016/j.geoderma.2016.10.019

Submitted

6. Veronesi F., **Schillaci C.**, Comparison between geostatistical and machine learning models to predict topsoil organic carbon with a focus to local uncertainty estimation. Submitted to ECOLOGICAL INDICATORS

Oral presentation at international/national congress:

1. **Schillaci C**., Saia S., Perego A. and Acutis M.- Estimating Soil Organic Carbon Of Arable Lands with Regional Legacy Soil Data And The 'Land Use And Coverage Area Frame Survey (LUCAS)', In Contrasting Areas Of Italy. Società Italiana di Agronomia XLVII Convegno Nazionale 12-14 September 2018.

2. **Schillaci C**. and Acutis M. 2017. Soil Organic Carbon mapping with Boosted Regression Trees. XII GIT 2017 Geosciences and Information Technologies, Gavorrano 25-27 June 2017.
3. **Schillaci C**., Lombardo L., Saia S., Fantappiè M., Märker M. and Acutis M.- Improving Soil Organic Carbon stock estimates in agricultural topsoil at a regional scale using a Stochastic Gradient Boosting technique- EGU General Assembly 2016.

Poster presentation at international/national congress:

1. **Schillaci C**., Perego A., Saia S., Bellieni M., Brenna S. and Acutis M.-Using regional pedological map to link the European soil organic carbon data from LUCAS with legacy data- EGU General Assembly 2018.
2. **Schillaci C**., Fastellini G., Diaz M., Iacono S. and Acutis M. - Optimizing nitrogen fertilization in corn fields using CropSpec proximal sensing sensor. 11° European conference on Precision Agriculture Edinburgh 15-17 July 2017.
3. Giovino A., **Schillaci C**., Saia S., Currò V., Vicari D., Reale S. and Caracappa S.- Protecting the habitat of the Mediterranean loggerhead sea turtle Caretta caretta by enhancing the ecosystem services provided by the European fan palm Chamaerops humilis. European Network of Palm Scientists (EUNOPS) 2017; 05/2017
4. **Schillaci C**., Acutis M., Lombardo L., Lipani A., Fantappiè M., Märker M. and Saia S. Mapping the variation of soil organic carbon (SOC) stock in time and space in Sicily, an extremely variable semi-arid Mediterranean region, highlighted that C was lost in area rich in organic C and gained in poor-C areas. EGU General Assembly 2017.
5. **Schillaci C.**, Saia S, Braun A. and Acutis M. Mapping of topsoil organic carbon in agro-ecosystems of a flat terrain area (Lombardy) by means of legacy soil data, climatic data and NDVI time series predictors with machine learning methods. EGU General Assembly 2017.

Summary

1. Introduction

1.1 Background and paradigms of soil carbon mapping in agricultural soils

Soil organic carbon (SOC) is an ecological indicator of soil health and ability to provide ecosystem services and soil fertility. Its study is mandatory for decisions making on sustainability (Dai et al., 2014; Panagos et al., 2013b). From a subsistence to a commercial level, agriculture soils have to provide crops, fibres and livestock (Bocchi, 2015). Soil data are the base for quantitative soil science. In this regards, there is a debate running on the use of existing and future data for non-experts by enhancing data access and providing easy-to-understand documentation, such as maps and supporting materials (Campbell et al., 2017). The scientific production about SOC in the last decade has been quantified in around 35 thousand papers (Smith et al., 2018). However, there is still a gap in the practical use by non-expert stakeholders and to find effective ways to share knowledge with soil managers and policy makers so that best management can be implemented (Lobry de Bruyn et al., 2017; Smith et al., 2015).

Maps conceal with a power of persuasion (Boria, 2017) by better reaching the aim of visual display. Nowadays the possibility to process and analyse big data allows for the rediscovery of legacy information stored in tables and in maps. At global scale, there is a lack of common procedures on protocols of soil surveying and lab procedures (Jandl et al., 2014), and difficulties in quantifying SOC pool changes over time (Köchy et al., 2015b, 2015a). This was underlined by a paper signed by 73 authors, in which an urgent need to reduce uncertainty associated with SOC management across terrestrial ecosystems is stated (Milne et al., 2015).

Many Earth System Models recognized that climatic variables are of primary importance in empirical and modelling approaches of SOC (Davidson and Janssens, 2006). Recently (Doetterl et al., 2015a) indicated that SOC turnover is likely to be related to geochemical factors. Based on data from field experiments across North America, Europe and Asia, it was recently demonstrated that global warming have induced changes in soil carbon stocks (Crowther et al., 2016). Also, these changes are likely to intensively impact soils in Mediterranean areas, which are fragile, and thus also reduce biodiversity in Mediterranean ecosystems and agro-ecosystems (Sokos et al., 2013; Underwood et al., 2009; Zamora et al., 2007).

Several studies were carried out to characterize soil diversity of the Mediterranean region, especially in cropland and knowing its status of conservation is a pre-requisite to improve land management. In particular, the knowledge of soil status and its modelling can allow for the problem identification of the soil organic pool and thus establish tools to reduce the pressure on agro-ecosystems while preserving its functionality (Brilli et al., 2017; Hijbeek et al., 2017). SOC mapping is one of the most

important tasks if considering that the Mediterranean region suffer by various agro-environmental problems, including climate change; land reclamation for industry and buildings; and high intensity of intensive crop production. These issues dramatically increase land degradation (Rodeghiero et al., 2011). Recently a worldwide initiative has been launched at the COP21 to offset CO_2 emission by increasing the global SOC stock by 4 per 1000 (or 0.4 %) per year (Minasny et al., 2017). This based the findings on the SOC stock estimates and sequestration potentials from 20 regions in the world and under best management practices. European Community has worked for the monitor and the adoption of new strategies of subsidies for farmers to slow the land degradation processes (Borrelli et al., 2016; Lugato et al., 2016).

1.2 Soil Organic Carbon in the digital soil mapping frame

To classify a soil and quantitatively measure their property fits in the general framework of soil mapping and pedometrics (McBratney et al., 2003; Vaysse et al., 2017; Xiong et al., 2015). Soil mapping as a branch of soil science has been developed mainly for the assessment of soil resource using geographic information systems (GIS), in order to limit the labour intensive work of soil analysis and field surveys. The temporal dimension is also important to understand the reasons of the SOC dynamics in space. Indeed, historical land assessment can give information on critical soil management aspects that strongly affect soil conditions (e.g. land use, tillage and fertilisation practices, plant residue management, crop rotation, flooding, low SOC values, change in topographic variables, construction of dams, etc.), and that should be taken into account for maintaining soil quality (Gregorich et al., 1994).

There are currently amount of mid-term (around 3-5 years) rather than long term (more than 10 years) experiments that have let to accumulate time or spatial clustered data. These data can be effectively treated with the statistical approaches. The discipline that offers different tool to understand such a variation in a quantitative way is Digital soil mapping (McBratney et al. 2003; Behrens & Scholter 2006a). DSM has multiple aims, including the provision of accurate estimates along with its accuracy regardless of the scale.

Understanding soil variation across the landscape and especially for agriculture management is of foremost importance in the actual changing climate conditions. For examples, topographical variables plays an important role in SOC prediction in heterogeneous landscapes (Behrens et al., 2010; Grimm et al., 2008; Mondal et al., 2016; Odeh et al., 1994).

After many years that the discipline was brought to an operational level, two reviews appeared important for its uniform use by researchers (McBratney et al., 2003; Scull et al., 2003). Scull et al. (2003), referred to predictive soil mapping and provided an overview of intents of the DSM to i) use

other ancillary variables in the process of estimating the property of interest and to collect soil data more effectively, ii) to produce a better representation of soil as a continuous, landscape variable and iii) to incorporate expert knowledge into predictive modelling. McBratney et al (2003) gave the theoretical framework of the discipline and their modern aspects (Malone et al., 2009; Minasny et al., 2013). The bunch of techniques DSM offers can drive knowledge in pedogenesis at local and regional scales. DSM needs as input digital data to represent the feature space for the quantitative models (Behrens et al., 2010). DSM seems to be based on the Jenny's well-known equation CLORPT (1941) which identified 5 major factors in the soil formation, subsequently modified in the actual SCORPAN equation where C is the Climate, O represents the Organisms, R the Relief, P the Parent material, A the Age, time, N the Geographic position. The SCORPAN model is a top-down data driven approach. This model was proved to be formalized firstly by Dokuchaev between 1899 and subsequently treated by Sergey Zakharov in its well-known fundamental textbook published in 1927 (Florinsky, 2012). The latter author argue that Jenny (1941) adopted the soil formation equation learning by the work of Sergey Zakharov that represent the work of Dokuchaev itself. This was the result of a recent investigation published in the anniversary year for publications of Dokuchaev and Jenny (Florinsky, 2012). DSM currently rely on finding geographical areas where soil properties are relatively constant. By exploring environmental covariates with multivariate statistical analysis and Machine Learning is the new frontier of soil science and pedology. Spatial distribution across multiple spatial scales is then searched in the interactions between soil forming factors, biophysical processes responsible for soil development and various other environmental proxies. Spatial scale in DSM could be summarized in High <20 m, Medium 20-200 m and Low > 200 m (McBratney et al., 2000). Temporal scale are better documented at detailed scale via intensive multiple year georeferenced sampling but are lacking to document the evolution of soil properties and processes at regional and continental scale (Grunwald et al., 2011).

World soil data have been first organized for a SOC and N map by Batjes (1996), followed by a global assessment (Hiederer and Köchy, 2012; Köchy et al., 2015a, 2015b), and subsequently mapped with a biome approach by Batjes (2014). Those studies are based on incorporated data points over the world, however there are biases in their spatial distribution and their density much different between the developed countries and the growing economies and rural ones. Latest worldwide SOC models has taken into account the capabilities of different observational databases using a data mining (Boosted regression trees) for the SOC mapping (Hashimoto et al., 2016). In the aforementioned paper, the relative importance of five groups of predictor (topographical, climate, soil properties, vegetation and land cover) were leading in the SOC distribution in Earth System Models ESMs.

At European level croplands plays an important role in acting as a potential carbon sink because of their area, their biological potential for carbon storage is of the order of 90–120 Mt C (Freibauer et al., 2004; Smith, 2004), by the fact that there is up to 45 Mt C per year of raw materials available. Furthermore, much effort has been made by the Joint research centre (JRC) of the European commission. European scale have benefited by ad hoc European sampling networks (Lugato et al., 2014a; Orgiazzi et al., 2017; Panagos et al., 2013a).

1.3 An introductory note on semi-arid Mediterranean agroecosystems and the study site

Drylands cover nearly of the half of the world and are inhabited by cca. 40 % of the world's population. Such lands, mostly occurring in undeveloped and developing countries, harbour a variety of soils whose net primary and agricultural production is limited by water scarcity and high temperatures in the area, low water holding capacity (WHC) and fertility of the soil and other soil-specific traits, including potential and actual soil erosion. In such conditions, the preservation of the soil organic carbon (SOC) pool, especially in the topsoil, has a striking potential to mitigate the loss of WHC and fertility and thus yield potential and variability among years, and also increase the CO_2 sequestration ability of the soil. These are further need at the light of the current climate change, which is mostly harming the fragile agro-ecosystems of drylands.

Mediterranean soils are the result of a complex genesis (Lagacherie et al., 2018), and thus SOC in the Mediterranean is determined by pedogenesis, erosion and especially in agro-ecosystems it is influenced by tillage. Many territories are cultivated and grazed by sheep and goats intensively. Catchment country scale assessments have a large importance if we consider the food security aspect. For this reason, focus of the work was the catchment scale. Frequent tillage have negative effects on SOC accumulation and soil resilience to erosion, desertification and climate change (Kämpf et al., 2016; Novara et al., 2013; Schillaci et al., 2017b), especially when tillage is intensive. This implies that SOC management plays a direct and crucial role in the world economy and is strategic to combat hunger and poverty. A number of agronomical management measures can be adopted to mitigate loss of carbon and preserve carbon-rich soil aggregates and preserve soil ecosystem services. The most important agronomical management measures include land use, land cover, the choice of crop species and genotypes, and soil management techniques, especially tillage. However, the role of each of these techniques and their interaction on SOC concentration, stock and spatio-temporal dynamics appear far to be clarified since it varies with the environmental traits of the site under study (Álvaro-Fuentes et al., 2012; Haddaway et al., 2016; Parras-Alcantara et al., 2016) and likely with the gross income of the population in the area and nation. In addition, the lack of data from many areas strongly impairs the ability to produce reliable indications on the site-specific management. Such information would

allow for the prescription of valuable actions able to preserve SOC and produce incentives to the actual agricultural production system and potential income derived from soil security.

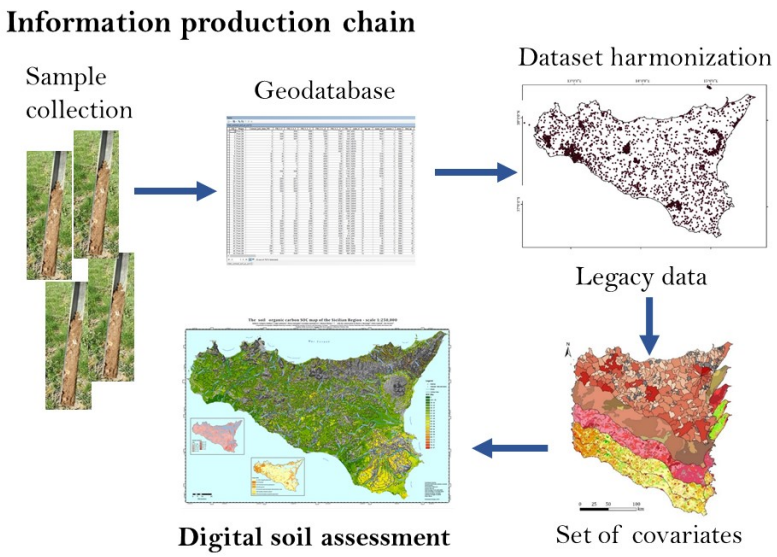


Fig.1 *Example of the conceptual workflow for this DSM experiment*

The study area of the dissertation is a reference semiarid area (Sicily, Italy). This area was used for the estimate of the importance of land use and soil erosion potential on SOC variation in time (both as % and absolute compared to the initial) at varying soil type and aridity of the environment. Sicily has great potential as an open laboratory for studies about ecological issues and anthropic pressure on the agro-ecosystems thanks to the variability of its traits and deep knowledge of its soils. Indeed, a total of about 7000 soil samples corresponding to ca. 2700 georeferenced points (more than 1 point each 10 km^2, Fig. 1A) are available, with information on SOC concentration, bulk density and soil texture. In addition, Sicily has variable, but on average high, demographic density and % area cropped in its territory, an ancient environmental and sociological history, a high climatic variability, several land uses and dominations from different populations, which introduced various plant species and management techniques and environmental heritages. The sum of such conditions makes Sicily an open and well suited laboratory to study the impact of anthropic pressure and of environmental variation at large scale (ecosystem level) and micro scale (few squared km), and of land cultivation and management on other environmental traits, including SOC distribution and dynamics. Finally, the present results can imply both agronomic and policy consequences at the district level and call for an intervention on soil fertility to maintain agriculture productivity (Acutis et al., 2014; Dono et al., 2016).

1.4 Objective of the thesis and Synopsis

The spatial scales in this dissertation range from plot to a third order catchment and the temporal scales from few years to almost three decades. Complexity is the word that best suit the Mediterranean soils (Yaalon, 1997). In the first chapter was provided an analysis of the literature available in the two most used databases: SCOPUS and Web of Science, out of that we can confirm that the two database have around 80 % of overlap, forcing the researcher looking at both. Non-peer reviewed reports were not taken into account. In the second chapter, the Legacy database have been carefully checked for potential errors of reporting and a 3D data clearing, screening and preparation for the modelling and mapping issues was performed. Such an optimization procedure was accompanied by a first test on the soil organic carbon (SOC) variation at varying depth, land use and soil and climate properties. In the third chapter, a digital soil mapping (DSM) experiment of the SOC stock in Sicily agro-ecosystems with Stochastic gradient treebost (SGT) was performed. It turned out an effective method proven for unbalanced, non-normal distributed data. SGT is in the family of the Boosted Regression trees models (BRT), with recursive partition model based on classification and regression trees CART (Elith et al., 2008). In this family of models, the effect of each predictor is represented after accounting for effects of other predictors. The splitting rules are based on binary splits on sequential explanatory variables. Input explanatory variables are weighted in subsequent trees and weights are applied in such a way that explanatory variables which has been poorly considered by previous trees has a higher probability of being selected in the new tree (Strobl et al., 2009). This sequential approach is the boosted (stochasticity) part of the regression tree meaning that the model takes in account as much as possible the explanatory variables available and avoid overfitting. This feature accounts for the main difference for another well-known Machine Learning method called Random Forest in which the explanatory variables have an equal probability of being selected for the next tree. This result was obtained by applying a BRT for the prediction of topsoil organic carbon stock (0-0.3 m) and examined the effects of 17 predictors, both continuous and categorical, that can be divided in five groups (climate, soil property, topography, vegetation, and land-use history). Aim of this paper was to produce a reliable and highly accurate topsoil SOC stock map by a robust mapping method and comparing this result with the newest global and European benchmarks available.

The fourth chapter dealt with SOC content mapping over Sicilian agro-ecosystem by using the two legacy most widespread soil sampling campaign, namely 1993 and 2008, and computing the expected SOC dynamics in three main agricultural land covers: i) field crops; ii) Vineyards, Olive grove and fruit berry trees plantations; iii) semi-natural areas and complex cultivation patterns. Semi-natural areas are defined as patches of land that has around the 50% of their coverage occupied by natural vegetation such as shrubs, wood, grassland and mountains. In this experiment, BRT model has been

feed with a set of explanatory variables coming from: i) topography, ii) soil properties, iii) remotely sensed, iv) land cover. Based on the Legacy data of each sampling year, 1993 and 2008, the models translated to a better performance in deployment when remote sensing covariate has been used than without their use in model. Although the methods of analysis are the same for the two sets of data, the most recent sampling campaign was not originally aimed at discover the dynamics of SOC and the other agricultural related traits but to increase the spatial density to produce detail scale pedological map, that are not yet published. The predictive space (made up by the aforementioned covariates) differed slightly, so the discovery of a pattern of SOC increase was due to the input data (that are object of further laboratory analysis) that were since the beginning higher in the most recent sampling campaign. However, the two stratified sampling does not certainly capture the extreme values as well as one might anticipate. Where the majority of the data were aligned, by the use of a common model one could highlight differences due to change in land use or even change in the management directly, that is not possible for the constrains that the experiment had.

The last chapter is a progression of an aim of the dissertation, i.e. quantifying the goodness of the most used DSM techniques and the local uncertainty at some unsampled location. Having reached relatively high R^2 (0.68 for SOC prediction with BRT), the model results were compared with other four classes of models: i) linear models, ii) geostatistical kriging based models, iii) Random forest, iv) Hybrid, regression kriging and regression kriging boosted regression trees. The dataset used is the 2008 sampling campaign. The topic is relatively new and proved BRT as the best machine learning model solution for SOC mapping, which confirmed previous results.

Chapter 2- Modelling of Soil Organic Carbon in the Mediterranean area: a systematic map

From: Schillaci, C., Saia, S., Acutis, M., 2018. Modelling of Soil Organic Carbon in the Mediterranean area: a systematic map. Rend. Online della Soc. Geol. Ital. doi.org/10.3301/ROL.2018.68.

Keywords: digital soil mapping; soil carbon model; spatial modelling; systematic map; carbon stock.

Abstract

A general feature of soil health is the sustainment of soil organic carbon (SOC) concentration and its stock. Digital soil mapping (DSM) development allowed for the implementation of soil properties mapping at various spatial and time scales. However, many of these studies were made in temperate or cold environments from central and northern Europe or United States or in stably arid ecosystems of Australia. Geographical information on the SOC are often fragmented, and this does not allow for a comparison on SOC regional variability in contrasting areas. Here a systematic research of peer-reviewed papers in the Web of science (WoS) and Scopus databases was carried out to highlight knowledge gaps in SOC studies in the Mediterranean area. The systematic searches identified 500 articles in WoS and 750 in Scopus, but only few of them were eligible as ad hoc studies. Regarding WoS, after screening, 150 studies were further analysed for inclusion in the map and only 128 included in the final map (1995-2018). From Scopus, only 104 studies were included in the map (1995-2017). Of all the countries around the Mediterranean Basin, report studies on SOC are available for 15 countries, only. Data gaps identified included the absence of long-term monitoring networks in the south of Europe, a scarcity of information from countries on the eastern coast of the Adriatic and Mediterranean Sea and almost lack of detailed information on SOC models and maps from north Africa. Model exportation built in neighbourhood countries (e.g. from Sicily, Italy, to northern Tunisia, or Andalusia, Spain, to northern Morocco) are strongly needed.

2.1 Introduction

Mediterranean areas have experienced large anthropic pressure, including fires and intensive cropping and other inappropriate management. This caused damage to the natural and agricultural ecosystems that has brought to severe land degradation by means of desertification, soil erosion and landslides, and reduction in soil organic carbon (SOC) (Persichillo et al., 2017; Saia et al., 2017b). In these areas, soil organic carbon (SOC) accumulation is peculiar and hampered by an unfavourable water-energy balance, especially in agricultural land where bare soil often occurs for several months, including those characterized by hot summers and discontinuous rainfall. Nonetheless, such issues can strongly vary with land use, its management, and morphology (Schillaci et al., 2017a). However, information on the region variability of SOC are fragmented, and this often does not allow for a

comparison of SOC evolution trends in contrasting areas that share similar land use management or SOC accumulation trends in similar areas with contrasting soil use management. In addition, this lack of information also impair the policy decisions about the soil use and management to preserve SOC stock, maintain soil fertility and reduce the environmental impact at broad areas (Pielke et al., 2002). Indeed, many reports pointed out at the SOC pool as a major player of the greenhouse gases (GHGs) emissions (e.g. Don et al., 2012; Lal, 2004).

The fragmented knowledge of SOC drives research at country scale, such as Spain (Aguilera et al., 2018), France (Martin et al., 2010) and even global projects such as 4x1000 (Minasny et al., 2017). In the present work, a systematic map (James et al., 2016) of the SOC studies (mapping and modelling) in the Mediterranean region was made to highlight information gaps. Such gaps impair the modelling of SOC and other soil properties to highlight the relationships with environmental management, which, for these areas, can be extremely variable depending on the subdomain of SOC or predictors (Lombardo et al., 2018; Schillaci et al., 2017b). Thus, aim of the present work is to analyse the coverage of two research databases, namely Web of Knowledge (WoS) and Scopus of papers dealing with SOC maps in Mediterranean areas, systematically refine its list and summarise it in a conceptual map along with some cartographic representation.

Systematic map protocols are unequivocal means of collection of reports of a given topic and were indicated as a first step to perform a systematic review (James et al., 2016). For the standards in use in ecological evidence journal, huge investment are needed to perform a systematic map that take into account 'grey literature', since many reports, especially before the 1995, are not available online or are collected in private repositories. In the present work, fund and time limitations did not allow for a complete exploration of the grey literature repositories and only WoS and Scopus outputs were taken into account. A guideline to perform systematic maps was developed first by the Social Care Institute for Excellence (SCIE, at http://www.scie.org.uk/research/maps.asp) to collect and analyse results in social science. A formalization of the procedure related to environmental science was made by James et al. (2016), which offered a thorough description of how synthesize, investigate natural evidences, to collate data to make an inventory, and highlight research gaps. Thus, despite no grey literature was checked for, the systematic map here offers a feasible and reproducible means of selection of references pertaining to SOC maps in Mediterranean areas. Indeed, systematic maps may be helpful to observe trends in the literature.

With regards to the studies on properties that could be predicted in space and time, such as crop yield, ecosystem services, soil or bedrock properties such as geological strata, population flows, etc., systematic maps can also be coupled with an mapping 'as such'. In such kind of studies, systematic

maps can also provide synthetic information with an open framed question, e.g. about the amount of researches conducted in a place in term of density of results, reliability of former mapping/modelling procedures, spatial and temporal resolution available in the literature (James et al., 2016). Such information can be given as tables and graph with regards of the geographical coordinates of the studies or either can be coupled with a GIS to produce spatial re-usable information that can be integrated or it can address future spatial or spatio-temporal researches. Few systematic maps and systematic map protocols has been published yet as stand-alone research ideas along with cartographic outputs of the literature research (Bayliss et al., 2016; Haddaway, 2014; Randall and James, 2012; Thorn et al., 2016).

A systematic research was conducted with reports dealing with studies on soil organic carbon modelling and mapping in the Web of science (WoS) and Scopus (Mongeon and Paul-Hus, 2016). The second aim of the work was to build a geodatabase for the Mediterranean region and highlight areas uncovered by any study.

2.2 Materials e Methods

A Systematic map was made according to the systematic process described by James et al. (2016) applied to the WoS and Scopus databases. Queries were built with Boolean operators after multiple trials from simple to complex strings, looking for the maximum coverage of the research area, the maximum overlap of the researches, and the minimum number of researches out of the chosen topic (i.e. SOC) .

The general principles of making a systematic map is proceeding in sequential stages (Fig. 1).

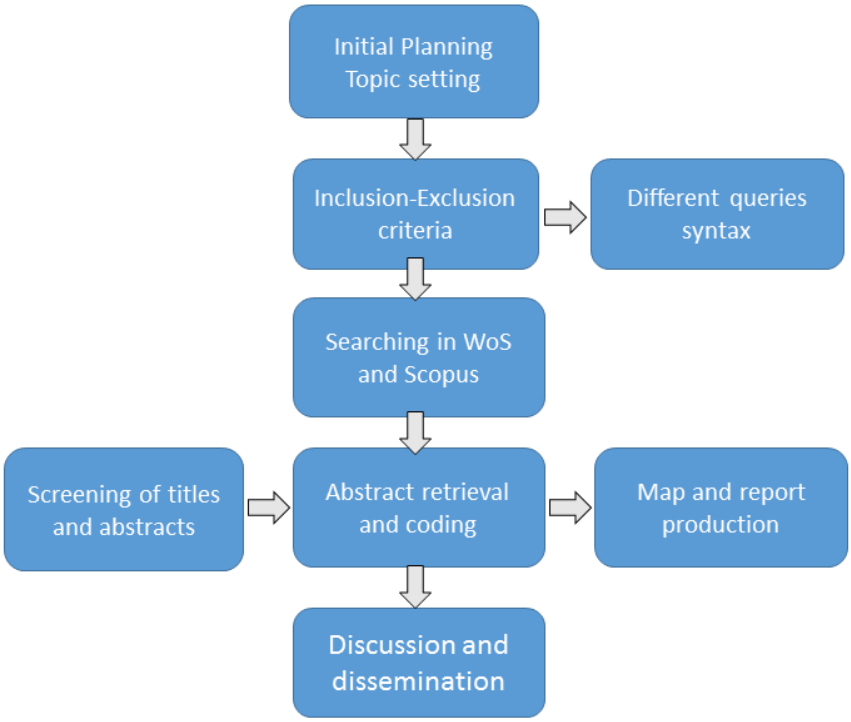


Fig. 1. *Workflow of the systematic mapping (adapted from James et al 2016)*

Following a systematic map approach, a search strategy has been defined and keywords were identified. Keywords of the research: soil organic carbon, mapping, model, Mediterranean, topsoil organic carbon. Boolean operators used; AND, OR and NOT.

Records dealing with SOC spatial and temporal mapping along with modelling at changing conditions (management, climate and land use) was done with an ad-hoc query reported in the supplementary material.

Records were exported as comma separated values and converted into an excel format to enable final editing of the strings. Title and abstract screening has been made, however no full text reading was performed with few exceptions (when the study area or methods were not clear) since this procedure does not imply systematic review of the results.

The database search was completed on the 10th October 2017. Using predefined categories assigned to each study for a suite of variables that describe the study's setting and design.

The following step called "coding" consisted in the database population with additional study metadata (e.g. extension of the study site in a qualitative way) using predefined classes, for a suite of variables that define the study's location and methodology design.

As the previous research schemes, the present work was a systematic search of potential relevant studies between 1995 and 2017.

Finally, a geodatabase was built along with cartographic output to make results easily intelligible for specialist and practitioner.

2.3 Results and discussion

The literature search on Web of Science "all collections" and Scopus yielded 128 and 104 experimental reports, respectively, of which only 19 in common (Supplementary Table 1), for 213 papers in this topic. Many of the original reports retrieved in the databases by means of the search were discarded since they did not include a Mediterranean area. More than three quarter of reports were however from France (20%), Italy (31%) and Spain (27%) and very few or no reports were found from countries on the eastern side of the Adriatic sea and North Africa, especially Libya. Review papers were discarded. In addition, 12 papers were reporting on the whole Europe and one on the whole world. The results of this work was in the form of a geographical database (**Fig. 2**).

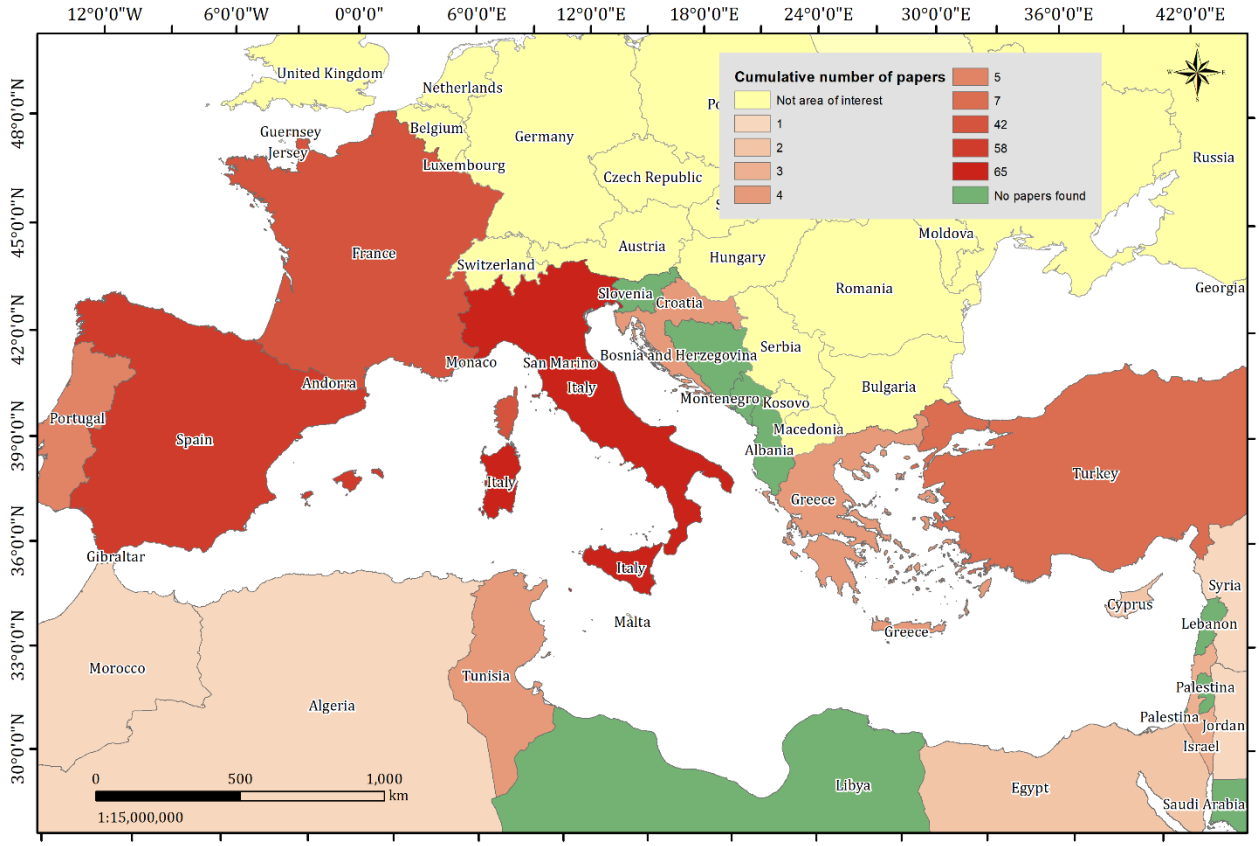


Fig. 2. *Map of the total number of articles per country. Number in the legend indicates total number of countries.*

This high discrepancy on the provenance of data have implication for the relationship between European soil(s) and its impact on the Mediterranean area. Haddaway et al. (2014) showed that systematic map targeting key environments could give indication on the potential impact of lack of knowledge in an area. Indeed in the Mediterranean areas taken into account in the present study, and especially those with semi-arid and arid ecosystems, the area covered by arable land is predominant compared to other land use (Schillaci et al., 2017b) and such area appeared to change few with time (Schillaci et al., 2017a). Arable land in these areas are mostly cropped with cereals and legumes growing from early winter to late spring and left with bare soil during summer and fall, in which rainfall have high intensity; permanent plantation which cover also an high percentage of the cultivated surface are offering a protection to soil aggregates and then lower the water erosion. This implies an impact in term of environmental acidification, eco-toxicity for aquatic fresh water, freshwater and marine eutrophication and Land use/soil organic matter and soil loss that can be high even for the less intensive system (Napoli et al., 2017; Novara et al., 2013; Saia et al., 2017a). The data in the present systematic map were mostly published after the year 2000 (**Fig. 3**) and rarely report data on SOC change in time. In addition, Scopus and WoS coverages showed some discrepancies, as also showed by Mongeon & Paul-Hus (2016).

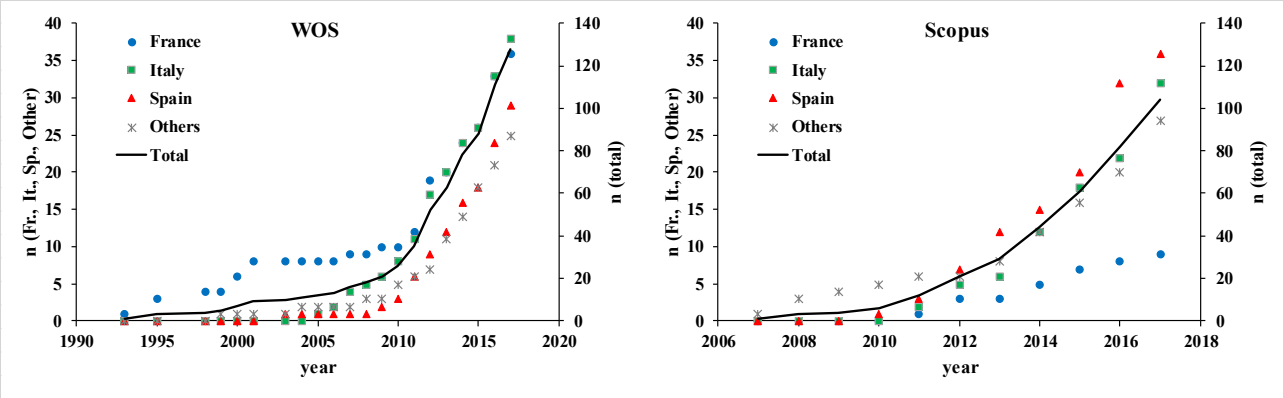


Fig. 3. *Number of articles by Country per year in the WOS (left) and Scopus (right) databases. Left ordinate axis for articles from France (Fr), Italy (It), Spain (Sp) and the sum of other countries. Right ordinate for total articles.*

With regards to land use change, other authors (Evrendilek et al., 2004; Novara et al., 2013) showed that it can imply very high change in C loss rate and total C loss, especially when natural/semi-natural stands are converted into crops. Such trends were confirmed for wider areas in both Mediterranean and continental climates (Barré et al., 2010; Schillaci et al., 2017a). Land use is likely changing in the northern, southern and eastern coasts of the Mediterranean Sea due to a number of European and

worldwide policy and economic issues (such as the European Union [EU] Common Agricultural Policy [CAP]).

Limitations of this study and survey include the inability to detect the grey literature and those papers that not reporting in the keywords, topic or title the search terms used here. In addition, there is scarcity of papers reporting on the whole Mediterranean area (by a world scale paper) and Europe, of which the coast is more populated than that of north Africa or the eastern coast, and can thus contributes more to the Mediterranean pollution and soil loss. Such scarcity implies that continental/broad studies are needed to normalize local/regional studies on SOC and its stock space-time variation (Malone et al., 2017), given that model exportation or downscaling imply the knowledge of the predictors behaviour at both the subdomains of the predicted (Lombardo et al., 2018) or predictor variable (Vaysse and Lagacherie, 2017).

In addition, such maps are needed to normalise for method diversity (sample density, modelling/mapping algorithms, area coverage; information gaps resulting from area covered at extremely low or uneven sample density and uncovered areas) in local modelling. Nonetheless, model exportation could be a rapid way to gain insight on the soils of northern Africa, given the similarities of these environments with those of southern Italy or Spain, as also confirmed by the share in key plant species and similarities in the natural populations of these species (Giovino et al., 2014; Mateu-Andrés et al., 2013; Said et al., 2016). The maps of SOC and other soil properties built for the Mediterranean area and an increased amount of local or regional reports are also needed to infer on SOC and soil relationship with release of GHGs in this peculiar environment (Leip et al., 2008). Peculiarities of the Mediterranean soils are indeed the richness in carbonates (Chevallier et al., 2017; Perri et al., 2016) and variation of soil pH upon topography, climate and management, which can affect CO_2 and N release in both cropped and natural soils from inorganic source, plant-derived, and native carbon pools (Badagliacca et al., 2017; Bleuler et al., 2017; Heinze et al., 2018).

2.4 Conclusion

This work has employed a systematic procedure to map the research paper on the WoS and Scopus databases about SOC mapping or modelling conducted in the countries around the Mediterranean Sea. An online platform and a GIS environment were combined. Out of our systematic procedure of selection, a thorough vision of the present published scientific developments could be easily achieved, and it is easily updatable and extendable to other environments, although it presently lacks of reports from the grey literature.

The main result of this systematic map is that many studies reported are not directly comparable due to very different methodology, time or area extent or sample resolution. The database generated here comprises 213 studies, which offers an overview geographical literature base, and provides evidence concerning a wide range of conditions of SOC in Mediterranean ecosystems and land uses. However, the studies available are unevenly distributed among countries, with France, Spain and Italy well sampled/studied, and the southern Mediterranean countries and the east cost of the Adriatic seas strongly underrepresented. Most notably, no reports in these databases was found from areas such as Libya, Lebanon, Palestine, and the eastern coast of the Adriatic seas, which lack hampers the exportation of models to these countries. This poses a concern on the implication of SOC management for both freshwater and marine environments, especially if taking into account of the shallow seabed and scarce water exchange with other sea areas.

Different output in some European region could have depended on the effect of European aids (e.g. CAP, LIFE projects, European Neighbourhood Policy, act.) that contributed to the intensification of agriculture in agro-ecosystems or the abandonment of lands (either degraded or interested by secondary successions) or to the sampling of soils.

Results from this systematic map also suggests that some studies shall be soon addressed to model exportation among similar countries and confirmation with ad-hoc soil sampling and analysis with robust methodologies (Conforti et al., 2017). Also, mapping procedures of the whole Mediterranean area should be produced, given the importance of maps for a plenty of policy aims (but see Pereira et al. 2018). For example, many areas from Sicily (south of Italy) and southern Spain frequently share geological strata, bedrock, soil types, and other environmental (climatic) variables with those of northern Tunisia or other parts of northern Africa (Brahim et al., 2011; Darwish and Fadel, 2017; Gargouri et al., 2013; Henry et al., 2009; Lal, 2007).This systematic map can be easily updated (Bayliss et al., 2016) and extended for additional coding or used as a basis of future secondary researches in the forms of systematic reviews.

Acknowledgments

Geosciences and Information Technologies - Sezione della Società Geologica Italiana. The authors are grateful to the society for the invitation to contribute to this special issue. This work was also improved after fruitful discussion with Dr. Katy James [Harper Adams University], whom the authors are very grateful to.

Supplementary table caption

Supplementary Table 1. Table of the results of the queries in Web of Knowledge ("organic carbon" AND "mapping" AND "Mediterranean" NOT "sea") *OR* **TOPIC:** ("organic carbon" AND "model" AND "Mediterranean" NOT "sea") and in Scopus: "soil organic carbon" AND mapping AND modelling AND Mediterranean OR "topsoil carbon mapping" AND NOT sea.

Chapter 3- A simple pipeline for the assessment of legacy soil datasets: an example and test with soil organic carbon from a highly variable area

From: Calogero Schillaci, Marco Acutis, Fosco Vesely, Sergio Saia, (CATENA doi.org/10.1016/j.catena2018.12.015.)
Keywords: Bulk density, Texture, R, GIS, CORINE land cover, LASSO.

Abstract

Legacy databases provide unique information on soil properties and act as a guide for the setup of monitoring processes. However, their use requires an evaluation of their drawbacks, especially when aiming to model the soil traits by depth. We set up a procedure for the integration and error correction of a soil legacy database. This database consisted of 6994 records in its original form and 6674 records after correction. These records were collected from 2886 locations in the south of Italy on a 25711-km^2 island (Sicily, Italy). Samples were taken in arable lands (5471 records), orchards, vineyards and seminatural lands (3010 records), and woodland and natural areas (1203 records). The procedure for the integration and error highlighting improved the prediction of soil organic carbon (SOC), and a general linear model with covariate selection by Least Absolute Shrinkage and Selection Operator (LASSO) tested the procedure. We focussed on exploring the amount of legacy information as georeferenced soil properties. SOC and fine earth fractions were analysed for each sample. Bulk density was provided for only 20% of the samples. These results will help to account for the legacy data available and propose an analysis to harmonize an SOC dataset; highlight missing or incorrect data; summarize data; and offer synthesis criteria for benchmarking SOC in different land uses and pedological areas. In addition, the results may stimulate funding bodies to support research in an open data frame, which can be turned into more sustainable use of resources, improved communication between governments and farmers, and the production of standard datasets that meet and facilitate the requirements for regional agro-environmental modelling.

3.1 Introduction

Information on soil organic carbon (SOC) and its dynamics is of growing interest to many different user groups. The increasing interest is because SOC is the main soil fertility indicator, plays a pivotal role in CO2 sequestration and other greenhouse gases, and therefore is related to global warming. SOC stock is a potential sink of C considering both the biosphere and atmosphere (Batjes, 2016). SOC was recently recognized from a survey conducted by Campbell et al. (2017) as among the first three important soil functions by non-science stakeholders.

However, data on soil properties and its past trends can be absent or difficult to achieve for several reasons that include the cost of sampling and analysis and the risk of disturbing fragile ecosystems (Lobry de Bruyn et al., 2017; Rial et al., 2017b). Legacy soil data are considered as a scientific priority

for digital soil mapping (DSM) (Odeh et al. 2012). However, soil sampling included in legacy databases can be very heterogeneous because of the aims of sampling campaigns, precision of the measures and data reported, distribution of samples by depth and maximum sampling depth, spatial coverage, and sample design (Arrouays, 2017; Dobos et al., 2010).

For instance, a sampling campaign can be conducted to obtain information on cropped or non cultivated soils, on soils that underwent erosion (Borrelli et al., 2017), for building pedological information or checking for soil fertility, for communicating the soil fertility status to farmers, etc. (Ingram et al., 2016; Sánchez et al., 2016). Lack of quality control can produce not usable legacy databases or provide information not closely responding to the soil conditions (Ramos et al., 2017).

When unevenly distributed, subsets of data can be affected by local tendencies, seasonality and general climate trends (Rial et al., 2017b). Extensive soil surveys have been undertaken in Europe (Bradley et al., 2005; Orgiazzi et al., 2017) and the Americas (Guevara et al., 2018; Liu et al., 2013; Sperow, 2016). In particular, Guevara et al. (2018) showed that country-specific soil estimates can depend on both geopolitical issues and environmental traits. In the Mediterranean area, many studies about SOC mapping and modelling have been undertaken: Schillaci et al. (2018) reported more than 300 studies from this area after a systematic map was prepared in the 'Scopus' and 'Web of Science' databases. In some of these areas, monitoring networks are presently operating and used for DSM. However, few studies have been merged in a unique mapping and modelling procedure (Grunwald, 2009), which is strongly needed for i) land management, ii) programming of ecosystem service provision, iii) present and newly to-build infrastructures and above all, iv) direct agronomical applications. In addition, limitation in the SOC accumulation in Mediterranean areas frequently occurs due to an unbalanced energy (i.e., solar radiation) per unit water available in the soil and long periods of bare soil and crop stubble burning (Egli et al., 2007; Huang et al., 2018; Lombardo et al., 2018), as well as application the Set-Aside directive of the Common Agricultural Policy. Indeed, stubble burning can dramatically reduce SOC accumulation, since straw is the main C input in some soils (Li et al., 2018).

Odeh et al. (2012) increased the attention on legacy data and summarized some of the main procedures necessary to make legacy data usable. However, simple procedures for error checking and data harmonization when creating a legacy database are few and are mostly provided for tropical or cold regions (Aitkenhead and Coull, 2016; Hendriks et al., 2016; Kempen et al., 2015; Stumpf et al., 2016; Sulaeman et al., 2013; Sun et al., 2015), and few papers deal with Mediterranean soil inventories and strategies for their building (Bogunovic et al., 2017; Fernández-Getino and Duarte, 2015; Francaviglia et al., 2017; Ramos et al., 2017). Indeed, procedures for checking putative errors

strictly depend on the area and traits chosen (e.g., when checking for soil pH or reliability of SOC measures). In particular, important features to check the quality of soil legacy databases pertain to various aspects of the database compilation, soil survey, and reporting of methodology (Krol, 2008), which implies that crucial pre-processing and harmonization of the data are needed for any further application (Gosling et al., 2017). Lack of these procedures impairs the extraction of knowledge from the data (Rivera et al., 2015).

Here, we present key points for a technical screening of legacy datasets to facilitate access to the digital soil database repository (Huang et al., 2017), optimize land suitability assessment (Hallett et al., 2017), and identify patterns of SOC in a dataset that can constrain modelling. Such an analysis aimed at assessing the measurement quality of the SOC concentration and texture and their accuracy for further DSM and updating. For the analysis, we used the database of the sole mainland area of Sicily, the most southern Italian region, which accounted for 6674 records. We applied a pipeline of data integration and error correction, which led to discarded records, and tested the effect of the optimization by predicting the log-linearized SOC by general linear models. Additional information was provided such as accuracy of texture, land use at the time of survey and CORINE land cover close to the time of survey, accuracy of the geographical position of each site and land use recent history (from CORINE, 1990 to 2012) of the sampling locations. The soil legacy dataset of Sicily was already used, with regards to the sole A and Ap soil horizons of the cropped soil, in some previous DSM applications (Schillaci et al., 2017b).

3.2 Materials and Methods

3.2.1 Study area, climate, soils and sampling database

Sicily is the largest Mediterranean island (25,711 km2). The island has high soil heterogeneity due to a high variability in geological strata, intraregional climatic features, agronomical techniques applied and land use variability. In addition, various land uses coexist closely (Fig. 1). We referred to the Sicily mainland soil legacy database (SMSLD). This database originally included 6674 records of SOC and 1426 of bulk density (BD) collected in 2843 and 1080 sites, respectively, up to a mean depth of 83.8 cm (median=75.0 cm, s.d. of maximum depth sampled=44.6 cm). Assessorato Regionale Territorio Ambiente (ARTA) provided the SMSLD as georeferenced values derived by pedological profiles and soil pits spanning 1967 to 2008 and containing 44 sampling campaigns. The BD was not sampled before 1993. The density of sampling was on average 0.12 sampling sites per km2. A physiographic description of the island along with a physiographic map is given in Schillaci et al. (2017a) and Fantappiè et al. (2016).

Fig. 1 *Main agricultural land uses, a) cropland (during early spring, a wheat field reported as CORINE 2.1.1 or 2.1), b) vineyards (during early summer, reported as CORINE 2.2.1 or 2.2), c) olive grove (during early summer, reported as CORINE 2.2.3 or 2.2), d) peach orchard (during early fall, reported as CORINE 2.2.2 or 2.2).*

3.2.2 Legacy database

The SMSLD contained the following: sampling code, sampling year, site code, latitude, longitude, upper and lower limit of the sample in depth, SOC, sand, silt and clay content, actual land use (as CORINE codes), and values in rows. Data were structured as a matrix. In this study, we made computations and individuation of drawbacks of the SMSLD by Microsoft Excel and R, data visualization with ArcGIS® by ESRI and SAGA GIS (Conrad et al., 2015), and testing of the procedure for the integration and error correction with the SAS 9.2 statistical software package. Latitude and longitude were measured with handheld GPS receivers since the 90s, whereas coordinates of soil profiles and soil samples collected before 1990 were likely reported from the 1:10,000 topographic maps. The original coordinate system of the database was geographic WGS 84, and we converted the system to UTM, because the metric system was used for the calculation of the topographical attributes.

3.2.3 Detailed dataset description

General information on the sampling campaigns was drawn by the SMSLD itself; therefore, the description was made in the column of the original database. The SMSLD originally provided the following information:

1. "Survey code" (Fig. 2): the survey code was of pivotal importance because samples taken a few months apart could be grouped together and was also useful to merge surveys that spanned a few years of separation.
2. "Observation type": profile, denoted as P, or soil quality (as Q). Soil quality referred to the sampling campaigns aimed at an agro-pedological characterization.
3. "Number soil survey" was an identification number.
4. "Horizon number" indicated the position of the samples in the profile, with 1 the upper surface horizon.
5. "year" of sampling (Fig. 3). Unfortunately, no further information on the month or season of sampling was given. Therefore, for such a reason at a particular moment, the role of soil management in cropped soil on SOC and BD was uncertain. In particular, the entire region suffers from a scarcity of rainfall (and in general water) during late spring and summer.

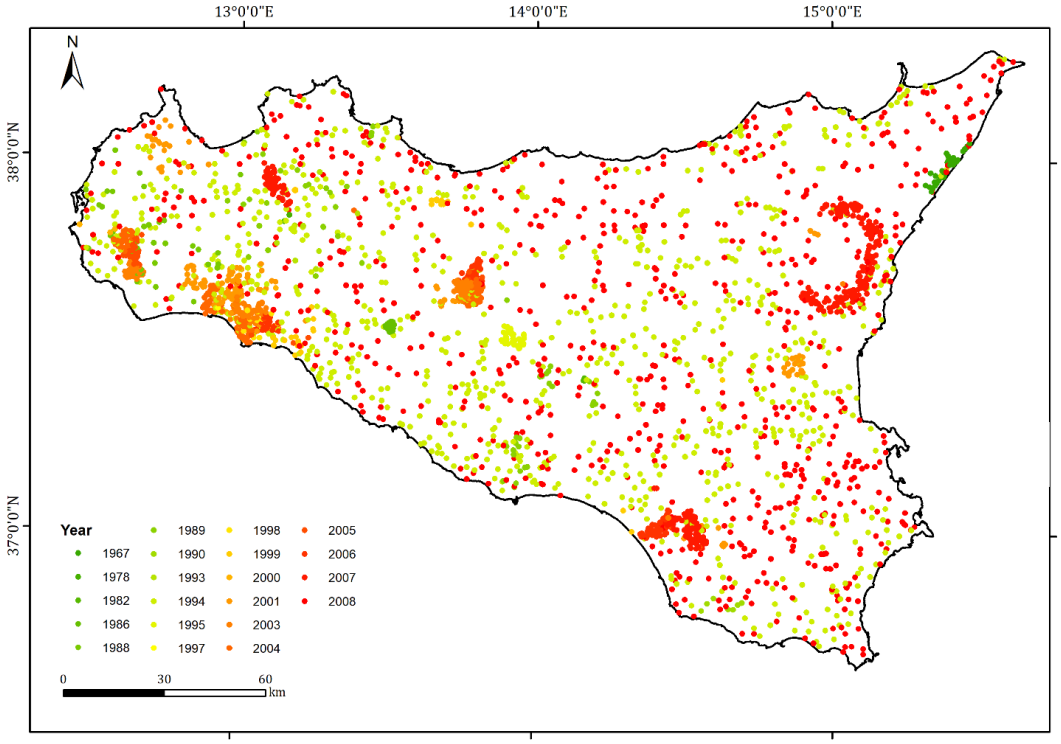
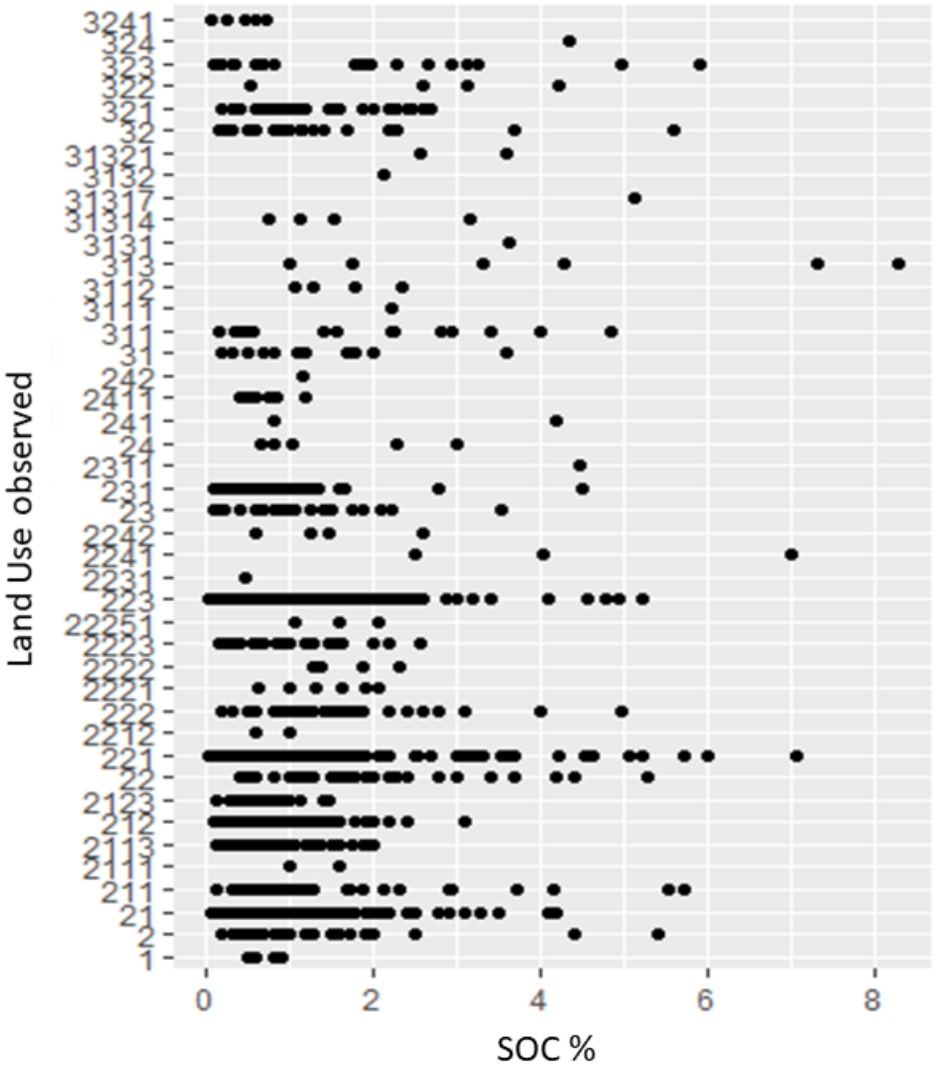


Fig. 2. *Traits and distribution of sampling campaign across the region*

6. “Sampling depth” was expressed as upper and lower limits of a sample. Since horizon depth may change in space, the depth referred to for each sample may also change with space and in different campaigns.
7. "Latitude" and "Longitude". Horizons belonging to the same profile had the same coordinates (See Suppmat. Section “R codes used”).
8. The observed “land use” was reported as CORINE codes. In the original SMSLD, the land use code was reported up to 5 digits from level 1 to 5, and then, we split it in single digits for purposes of analyses. See Suppmat. Fig. 1.



Supplementary Fig. 1. *Soil organic carbon (% of the fine earth fraction) by land use observed. These new fields allowed for the selection of levels with contrasting depth of information (e.g. a layer with Land use 2 and another with land use 2.1, sharing the level one information, i.e. 2, agricultural use). Thus, categorization was made according to a first level grouping: 2 (agriculture) or 3 (natural and semi-natural); second level: e.g. 2.1 (arable) or 2.2 (permanent) crops; and third level: e.g. 2.1.1 rainfed field crops or 2.2.1 vineyards.*

9. "SOC" was expressed as % (i.e., dag SOC kg-1 soil dry weight) of the fine earth fraction. The method reported was Walkley-Black with the chromic acid wet oxidation method (Walkley and Black, 1934).
10. Texture was reported as clay (<2 µm), total silt (2 to 20 µm), and total sand (20 to 2000 µm). To confirm that sums of fractions would sum to 100%, the actual sums were considered valid when between 97.5 and 102.5%. Data out of this range were discarded. The rest of data were recomputed to 100% by weighting each fraction. The textural plot was plotted in R with the package "ggtern" (See Suppmat. Section "R codes used").
11. The "% of gravel in the total soil". This field reported the amount of gravel (in weight) per unit weight of each sample.
12. Soil texture type according to the USDA description (referred to as "usda" in Suppmat. tab. 2 and Suppmat. tab. 3).

3.2.4 Data integration and error correction

The first steps of the analysis were for quality control and outlier detection, harmonization and duplicate removal (see Suppmat. Section "Data harmonization detail"). The steps consisted of the following:

- exclusion of samples without complete texture data;
- exclusion of samples from litter layers;
- exclusion of samples with land use =1, i.e., "Artificial surfaces";
- correction of likely wrong samples that indicated a null or negative thickness, i.e., a lower boundary shallower than the upper boundary.
- elimination of doubled samples (by means of the "pivot" tool in Excel including the sampling campaign, year, coordinates and depth at one time as discriminant fields).
- identification of clustered sampling campaigns, in GIS, to consolidate those sampling campaigns spread in close years.

3.2.5 Data consolidation and synthesis

Samples were aggregated by sampling code. Spatial coverage statistics were assessed (density per sampling campaign, year and land use). Descriptive statistics, statistical distribution of the SOC, texture, thickness and main depth were also computed. Temporal and spatial distribution of the georeferenced observations were shown. Accuracy assessment of the coordinates reported was made by checking for the number of decimal digits of each site. See also Suppmat. section "Metadata" for additional information.

3.2.6 Covariates

We created a list of soil properties predictors (Suppmat. Tab. 1), which include a range of climatic traits (mean annual rainfall and temperature and some climatic indices), geographical traits (slope, altitude, catchment area, etc.), land use by the CORINE land cover. These 14 covariates were chosen among those highly or scarcely important in SOC or SOC stock prediction in the same area as observed in previous experiments (Lombardo et al., 2018; Schillaci et al., 2017b, 2017a).

3.2.7 Test of the corrected dataset predictive ability

To test for the correction procedure, we compared the goodness-to-fit of 4 models built on the optimized and uncorrected (i.e., original) databases. Such a comparison was made by applying to each dataset (corrected or original) a general linear model with two strategies of predictor selection: no selection method (i.e., compulsorily retaining all predictors in the modelling process) and Least Absolute Shrinkage and Selection Operator (LASSO) (Camilo et al., 2017; Veronesi et al., 2016). The sequence of models in LASSO was determined including at any step the coefficients of the parameters for the model by using ordinary least squares. The test was conducted both on the complete database and separately after splitting the complete databases into two sub-databases: one pertaining to samples for which the deepest information was shallower than 50 cm (named "DIS50") and another with information deeper than 50 cm (named "DID50", which was the counterpart of the former), irrespective of the mean depth of the layer and sample thickness. Therefore, layers included in the 'subsoil' information stratum could also have a wide part between the field surface and the 50-cm limit (e.g., a layer from 10 to 55 cm). DIS50 and DID50 were needed to highlight the number of samples with minimum information content up to the 50 cm depth. Such a depth was chosen since most of the agricultural soil, especially in field crops, is usually ploughed up to various cm depths. Thus, the split into 2 databases allowed accounting for the greater amount of data in DIS50 than that in DID50 and the likely role of tillage in SOC accumulation in agricultural soil, which rarely went beyond a 50 cm depth. In the modelling process, land use was included as CORINE level 1 as a main factor (referred to as SU) and as CORINE level 2 as a nested factor (referred as LU). Furthermore, metadata were created.

To do so, we used the GLMSELECT procedure in the SAS/STAT 9.2 statistical software program. GLMSELECT is a general linear model for normal responses capable of individuating regression coefficients between a dependent variable and various independent variables. Both the dependent and independent variables can be dichotomous, class or continuous variables. Such a procedure does not handle random effects. Thus, SOC was linearized by log transformation. Within the LASSO selection method, the LSCOEFFS option was applied.

LASSO avoided overfitting, discarded noninformative covariates, and handled multicollinearity of predictors. LASSO was stopped according to the minimization of the Schwarz Bayesian information criterion (SBC). See also Camilo et al. (2017) for additional information on the LASSO criteria of variable selection and variable estimation.

Model statistics were provided, including the R2 and adjusted R2 statistics (ADJRSQ), Akaike's information criterion (AIC), corrected Akaike's information criterion (AICC), the Sawa Bayesian information criterion (BIC), the Mallows C(p) statistic, the predicted residual sum of squares statistic (PRESS), and the Schwarz Bayesian information criterion (SBC). These statistics are directly proportional to the increase in model-to-data error.

Average square error (ASE) and the F and p statistics of the retained effects were computed. The variation, if any, of the standardized coefficients of the retained predictors at each step of each model building was plotted.

3.3 Results

3.3.1 Database assessment

The first correction procedure aimed at individuating layers reporting null or negative thickness. We highlighted that such an error was reported for 12 samples and that their coordinates included a total of 33 samples. In total, 317 lines were deleted from the original dataset (see Suppmat. Section "procedure for the integration and error highlighting" for details).

A procedure for highlighting likely 'very wrong' coarse fraction and land use reported was set up as follows:

- samples with coarse fraction higher than 80% were selected;
- in this selection, samples with SOC content higher than 1.0% or sand content lower than 80% were selected;
- we assumed that these fractions could be flawed by reporting errors, since high coarse fraction contents are unlikely associated with high organic C contents (Anderson et al., 1981); thus, these samples were deleted.

at the same time, with BD information (Fig 3).

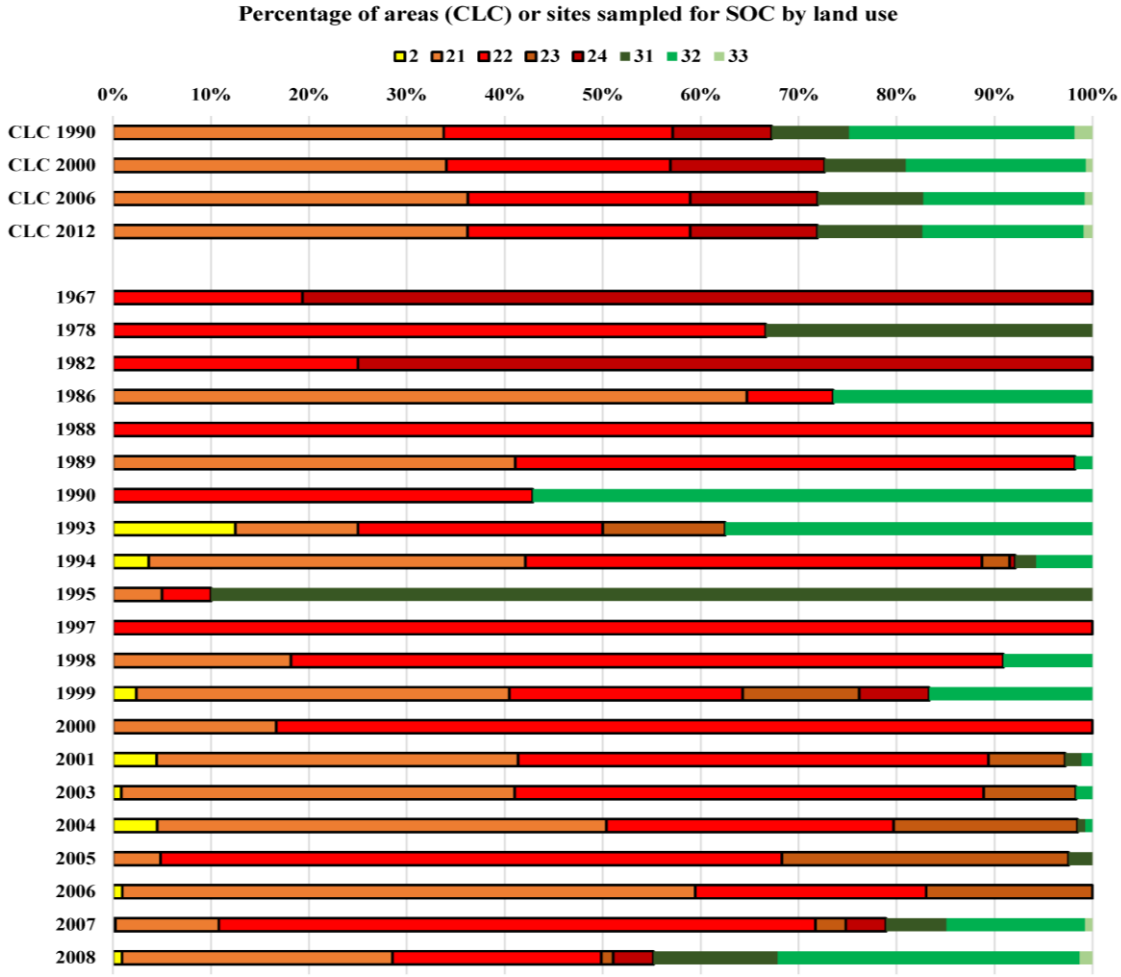


Fig. 3. *Percentage distribution of sample per land use observed by year of the sampling campaign. CLC 1990, 2000, 2006, 2012 indicates the percentage area distribution of the land use according to the CORINE maps of the relevant year. See the CORINE manual for correspondence between codes and land use.*

Tab. 1. *Number of sites sampled in each land use in which various soil layers were sampled, with information on the bulk density, if any, and minimum and maximum depth sampled.*

amount of layers sampled	without bulk density data						with bulk density data					
	Land use = 2			Land use = 3			Land use = 2			Land use = 3		
	amount of sites sampled	minimum depth sampled	maximum depth sampled	amount of sites sampled	minimum depth sampled	maximum depth sampled	amount of sites sampled	minimum depth sampled	maximum depth sampled	amount of sites sampled	minimum depth sampled	maximum depth sampled
1	676	6	200	215	3	110	779	10	130	84	10	115
2	460	25	205	139	14	160	121	20	180	14	20	140
3	464	40	180	100	35	165	44	44	165	2	50	150
4	200	50	210	35	80	220	25	65	230	0		
5	51	90	210	7	70	180	11	110	270	0		
6	13	90	240	4	140	280	0			0		
7	0	n.a.*	n.a.	0	n.a.	n.a.	0			0		
8	3	170	250	1	120	120	0			0	n.a.	
9	0	n.a.	n.a.	2	100	120	0	n.a.		0		
10	1	120	120	1	120	120	0			0		
11	1	220	220	3	120	130	0			0		
12	0	n.a.	n.a.	2	110	120	0			0		

* not available.

Tab. 2. *Land use depth of information (CORINE code level) by site for soil organic carbon (SOC) and bulk density (BD) data*

Land use depth of information	**Land use = 2**		**Land use = 3**	
	SOC	**BD**	**SOC**	**BD**
	number of sites			
1	58	35		
2	688	355	75	36
3	1319	523	419	58
4	218	74	60	10
5	18	8	18	10
6	7	6		

Number of non integer digits in the GPS datum					
Lon	**Lat**	***number of sites***			
2	**2**	1	1		
	5	1		1	
3	**2**	1	1		
	3	1	1		
	4	3	2		
	5	57	51	2	2
	6	1	0		
4	**2**	1	1		
	3	3	1		
	4	6	1	1	
	5	56	36	5	4
	6	14	4	3	
5	**2**	21	18	2	1
	3	13	12	4	2
	4	49	40	7	4
	5	717	545	55	40
	6	121	21	48	4
6	**4**	11	2	7	
	5	117	33	39	1
	6	1093	210	383	42

Tab. 3. *Descriptive statistics of the samples by soil organic carbon (SOC) and bulk density (BD) and Land use.*

	Land use 2 (Agricultural Areas)											
	all data from LU2		**"2" only indication**		**21**		**22**		**23**		**24**	
	SOC	**BD**	**SOC**	**BD**	**SOC**	**BD**	**SOC**	**BD**	**SOC**	**BD**	**SOC**	**BD**
n	5471	1308	112	37	2162	564	2671	606	336	87	190	14
Mean	1.067	1.204	1.048	1.160	0.907	1.238	1.201	1.165	0.900	1.256	1.331	1.348
Min	0.012	0.500	0.160	0.900	0.020	0.600	0.012	0.500	0.020	0.580	0.030	0.900
Q 0.025	0.100	0.700	0.186	0.900	0.100	0.700	0.100	0.700	0.100	0.900	0.115	0.965
Q 0.05	0.120	0.800	0.217	0.900	0.130	0.800	0.110	0.700	0.150	0.928	0.164	1.030
Q 0.25	0.500	1.000	0.500	0.900	0.500	1.099	0.500	0.900	0.490	1.099	0.630	1.303
Median	0.830	1.200	0.905	1.100	0.800	1.260	0.886	1.137	0.730	1.287	1.005	1.390
Q 0.75	1.280	1.400	1.300	1.300	1.120	1.400	1.420	1.380	1.053	1.400	1.670	1.440
Q 0.95	3.010	1.618	2.225	1.800	1.999	1.620	3.714	1.600	2.100	1.603	3.650	1.592
Q 0.975	4.136	1.700	3.584	1.800	2.789	1.660	4.703	1.711	2.800	1.696	4.299	1.621
Max	12.50	2.30	5.40	1.80	8.10	2.16	12.50	2.30	6.01	1.91	7.75	1.65
S.D.	0.99	0.27	0.81	0.30	0.68	0.26	1.17	0.28	0.77	0.21	1.17	0.19
Skewness	2.91	0.14	2.66	1.03	2.96	-0.04	2.53	0.33	3.29	0.06	2.26	-0.90
Kurtosis	13.51	-0.15	10.05	-0.18	16.19	-0.13	9.91	-0.05	15.27	1.01	7.07	1.41

	Land use 3 (Forest and semi natural areas)							
	all data from LU3		**31**		**32**		**33**	
	SOC	**BD**	**SOC**	**BD**	**SOC**	**BD**	**SOC**	**BD**
n	1203	118	477	47	699	71	26	0
Mean	1.576	1.206	1.587	1.195	1.563	1.213	1.615	
Min	0.010	0.550	0.010	0.550	0.020	0.800	0.060	
Q 0.025	0.100	0.696	0.157	0.606	0.100	0.878	0.110	
Q 0.05	0.151	0.729	0.200	0.658	0.140	0.900	0.155	
Q 0.25	0.510	0.900	0.630	0.800	0.460	0.900	0.640	
Median	1.050	1.203	1.140	1.330	1.000	1.200	1.120	
Q 0.75	1.970	1.485	1.978	1.605	1.940	1.450	2.273	
Q 0.95	4.679	1.650	4.076	1.664	4.823	1.590	4.420	
Q 0.975	5.969	1.671	5.183	1.670	6.455	1.620	5.733	
Max	20.10	1.70	15.64	1.70	20.10	1.70	7.07	
S.D.	1.74	0.32	1.58	0.39	1.85	0.27	1.58	
Skewness	3.65	-0.11	3.70	-0.17	3.65	0.10	2.02	
Kurtosis	23.02	-1.33	23.60	-1.65	22.67	-1.42	5.08	

Tab. 4. *Performance statistics of the models run with the complete database (Unsplit) or split by in two sub-databases: samples for which the deepest information was shallower than 50 cm (DIS50) and for which the information was deeper than 50 cm (DID50). Statistics are root mean square error (MSE), dependent mean, R^2 and adjusted R^2 statistics, Akaike's information criterion (AIC), corrected Akaike's information criterion (AICC), the Sawa Bayesian information criterion (BIC), the Mallows C(p) statistic, the predicted residual sum of squares statistic (PRESS), the Schwarz Bayesian information criterion (SBC), the model average square error (ASE) and the model F statistic.*

	Unsplit (Total database)				Split							
					DIS50				**DID50**			
	No Selection		**LASSO**		**No Selection**		**LASSO**		**No Selection**		**LASSO**	
Performance statistics	Corr	Uncorr	Corr	Uncorr	Corr	Uncorr	Corr	Uncorr	Corr	Uncorr	Corr	Uncorr
Root MSE	0.320	0.317	0.320	0.318	0.270	0.269	0.278	0.277	0.348	0.345	0.349	0.345
Dependent Mean	-0.102	-0.097	-0.102	-0.097	0.065	0.066	0.065	0.066	-0.265	-0.256	-0.265	-0.256
R^2	0.364	0.356	0.362	0.352	0.278	0.271	0.233	0.222	0.271	0.264	0.263	0.259
Adjusted R^2	0.362	0.354	0.360	0.351	0.272	0.264	0.232	0.220	0.265	0.258	0.260	0.255
AIC	-8515	-9029	-8507	-9014	-5296	-5572	-5138	-5393	-3715	-3970	-3707	-3971
AICC	-8515	-9029	-8507	-9014	-5295	-5571	-5138	-5393	-3714	-3970	-3707	-3971
BIC	-15187	-16014	-15179	-15999	-8593	-9016	-8437	-8839	-7088	-7510	-7081	-7511
C(p)	29.00	32.00	36.98	46.58	29.00	32.00	190.62	216.07	29.00	31.00	36.18	30.00
PRESS	685	705	686	707	244	252	256	265	413	424	414	424
SBC	-14992	-15797	-15052	-15878	-8419	-8822	-8383	-8785	-6913	-7321	-6992	-7402
ASE	0.102	0.100	0.102	0.101	0.073	0.072	0.077	0.076	0.120	0.118	0.121	0.119
F Value	136.04	124.16	209.49	222.99	44.93	40.94	125.13	122.51	44.42	41.98	85.79	72.36

From the second deletion procedure and on, we always checked whether the coordinate of the sample to delete was already present in the other deleted samples. No correspondence occurred.

Only 25 sites reported low detailed geographical information (i.e., with less than a 2 non integer digit precision in the coordinate reported, Tab. 2)

The sampled sites were widely distributed in the region only in 3 sampling campaigns (Fig. 2): a campaign in 1993-94 and two in 2008, one of which was a gridded scheme with a 20 km spatial resolution. Across these campaigns and the whole dataset, SOC had a wide range of variation in arable lands (2.1), vineyards (2.2.1), fruit trees and berry plantations (2.2.2), complex cropping systems (2.4), and natural areas (3.1 and 3.2) (Fig. 4 and Fig. 5).

Strong and well-represented information on soil texture occurred only for the agricultural lands (CORINE code 2) (Fig. 6), irrespective of the assigned category "Topsoil" (i.e., those samples with no information deeper than 50 cm) or "Subsoil" (i.e., those samples with information deeper than 50 cm). The CORINE code 3 lacked data with clay content lower than 20% and higher than 80%. Maximum depth sampled (Fig. 7) differed by land covers (see Suppmat. Section "procedure for the integration and error highlighting" for details).

3.3.2 Descriptive statistics of SOC and BD

SOC was on average 1.067±0.013% (mean ±standard error) in the land cover agriculture (CORINE code 2) and 1.576±0.050% in the land cover forests and seminatural areas (CORINE code 3) irrespective of the depth of sampling (Fig. 8 and Tab. 3). Bulk density was similar between the two land use groups within the CORINE code 2, and SOC was lower in arable land (2.1) and pastures (2.3) than that in permanent crops (2.2) and heterogeneous agricultural areas (2.4), whereas within the CORINE code 3, few differences in SOC among the forests and seminatural areas (3.1), shrubs and/or herbaceous vegetation association (3.2) and open spaces with little or no vegetation (3.3) were found. Skewness and kurtosis were high in all sub-datasets of SOC and relatively low in those of BD.

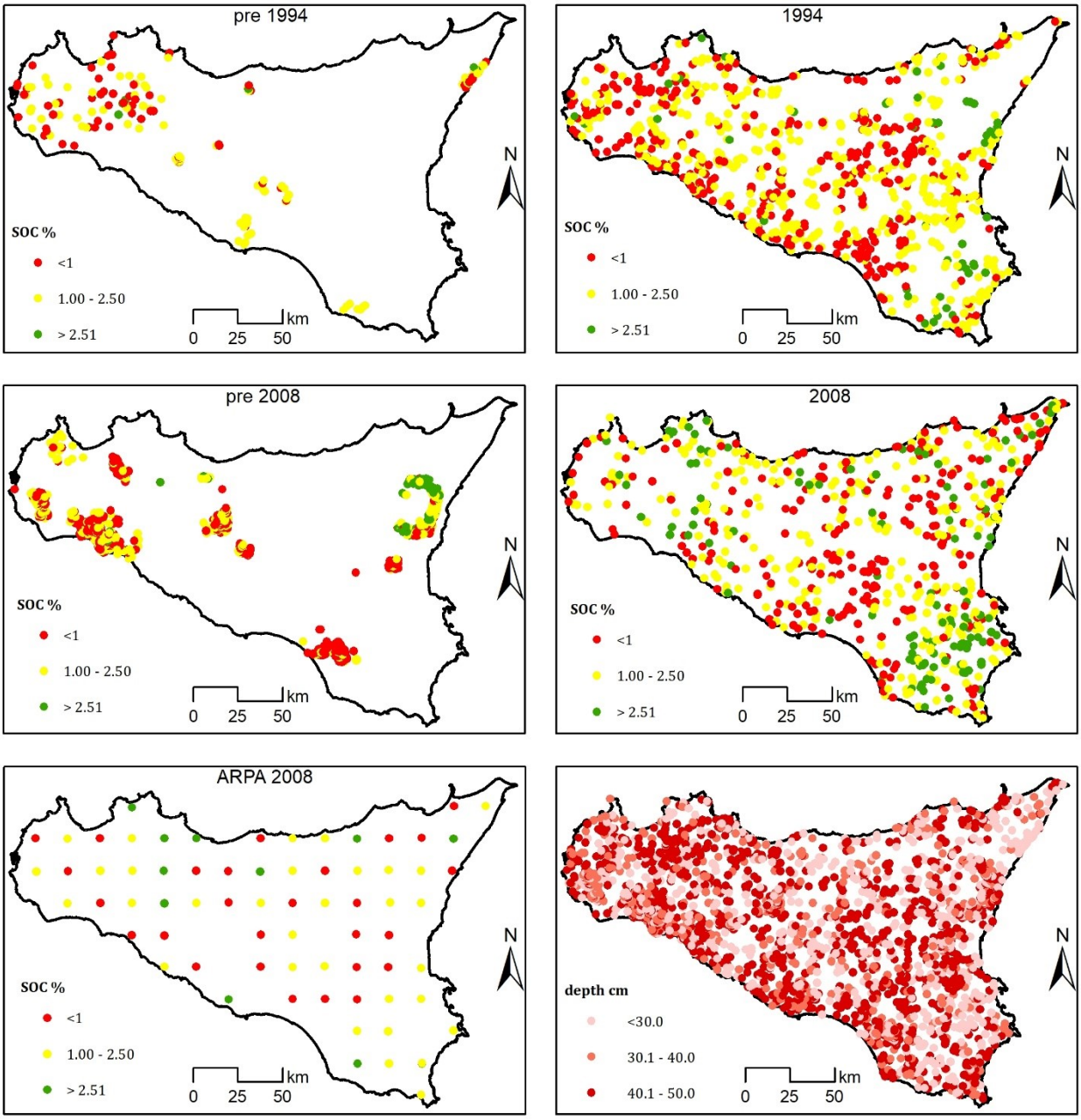


Fig. 4. *Distribution of the SOC samples categorized as TOPSOIL by sampling campaign indicated with year of sampling (a-e): pre1994 is for the sum of campaigns before the 1993-94; 1994 is for the 1993-94 campaign, pre 2008 is for the sum of campaigns between 1995 and 2007; 2008 is for the ARTA campaign in 2008, ARPA 2008 is for a campaign from the regional agency for the environmental protection (ARPA) conducted in the 2008. In the panel f) maps of the mean sampling depth is shown. See text for a correct definition of "TOPSOIL" samples.*

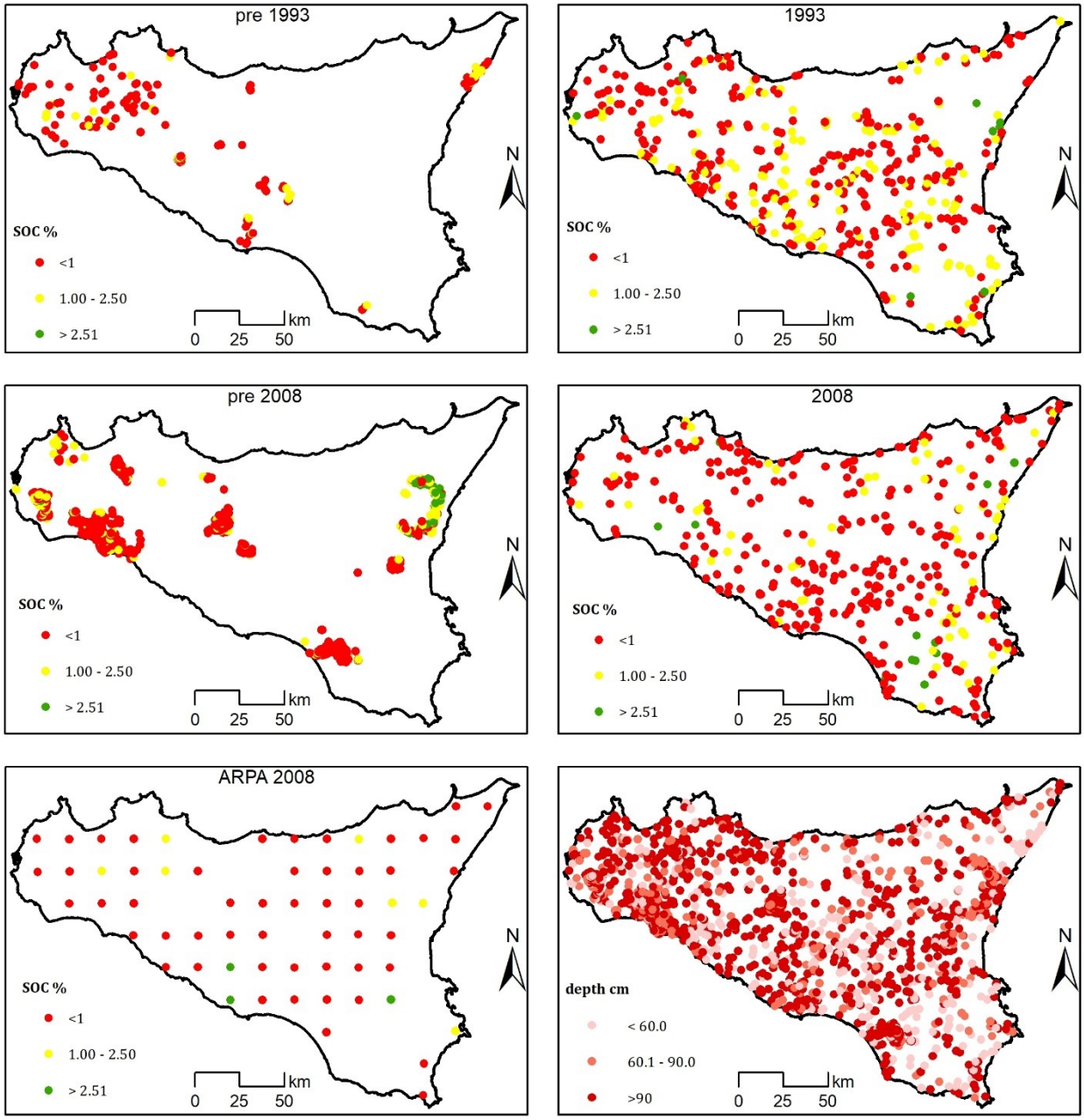


Fig. 5. *Distribution of the SOC samples categorized as SUBSOIL by sampling campaign indicated with year of sampling (a-e): pre1994 is for the sum of campaigns before the 1993-94; 1994 is for the 1993-94 campaign, pre 2008 is for the sum of campaigns between 1995 and 2007; 2008 is for the ARTA campaign in 2008, ARPA 2008 is for a campaign from the regional agency for the environmental protection (ARPA) conducted in the 2008. In the panel f) maps of the mean sampling depth is shown. See text for a correct definition of "SUBSOIL" samples*

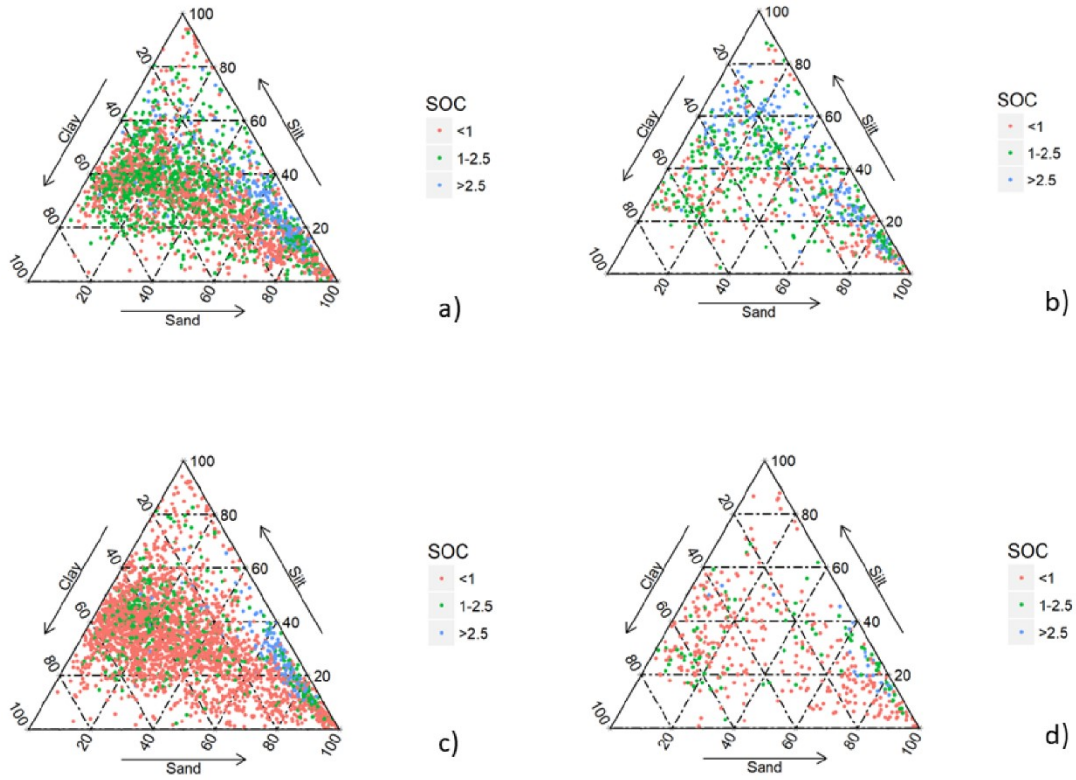


Fig. 6. *Distribution of samples by texture and SOC ranges in the TOPSOIL and SUBSOIL by each CORINE level 1 (agricultural or natural) land use. See text for a correct definition of "TOPSOIL" and "SUBSOIL".*

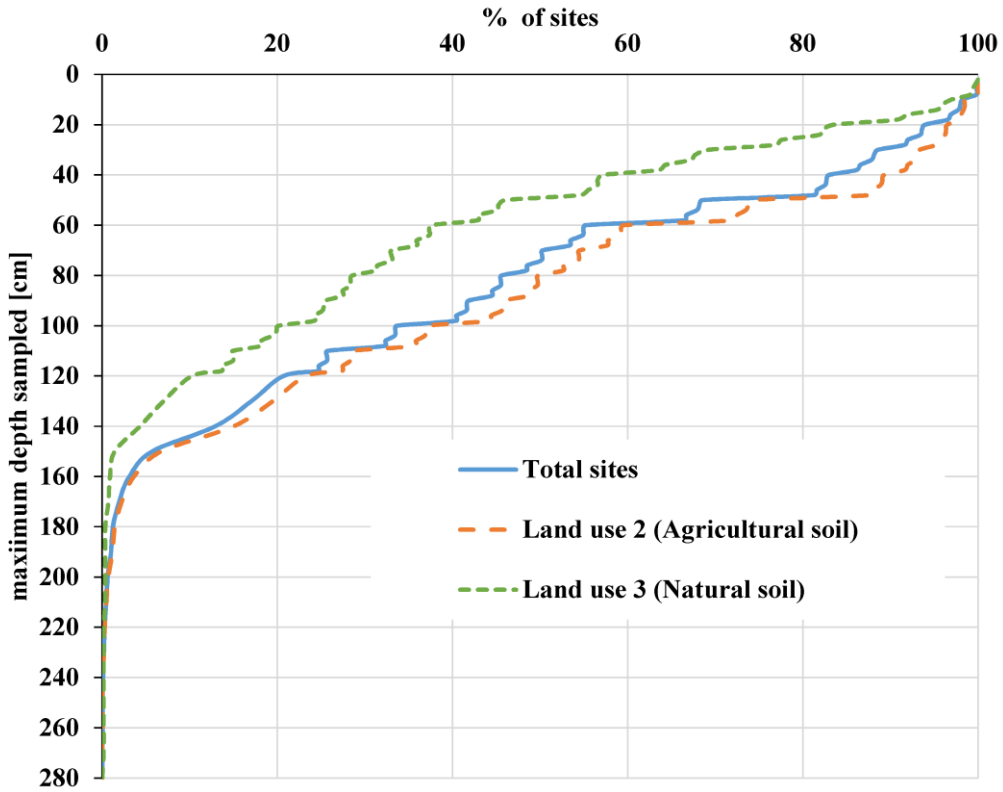


Fig. 7. *Maximum depth sampled by number (as percentage of the total or each fraction) of sites in each land use.*

3.3.3 The predictive ability of the corrected compared to that of the original dataset

When no covariate selection method was applied, the GLMSELECT retained 12 of the 13 covariates (Suppmat. Tab. 2). Adjusted R^2 values of the models built from the corrected (number of records used = 6674) and noncorrected (number of records used = 6985) datasets were 0.36 and 0.35, respectively (Tab. 4). Differences in the model performance statistics varied slightly by the effects of the correction (ranging from –9.4% to +9.6% depending on the statistics, see Tab. 4). Most of the standardized coefficients varied slightly at each step of the model building, with mean depth of the sample sturdily but constantly affecting the prediction at each step of the model building (Suppmat. Figs. 4 and 5). Notably, inclusion of the SRTM increased the contribution of CNBL to the SOC estimation.

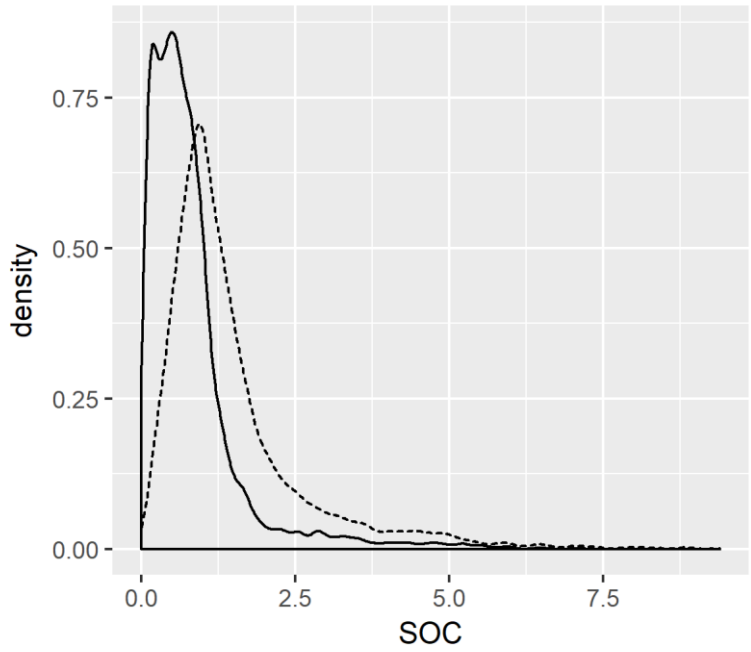


Fig. 8. *Density distribution of SOC (as % of the fine earth fraction) by TOPSOIL and SUBSOIL. See text for a correct definition of "TOPSOIL" and "SUBSOIL".*

When LASSO was applied, GLMSELECT retained 9 predictors and 7 predictor levels for the corrected and 9 predictors and 6 predictor levels for the non corrected datasets (Suppmat. Tab. 3 and Fig. 9). Adjusted R^2 values of models did not vary compared to the no variable selection method, but the F statistic increased by 54 and 80% in the corrected and uncorrected databases, respectively. After LASSO application, the contribution of each retained predictor changed. In particular, the contribution of depth and CNBL increased, whereas soil type was discarded and only some of its levels of coefficients retained (Fig. 9 and Suppmat. Figs. 6 and 7). In particular, the negative

association between SOC and mean depth of sampling and the positive association between average annual rainfall or CNBL and SOC were confirmed. Application of LASSO reduced the number of soil texture types relevant for the prediction, but in the optimized dataset, it highlighted a similar contribution by the percentage of sand in the fine earth or the presence of sandy soil. When procedures were run with splitting the database into the DIS50 and DID50 sub-databases, both the adjusted R^2 values of the models (Tab. 4) and contribution of depth to the modelling strongly decreased, with few differences by the sub-database. In each sub-database, the variable selection method had an effect on the performance statistic similar to that found in the complete database. However, DIS50 showed a lower ASE than both DID50 and the complete dataset (Suppmat. Fig. 8).

In each modelling process run for every sub-database (Suppmat. Figs. 9-12), CNBL frequently showed highly positive coefficients of association in either the corrected database or not and with the application of LASSO or not.

When no covariate selection was applied, clay and sand percentages in the fine earth fraction were strongly and negatively associated with SOC in DIS50 and with SRTM in DID50 (Suppmat. Figs. 9-10). When LASSO was applied, similar results were found. However, LASSO-models included fewer covariates than those with no variable selection for DIS50 but similar or more for DID50 (Suppmat. Figs. 11-12).

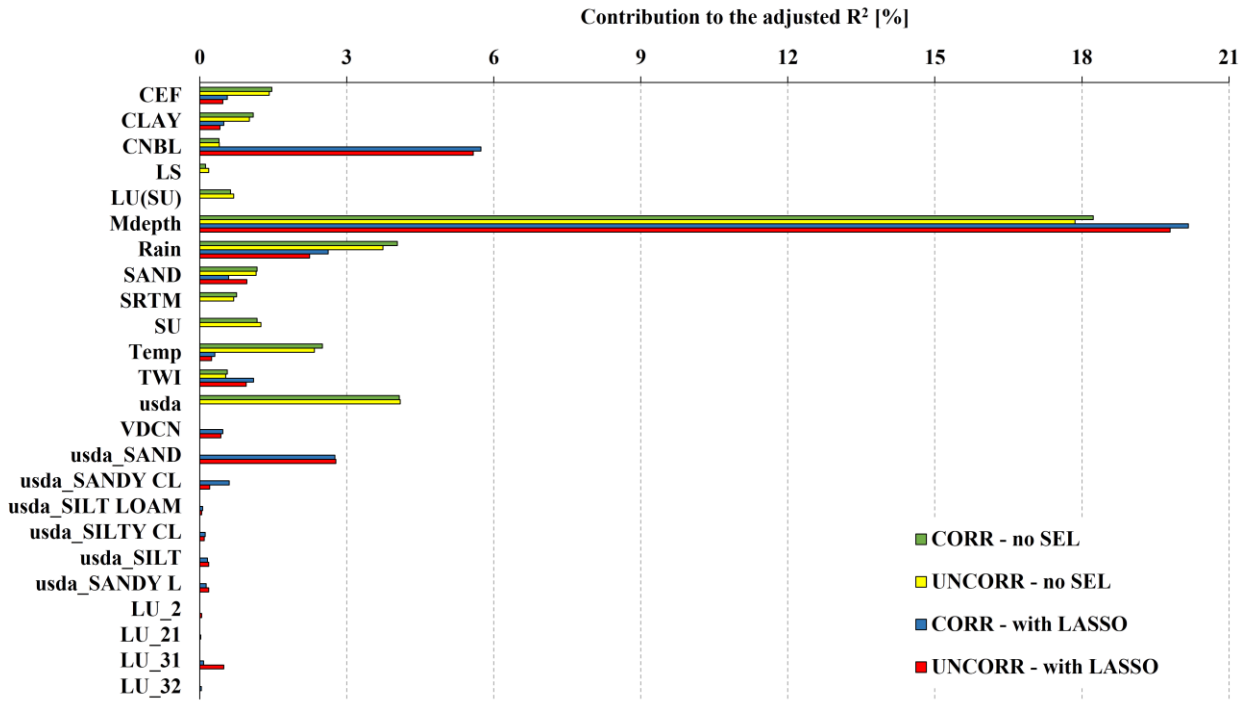


Fig. 9. *Contribution (expressed in percentage) of the inclusion of each predictor to the total adjusted R^2 of the models built with the optimized (+OPT) or non-optimized (unOPT) databases by applying a LASSO predictor selection strategy ("with LASSO") or not ("no SEL").*

3.4 Discussions

A wealth of studies are addressing the use of soil legacy data to estimate important soil traits, especially SOC (An et al., 2018; Dobos et al., 2010; Dong et al., 2018; Morvan et al., 2008; Odeh et al., 2012; Rial et al., 2017a; Suuster et al., 2012; Vaysse and Lagacherie, 2015). Such studies are important to fill data gaps in certain (fragile or hard to reach) areas, act as a guide for the harmonization of content-wide or worldwide sampling campaigns, and take the most from existing data from older times (Karunaratne et al., 2014).

In particular, soil legacy data can be uniquely adjusted for the space-time effects on soil properties (Dong et al., 2018) and jointly with newly computed or derived variables, can make important inferences on the quality of past and present soil representation (Teng et al., 2018). Such aspects fostered the creation of a soil legacy database at the global scale (Arrouays et al., 2017). In Europe, Morvan et al. (2008) indicated that the implementation of a monitoring system for the Mediterranean semiarid regions is urgent. In addition, the European monitoring network could benefit from local and country-scale sources of data, but this gain depends on the quality of the legacy data provided.

The procedure for the integration and error highlighting in this study was valid and improved goodness-to-fit, which was confirmed by the modelling procedures (either including or not a predictor selection method) and the marked reduction in fit statistics (such a reduction indicated a reduction in the model-to-data error). We retrieved relatively high SOC values in the quantiles >0.975, especially in the agricultural areas. In these areas, such an amount is likely high but it may still occur in irrigated lands under a conservative soil management system. These land uses were unfortunately not clearly recorded since the CORINE land cover (e.g., code 2.1 or 2.4) and not the actual standing cropland cover (e.g., wheat, tomato, peach orchards, or rosemary) was used; in addition, information on soil management was not provided. The discrepancy in land use attribution (data not shown) between the data in the present study and those used in previous experiments (Schillaci et al., 2017a) is because in the latter experiments, land cover was chosen by the CORINE land cover and not the land use actually observed in the field.

In the database used here, the sampling campaigns were mostly addressed to agricultural soils (bold yellow line and reddish fractions in Fig. 3), and only 3 surveys were mostly addressed to non agricultural land uses. Similarly, distribution of land use observed for each sample by year of sampling was mostly in agricultural areas. When comparing the share of area of each land use from the CORINE land cover maps and the percentages of sites in each land use from each sampling campaign, a discrepancy between datasets occurred. In particular, samples from the land cover forests and seminatural areas were often underrepresented in the sampling campaigns compared to the

CORINE, and samples from land uses arable land (2.1), permanent crops (2.2), pastures (2.3), and heterogeneous agricultural areas (2.4) were often unbalanced compared to the CORINE information.

Such discrepancy should be taken into account when modelling with remote-sensed or spatial data, since agricultural soils are often ploughed up to various cm depths (usually between 20 and 40 cm). Ploughing can homogenize and reduce some soil properties, including both the stable SOC fraction and BD. Such a homogenization can lead to an overestimation of some environmental properties, including SOC stock (Akpa et al., 2016; Chen et al., 2018b; Lee et al., 2009).

In the area under study, the low SOC content in the shallow (0-50 cm depth) layers of many cropland samples (CORINE code 2.1, mean SOC=1.2% of 842 data) could limit the potential exploitation of these soils for intensive agricultural production and further indicated a potential risk of degradation for these soils. At the same time, the low SOC indicated a high storage potential, as recently shown in a France national survey by Chen et al. (2018). Tree crops (2.2.1, 2.2.2, and 2.2.3, mean SOC= 1.5% of 1091 data), seminatural areas (2.3 and 2.4, mean SOC= 1.2% and 1.7%, respectively, of only 120 and 82 data) and woodland and forestry areas (CORINE code 3, mean SOC= 2.2% of 570 data) showed higher SOC content than that of arable lands. Arable lands in the region are mostly grown with winter-growing cereals (durum wheat and barley) and pulses (especially fava bean) with a fallow time-lapse of 6-7 months per year.

When aiming to fill gaps in the soil BD within databases, the application of a threshold of the maximum depth of information from each sample in this study agrees with the findings by Sequeira et al. (2014). Such thresholds, jointly with mean sample depth and sample thickness, can help in the modelling processes of BD and thus for SOC stock, since BD estimation is a major limit to the correctness of SOC stock computation (Poeplau et al., 2017). Indeed, we found that splitting the databases by maximum depth of information into two sub-databases dramatically reduced the contribution of depth to the modelling process and increased those of topographic and morphometric traits, including CNBL and SRTM, mean rainfall and temperature, and soil type.

The procedure for the integration and error highlighting used in this study can thus help in the harmonization of data in wider databases (Ribeiro et al., 2015) and their merging with databases built at different scales (Bardy et al., 2018; Wiesmeier et al., 2019). Last, the procedure can remove the noises that hamper the study of the temporal by depth dimension and achieve a clearer view of the land use/soil effects on SOC stocks at various scales, as also suggested by other authors for similar conditions (Di Bene et al., 2016; Francaviglia et al., 2017; Rabbi et al., 2016).

3.5 Conclusion

The analysis carried on in this paper looked at the real needs of a monitoring network that will allow better planning of agriculture and environmental management at a regional and sub-regional extent. A wider usage of the SMSLD in environmental modelling and DSM for national and international projects is expected. From the description of the data collection, we obtained useful information about the limits of a sampling campaign organization and the purpose of a sampling. The supplemental information strata created from our analysis fixed some important shortcomings in the analytical values of soil texture, coordinate precision and land cover, and this approach should be used in further processes of bulk density modelling. The coordinate precision was assessed by a new field in the database that allowed the user to measure the reliability of the precision when aiming to derive data from other sources (e.g., the CORINE land cover). The land cover of the sampled data was provided at the most detailed CORINE level when available and for all the data, was provided up to the second level. These pre-processing procedures allowed the compilation of a solid database from which some physical properties have been used together with topographical indices to fit an SOC model. Last, the application of the modelling processes to the complete database or the two sub-databases differing per deepest information highlighted that dominant covariates changed with a change in the sample pools considered, and such changes should be carefully taken into account when planning new sampling campaigns.

Thus, future perspectives include the following:

- to move towards better methodology of capturing (historical analysis) and sampling (renewing) the legacy data,
- to study in detail the temporal by depth dimension of the SMSLD,
- to highlight the statistical correction terms to apply for unbalanced data,
- to explore the modelling process at varying key aspects of the data considered that include proximity, density, layer, subdomains of environmental variables (such as rainfall or temperatures), etc.,
- to foster land management at the administrative level (e.g., by province),
- to suggest a data sharing agreement form at the involved institutions.

Data harmonization supplemental materials

Correction of likely wrong samples of coarse fraction (diameter higher than 2 mm). This step was done by checking for samples with coarse fraction higher than 80% and at the same time sand lower than 80%, and SOC>1% (by means of the "IF" tool in Excel);

This procedure allowed the univocal identification of doubled records, which were ordered by the internal code. Such code was then used to create a list of doubled records. The creation of such list was made by identifying, in the pivot system, those internal codes which were not aligned to the coordinate identification were selected and used as an ordination table of the internal code in the database by means of the combination of the tools "VLOOKUP" and "FILTER". Briefly, a table of the doubled lines was made and ordered from the lower to the higher number. Then an additional column was created in the database. In this column, we applied the "VLOOKUP" to the total internal codes. Those internal codes corresponding to a doubled code were indicated by 1, else they were indicated by 0. Then we selected the "1"-indicated records by the "filter", which were copied in another sheet and deleted from those of the main database.

Metadata

In its original form, the dataset had no metadata reporting instruction to how to use the data, how and why the data were collected, and which was their reliability in terms of coordinate precision. Every time that we performed a search for mistakes, outliers or correction (e.g. texture sum up to 100% or duplicate removal) a new field in the SMSLD was added with information on the change. Deleted lines were copied in a new sheet. We also collected several auxiliary data that were projected in UTM33N (http://spatialreference.org/ref/epsg/wgs-84-utm-zone-33n/). These data include spatial extent, geographic projection, data owner, accessibility.

The data from the various sampling campaigns provided by ARTA and included in the legacy SMSLD had the same spatial reference system.

Covariates

Supplementary Material Tab. 1. Predictors used to build the models to test the database optimization procedure.

Trait (Acronym)	reference
Land Use (LU)	Regional Bureau of Agricuture, Sicly, Italy
Texture (USDA)	
Mean depth of the sampling (Mdepth)	
Coarse earth fraction % (CEF)	
Soil type	
Clay Content (CLAY, particle size <0.002 mm)	

Sand content (SAND, particle size from 0.02 mm to 2 mm)	
Mean Annual Temperature (Temp)	Fick & Hijmans, 2017. Int. J. Climatol. 37, 4302–4315. doi:10.1002/joc.5086
Mean Annual Rainfall (Rain)	
Slope aspect (asp)	Zevenbergen, & Thorne, 1987. Earth Surf. Process. Landforms 12, 47–56. doi:10.1002/esp.3290120107
Slope length factor (LS)	
Channel network base level (CNBL)	Conrad et al., 2015. Geosci. Model Dev. 8, 1991–2007. doi:10.5194/gmd-8-1991-2015
Vertical distance channel network (VDCN)	
Digital elevation model (SRTM)	Farr et al., 2007. Rev. Geophys. 45, RG2004. doi:10.1029/2005RG000183
Topographic wetness index (TWI)	Moore et al., 1991. Hydrol. Process. 5, 3–30. doi:DOI: 10.1002/hyp.3360050103

Correction and error highlighting

We checked the negative depth reported. E.g. in a sample, the first layer reported the depth from 0 to 2 cm and the deeper 3 layer steps of 25 cm thickness, i.e. from 25 to 50 cm, from 50 to 75 cm and from 75 to 100 cm. Thus 2 was changed to 25.

In addition, we created two additional information fields of the beginning of the soil layers within each sampling size with three classes (Supplementary Material Fig. 2):

- '0' for layers starting from 0 cm (i.e. soil surface) containing 2807 observations of SOC and 977 of BD;
- "SUP" for layers starting from 0.1 to 9.9 cm (i.e. close to the soil surface) and containing 61 observations of SOC and 16 of BD. Such stratum was due to an objective hardness in assuming that these soils are or not directly influenced by the atmosphere. Among these samples (corresponding to 51 sites with SOC measures and 14 of BD), 18 sites for SOC and 3 of BD showed the upper layer sampled not beginning from the soil surface;
- "INF" for layers starting from deeper than 10.0 cm (i.e. 'far' from the soil surface) and containing 3809 observations of SOC and 433 of BD. The coupling of these two latter information strata give a rapid information on the layer positioning in the profile.

Lines deleted were:

- 8 records lacking sand and silt information, since we hypothesized clay was not reliable. From the rest of data, sum of clay+silt+sand percentages was checked to be comprised between 97.5 and 102.5%. No data were out of these boundaries. Nonetheless, original data were harmonised to 100% to allow for statistical inference assuming that sum of fine earth fraction is, in theory, a constant value;

- 4 records of litter (i.e. with upper soil limit lower than 0). This was due since the scarce numerousness would not allow for any modelling approach and strongly and negatively affect the computations on the mineral soils;
- 11 records with observed CORINE Land use =1 (Artificial surface) for the same reasons above and because relationships between predictors and target variable can't be found in artificial, non-systematically made, soils;
- 294 doubled records.

Modification of the coarse fraction of 3 samples (all from Land use=2) was done as follows. Samples derived from 3 different sampling campaigns, they were sampled in layers 0-20, 20-55, and 25-50 cm depth, and showed SOC content of 1.8, 1.51 and 2.66%, respectively. The second and third samples were coming from close sites (distant around 30 km). Lastly, one sample (internal id: 5547) reported as a land use the CORINE code 344, which does not exist in the CORINE legend. After checking in the orthophotos, the indication was changed to 244.

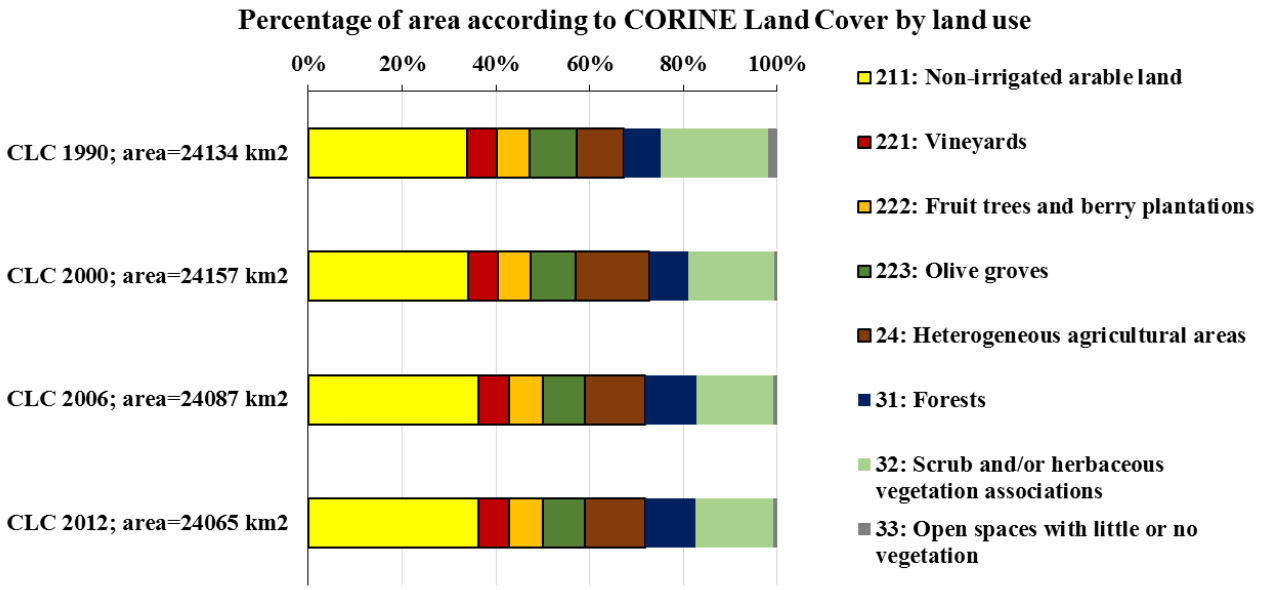
Profundity of information about land use and coordinates reported was in general high. Only 58 sites with SOC information and 35 with BD information reported only the level one of the CORINE information (i.e. agricultural land use, with no further specification on which agricultural land use; tab. 2). The rest of the sampling sites reported highly detailed information on the land use.

CORINE land cover maps of 1990, 2000, 2006, and 2012 (total area covered in land uses 2 and 3: 24134 km^2, 24157 km^2, 24087, and 24065 km^2, respectively) was 67-73% in land use 2 (Agricultural areas) and 27-33% in the land use 3 (Forest and semi natural areas). In the whole dataset, 82% of SOC samples were coming from soils categorised in the land use 2.

With regards to sampling depth, 80% of sites sampled in land use 2 (agricultural soils) gave an information up to 50 cm, whereas in land use 3 (natural soils), information on a 50 cm depth was available for less than 50% of sites sampled. Information on SOC concentration at 100 cm was available for only 40% and 20% of sites with land use 2 and 3, respectively.

Such correction made us to discard 317 data, the most of which due to duplication and only 23 for not being similar to others. These latter data were discarded due to the extremely low number of cases compared to the core of data (i.e. mineral soil).

Supplementary Material Fig. 3. Area share in by CORINE per land use in Sicily (Italy) from 1990 to 2012, and area covered by each map in the land uses indicated.



Acknowledgments

The authors would like to thank M.G Matranga, V. Ferraro and A. Guaitoli from the Regional Bureau for Agriculture, Rural Development and Mediterranean Fishery, the Department of Agriculture, Palermo- The authors thanks also Grant Campbell from Cranfield University for the proofreading and two anonymous reviewers that helped us improving the quality of the manuscript.

Supplementary Material Tab. 2. Fit-statistics for the models with **no covariate selection** built with the CORRECTED and UNCORRECTED (i.e., original) databases. See Supplementary Material Tab. 1 for codes. Procedure was run with the complete database, including samples from all layers. * Optimal Value Of Criterion

OPTIMIZED DATABASE

No covariate selection summary													
Step	Effect	Number **of parameters in the model**	R-Square	Adjusted **R-Square**	AIC	AICC	BIC	CP	SBC	PRESS	ASE	F **Value**	Pr > **F**
0	Intercept	1	0	0	-5549.41	-5549.408	-12225.4	3826.6994	-12219.6	1068.9719	0.1601	0	1
1	LU	8	0.019	0.018	-5663.532	-5663.505	-12344.41	3641.0862	-12286.08	1050.9358	0.1571	18.46	<.0001
2	usda	19	0.0613	0.0587	-5935.456	-5935.33	-12622.93	3219.3773	-12483.14	1009.263	0.1503	27.24	<.0001
3	Mdepth	20	0.2433	0.2411	-7371.921	-7371.782	-14053.37	1310.6003	-13912.8	813.8111	0.1212	1600.41	<.0001
4	CEF	21	0.258	0.2558	-7501.673	-7501.52	-14182.73	1157.3073	-14035.74	798.2898	0.1188	132.64	<.0001
5	CLAY	22	0.269	0.2667	-7599.032	-7598.866	-14279.8	1044.2046	-14126.3	786.7111	0.117	99.77	<.0001
6	SAND	23	0.2808	0.2784	-7705.251	-7705.07	-14385.62	922.7736	-14225.71	774.3004	0.1151	108.72	<.0001

7	Temp	24	0.3058	0.3034	-7939.989	-7939.794	-14619.05	661.6328	-14453.64	747.8519	0.1111	240.12	<.0001
8	Rain	25	0.3461	0.3438	-8337.304	-8337.093	-15013.69	240.3948	-14844.15	704.4477	0.1047	409.96	<.0001
9	asp	26	0.3486	0.3462	-8360.968	-8360.74	-15037.22	216.0501	-14861.01	701.9702	0.1043	25.61	<.0001
10	CNBL	27	0.3525	0.35	-8398.582	-8398.337	-15074.56	177.5827	-14891.82	698.074	0.1037	39.57	<.0001
11	LS	28	0.3537	0.3511	-8408.95	-8408.688	-15084.87	166.9974	-14895.38	697.0618	0.1035	12.33	0.0004
12	SRTM	29	0.3614	0.3587	-8486.81	-8486.53	-15162.07	88.2927	-14966.43	689.0079	0.1022	79.99	<.0001
13	TWI	30	0.3671	0.3644*	-8545.1018*	-8544.8031*	-15219.831*	30.0000*	-15017.918*	682.9866*	0.1013	60.29	<.0001
14	VDCN	30	0.3671	0.3644	-8545.102	-8544.803	-15219.83	30	-15017.92	682.9866	0.1013	.	.

*** Optimal Value Of Criterion**

UNCORRECTED DATABASE													
No covariate selection summary													
Step	Effect	Number **of parameters in the model**	R-Square	Adjusted **R-Square**	AIC	AICC	BIC	CP	SBC	PRESS	ASE	F **Value**	Pr > **F**
0	Intercept	1	0	0	-6013.664	-6013.663	-12999.63	3850.2569	-12993.81	1086.0409	0.1554	0	1
1	LU	11	0.0208	0.0194	-6140.231	-6140.186	-13132.98	3645.3105	-13051.86	1065.851	0.1522	14.79	<.0001
2	usda	22	0.0631	0.0603	-6426.947	-6426.788	-13425.86	3208.6649	-13263.21	1023.3605	0.1456	28.6	<.0001
3	Mdepth	23	0.2413	0.2389	-7898.473	-7898.3	-14890.43	1280.2995	-14727.89	828.9715	0.1179	1635.08	<.0001
4	CEF	24	0.2555	0.2531	-8028.902	-8028.715	-15020.37	1127.9372	-14851.47	813.7565	0.1157	133.23	<.0001
5	CLAY	25	0.2657	0.2632	-8123.301	-8123.099	-15114.42	1019.3982	-14939.01	802.789	0.1141	96.72	<.0001
6	SAND	26	0.2773	0.2747	-8232.545	-8232.328	-15223.17	895.7186	-15041.41	790.3589	0.1123	111.72	<.0001
7	Temp	27	0.3007	0.2981	-8459.833	-8459.6	-15449.06	644.9068	-15261.84	765.3471	0.1087	232.19	<.0001
8	Rain	28	0.338	0.3355	-8841.179	-8840.929	-15827.64	242.3351	-15636.34	724.4973	0.1029	392.48	<.0001

9	asp	29	0.3402	0.3376	-8862.479	-8862.211	-15848.8	220.4534	-15650.79	722.3152	0.1026	23.24	<.0001
10	CNBL	30	0.3442	0.3415	-8902.536	-8902.25	-15888.55	179.5477	-15683.99	718.2312	0.1019	42	<.0001
11	LS	31	0.3459	0.3431	-8918.984	-8918.68	-15904.87	162.8085	-15693.59	716.6174	0.1017	18.39	<.0001
12	SRTM	32	0.353	0.3501	-8992.945	-8992.623	-15978.17	88.1681	-15760.7	709.1018	0.1006	76.03	<.0001
13	TWI	33	0.3583	0.3553*	-9048.1499*	-9047.8074*	-16032.837*	33.0000*	-15809.050*	703.4984*	0.0997	57.17	<.0001
14	VDCN	33	0.3583	0.3553	-9048.15	-9047.807	-16032.84	33	-15809.05	703.4984	0.0997	.	.
*** Optimal Value Of Criterion**													

Supplementary Material Tab. 3. Fit-statistics for the models with LASSO selection procedure built with the CORRECTED and UNCORRECTED (i.e., original) databases. See Supplementary Material Tab. 1 for codes. Procedure was run with the complete database, including samples from all layers. * Optimal Value Of Criterion

OPTIMIZED DATABASE

LASSO Selection Summary											
Step	Effect	Model **R-Square**	Adjusted **R-Square**	AIC	AICC	BIC	CP	SBC	ASE	F **Value**	Pr > **F**
0	Intercept	0	0	-5549.4098	-5549.408	-12225.404	3826.6994	-12219.604	0.1601	0	1
1	Mdepth	0.1344	0.1342	-6510.5863	-6510.5827	-13186.79	2417.8694	-13173.974	0.1386	1035.82	<.0001
2	CNBL	0.1483	0.148	-6616.8487	-6616.8427	-13293.499	2273.6454	-13273.43	0.1364	109.1	<.0001
3	Rain	0.2463	0.246	-7430.7425	-7430.7335	-14107.044	1246.7374	-14080.518	0.1207	867.33	<.0001
4	usda_**SAND**	0.2827	0.2823	-7759.3264	-7759.3138	-14435.489	866.3513	-14402.296	0.1148	338.65	<.0001
5	CEF	0.2856	0.285	-7784.0845	-7784.0677	-14460.43	838.2208	-14420.248	0.1144	26.79	<.0001
6	usda_**SANDY CLAY LOAM**	0.2922	0.2916	-7844.226	-7844.2044	-14520.672	770.7115	-14473.583	0.1133	62.37	<.0001
7	SAND	0.2956	0.2949	-7874.8352	-7874.8082	-14551.413	736.4942	-14497.386	0.1128	32.65	<.0001
8	TWI	0.3048	0.3039	-7959.8128	-7959.7797	-14636.368	642.7535	-14575.558	0.1113	87.43	<.0001
9	asp	0.3069	0.3059	-7978.1925	-7978.1529	-14654.862	622.5003	-14587.131	0.111	20.38	<.0001
10	Temp	0.3281	0.3271	-8184.1437	-8184.0968	-14860.332	401.2742	-14786.276	0.1076	210.88	<.0001
11	usda_SILT **LOAM**	0.3306	0.3295*	-8206.7713*	-8206.7167*	-14882.983*	377.2951*	-14802.098*	0.1072	24.63	<.0001

*** Optimal Value Of Criterion**											
UNCORRECTED **DATABASE**											
LASSO **Selection Summary**											
Step	Effect	Model **R-Square**	Adjusted **R-Square**	AIC	AICC	BIC	CP	SBC	ASE	F **Value**	Pr > **F**
0	Intercept	0	0	-6013.6643	-6013.6625	-12999.627	3850.2569	-12993.813	0.1554	0	1
1	Mdepth	0.1325	0.1324	-7004.4979	-7004.4945	-13990.657	2416.8674	-13977.795	0.1348	1066.55	<.0001
2	CNBL	0.159	0.1587	-7219.0684	-7219.0627	-14205.577	2131.9565	-14185.514	0.1307	219.87	<.0001
3	Rain	0.2369	0.2366	-7896.4696	-7896.461	-14882.758	1289.5063	-14856.063	0.1186	713.13	<.0001
4	usda_**SAND**	0.278	0.2776	-8280.9251	-8280.913	-15267.018	846.5714	-15233.667	0.1122	397.06	<.0001
5	CEF	0.2791	0.2786*	-8289.6479*	-8289.6318*	-15275.937*	836.5736*	-15235.539*	0.1121	10.72	0.0011
*** Optimal Value Of Criterion**											

Supplementary Material Tab. 4. Fit-statistics for the models with **no covariate selection** built with the CORRECTED and UNCORRECTED (i.e., original) databases. See Supplementary Material Tab. 1 for codes. Procedure were run only on the database pertaining samples for which the deepest information was shallower than 50 cm (named "**DIS50**") or only on the database pertaining samples for which the deepest information was deeper than 50 cm (named "**DID50**").
* Optimal Value Of Criterion

CORRECTED DATABASE – **DIS50 - No covariate selection** summary

Step	Effect	Number of parameters in the model	Model R-Square	Adjusted R-Square	AIC	AICC	BIC	CP	SBC	PRESS	ASE	F Value	Pr > F
0	**Intercept**	1	0	0	-4277.87	-4277.87	-7576.56	1230.9534	-7571.77	331.4094	0.1004	0	1
1	**SU**	2	0.0269	0.0266	-4365.96	-4365.95	-7665.09	1110.974	-7653.75	322.7398	0.0977	91.27	<.0001
2	**LU(SU)**	8	0.0353	0.0332	-4382.39	-4382.34	-7684.43	1085.1541	-7633.59	321.1734	0.0969	4.75	<.0001
3	**usda**	19	0.0973	0.0924	-4579.5	-4579.25	-7884.91	826.4391	-7763.58	302.534	0.0907	20.48	<.0001
4	**Mdepth**	20	0.1284	0.1233	-4692.95	-4692.66	-7997.57	687.8753	-7870.92	292.3174	0.0875	116.77	<.0001
5	**CEF**	21	0.1566	0.1515	-4799.73	-4799.42	-8103.48	561.8479	-7971.6	283.0856	0.0847	109.89	<.0001
6	**CLAY**	22	0.1732	0.1679	-4863.1	-4862.76	-8166.36	488.9146	-8028.88	277.6901	0.083	65.59	<.0001
7	**SAND**	23	0.1935	0.1881	-4943.22	-4942.85	-8245.68	398.8723	-8102.89	271.0433	0.081	82.57	<.0001
8	**Temp**	24	0.2281	0.2226	-5085.58	-5085.18	-8386.29	244.5061	-8239.16	259.7934	0.0775	146.5	<.0001
9	**Rain**	25	0.2665	0.2611	-5252	-5251.57	-8550.33	72.5381	-8399.47	246.9913	0.0737	171.48	<.0001
10	**CNBL**	26	0.2691	0.2635	-5261.65	-5261.19	-8559.82	62.8261	-8403.02	246.3863	0.0734	11.58	0.0007
11	**LS**	27	0.2696	0.2638	-5261.82	-5261.33	-8559.96	62.6452	-8397.09	246.4255	0.0734	2.16	0.142
12	**SRTM**	28	0.2757	0.2697	-5287.49	-5286.95	-8585.16	37.0261	-8416.66	244.5805	0.0727	27.54	<.0001

13	**TWI**	29	0.2779	0.2717*	-5295.5848*	-5295.0155*	-8593.0704*	29.0000*	-8418.6537*	243.9786*	0.0725	10.03	0.0016
14	**VDCN**	29	0.2779	0.2717	-5295.58	-5295.02	-8593.07	29	-8418.65	243.9786	0.0725	.	.

105

106 UNCORRECTED DATABASE – **DIS50 - No covariate selection** summary

Step	Effect	Number of parameters in the model	Model R-Square	Adjusted R-Square	AIC	AICC	BIC	CP	SBC	PRESS	ASE	F Value	Pr > F
0	**Intercept**	1	0	0	-4544.82	-4544.81	-7990.49	1239.2033	-7985.67	338.6097	0.0982	0	1
1	**SU**	3	0.0255	0.025	-4629.92	-4629.9	-8076.51	1123.6601	-8058.48	330.3234	0.0957	45.09	<.0001
2	**LU(SU)**	11	0.0342	0.0314	-4644.87	-4644.78	-8095.22	1098.8456	-8024.28	328.5986	0.0949	3.87	0.0001
3	**usda**	22	0.0965	0.091	-4852.47	-4852.15	-8305.71	829.2991	-8164.29	309.4579	0.0888	21.45	<.0001
4	**Mdepth**	23	0.1288	0.1232	-4975.86	-4975.5	-8428.05	680.1032	-8281.53	298.5522	0.0856	126.84	<.0001
5	**CEF**	24	0.1569	0.1512	-5086.63	-5086.25	-8537.77	550.7348	-8386.16	289.1547	0.0828	113.84	<.0001
6	**CLAY**	25	0.1721	0.1662	-5147.24	-5146.83	-8597.84	481.6282	-8440.63	284.0626	0.0813	62.73	<.0001
7	**SAND**	26	0.1935	0.1876	-5235.8	-5235.36	-8685.37	383.0572	-8523.04	276.8859	0.0792	91.06	<.0001
8	**Temp**	27	0.2256	0.2197	-5373.37	-5372.9	-8821.08	235.1267	-8654.47	266.2287	0.0761	141.33	<.0001
9	**Rain**	28	0.2606	0.2547	-5530.79	-5530.28	-8976.07	73.1507	-8805.74	254.3212	0.0726	161.84	<.0001
10	**CNBL**	29	0.2632	0.2572	-5541.25	-5540.7	-8986.33	62.6548	-8810.05	253.6648	0.0724	12.37	0.0004
11	**LS**	30	0.264	0.2577	-5542.59	-5542.01	-8987.62	61.3054	-8805.25	253.6213	0.0723	3.32	0.0686
12	**SRTM**	31	0.2695	0.263	-5566.49	-5565.87	-9011.05	37.4969	-8823.0046*	251.9371	0.0718	25.76	<.0001
13	**TWI**	32	0.2711	0.2644*	-5572.0485*	-5571.3906*	-9016.4486*	32.0000*	-8822.42	251.5345*	0.0716	7.5	0.0062
14	**VDCN**	32	0.2711	0.2644	-5572.05	-5571.39	-9016.45	32	-8822.42	251.5345	0.0716	.	.

107 CORRECTED DATABASE – **DID50 - No covariate selection** summary

Step	Effect	Number of parameters in the model	Model R-Square	Adjusted R-Square	AIC	AICC	BIC	CP	SBC	PRESS	ASE	F Value	Pr > F
0	**Intercept**	1	0	0	-2703.99	-2703.98	-6078.66	1216.8362	-6073.86	556.5978	0.1649	0	1
1	**SU**	2	0.0005	0.0002	-2703.52	-2703.51	-6078.72	1216.753	-6067.27	556.7006	0.1648	1.53	0.216
2	**LU(SU)**	8	0.0149	0.0129	-2740.7	-2740.65	-6118.89	1162.3759	-6067.71	550.5151	0.1624	8.24	<.0001
3	**usda**	19	0.0541	0.049	-2855.54	-2855.29	-6238.1	1004.7094	-6115.19	532.5227	0.156	12.62	<.0001
4	**Mdepth**	20	0.1192	0.1143	-3094.4	-3094.13	-6475.07	707.6396	-6347.93	496.0846	0.1452	248.19	<.0001
5	**CEF**	21	0.1316	0.1264	-3140	-3139.7	-6520.49	653.0265	-6387.4	489.6042	0.1432	47.63	<.0001
6	**CLAY**	22	0.1413	0.136	-3176.1	-3175.77	-6556.46	610.2796	-6417.37	484.3479	0.1416	38.07	<.0001
7	**SAND**	23	0.1516	0.1461	-3214.79	-3214.43	-6594.96	565.0467	-6449.94	478.8329	0.1399	40.66	<.0001
8	**Temp**	24	0.1776	0.172	-3317.8	-3317.41	-6696.91	447.7485	-6546.83	464.8916	0.1356	105.9	<.0001
9	**Rain**	25	0.2302	0.2247	-3538.78	-3538.36	-6915.03	208.4062	-6761.69	434.9704	0.1269	228.81	<.0001
10	**CNBL**	26	0.2375	0.2318	-3569.01	-3568.56	-6944.87	176.8249	-6785.79	431.1158	0.1257	32.13	<.0001
11	**LS**	27	0.2435	0.2376	-3593.53	-3593.05	-6969.05	151.4322	-6804.19	428.0602	0.1247	26.41	<.0001
12	**SRTM**	28	0.2606	0.2546	-3668.67	-3668.15	-7042.98	74.9718	-6873.2	418.7657	0.1219	77.37	<.0001
13	**TWI**	29	0.2711	0.2650*	-3714.7096*	-3714.1532*	-7088.2069*	29.0000*	-6913.1178*	413.0039*	0.1202	47.97	<.0001
14	**VDCN**	29	0.2711	0.265	-3714.71	-3714.15	-7088.21	29	-6913.12	413.0039	0.1202	.	.

108

UNCORRECTED DATABASE – **DID50 - No covariate selection** summary

Step	Effect	Number of parameters in the model	Model R-Square	Adjusted R-Square	AIC	AICC	BIC	CP	SBC	PRESS	ASE	F Value	Pr > F
0	**Intercept**	1	0	0	-2944.82	-2944.81	-6485.47	1230.2704	-6480.65	566.4834	0.1599	0	1
1	**SU**	3	0.0002	-0.0004	-2941.57	-2941.56	-6483.25	1233.2536	-6465.06	566.9292	0.1599	0.38	0.6858
2	**LU(SU)**	10	0.0224	0.0199	-3006.86	-3006.78	-6551.78	1141.6703	-6487.14	555.9497	0.1564	11.42	<.0001
3	**usda**	21	0.0602	0.0549	-3124.79	-3124.5	-6673.7	982.996	-6537.18	538.2799	0.1503	12.9	<.0001
4	**Mdepth**	22	0.1198	0.1145	-3354.45	-3354.14	-6901.39	701.1431	-6760.67	504.4701	0.1408	237.92	<.0001
5	**CEF**	23	0.1313	0.1259	-3399.14	-3398.8	-6945.87	648.1448	-6799.19	498.3236	0.1389	46.7	<.0001
6	**CLAY**	24	0.14	0.1344	-3432.93	-3432.56	-6979.53	608.4853	-6826.8	493.5371	0.1375	35.72	<.0001
7	**SAND**	25	0.1495	0.1437	-3470.03	-3469.63	-7016.42	565.4404	-6857.73	488.4024	0.136	39.04	<.0001
8	**Temp**	26	0.1735	0.1676	-3569.48	-3569.05	-7114.79	452.8656	-6951.01	475.3851	0.1322	102.16	<.0001
9	**Rain**	27	0.2225	0.2167	-3783.52	-3783.06	-7325.97	221.5522	-7158.88	447.0694	0.1244	221.07	<.0001
10	**CNBL**	28	0.2299	0.224	-3815.62	-3815.12	-7357.64	188.0131	-7184.81	443.1276	0.1232	33.99	<.0001
11	**LS**	29	0.2377	0.2316	-3849.47	-3848.94	-7390.99	153.0161	-7212.48	438.9598	0.1219	35.73	<.0001
12	**SRTM**	30	0.2537	0.2476	-3922.83	-3922.26	-7463.14	78.4502	-7279.67	430.1122	0.1194	75.52	<.0001
13	**TWI**	31	0.2641	0.2578*	-3970.3677*	-3969.7655*	-7509.8201*	31.0000*	-7321.0393*	424.2663*	0.1177	49.45	<.0001
14	**VDCN**	31	0.2641	0.2578	-3970.37	-3969.77	-7509.82	31	-7321.04	424.2663	0.1177	.	.

-

Supplementary Material Tab. 5. Fit-statistics for the models with LASSO selection procedure built with the CORRECTED and UNCORRECTED (i.e., original) databases. See Supplementary Material Tab. 1 for codes. Procedure were run only on the database pertaining samples for which the deepest information was shallower than 50 cm (named "**DIS50**") or only on the database pertaining samples for which the deepest information was deeper than 50 cm (named "**DID50**"). * Optimal Value Of Criterion

CORRECTED DATABASE – **DIS50** - LASSO selection summary

Step	Effect	Model R-Square	Adjusted R-Square	AIC	AICC	BIC	CP	SBC	PRESS	ASE	F Value	Pr > F
0	**Intercept**	0	0	-4277.8713	-4277.8676	-7576.5616	1230.9534	-7571.7702	331.4094	0.1	0	1
1	**Rain**	0.0761	0.0758	-4536.9092	-4536.9019	-7835.8445	888.4567	-7824.707	306.4676	0.093	271.48	<.0001
2	**SRTM**	0.1106	0.1101	-4660.4486	-4660.4365	-7959.5993	734.2422	-7942.1454	295.216	0.089	127.84	<.0001
3	**CNBL**	0.1118	0.111	-4662.7572	-4662.7389	-7962.2603	730.9857	-7938.3529	295.0963	0.089	4.31	0.0381
4	**Mdepth**	0.1523	0.1513	-4814.7229	-4814.6974	-8114.1685	549.582	-8084.2176	281.8275	0.085	157.38	<.0001
5	**CEF**	0.1685	0.1673	-4876.5829	-4876.5488	-8176.0991	477.9888	-8139.9764	276.7089	0.084	64.36	<.0001
6	**usda_SILT LOAM**	0.1876	0.1861	-4951.0408	-4950.997	-8250.5076	393.7316	-8208.3333	270.5399	0.082	77.19	<.0001
7	**usda_SAND**	0.2257	0.224	-5107.3777	-5107.323	-8406.3311	223.3352	-8358.5692	258.1647	0.078	161.81	<.0001
8	**Temp**	0.2333	0.2315*	-5138.2057*	-5138.1388*	-8437.1059*	190.6165*	-8383.2961*	255.8591*	0.077	32.9	<.0001

UNCORRECTED DATABASE – **DIS50** - LASSO selection summary

Step	Effect	Model R-Square	Adjusted R-Square	AIC	AICC	BIC	CP	SBC	PRESS	ASE	F Value	Pr > F
0	**Intercept**	0	0	-4544.8183	-4544.8148	-7990.4864	1239.2033	-7985.6736	338.6097	0.098	0	1
1	**SRTM**	0.0719	0.0716	-4799.9404	-4799.9334	-8245.8551	904.4638	-8234.651	314.4885	0.091	266.81	<.0001
2	**Rain**	0.1042	0.1037	-4919.9179	-4919.9063	-8366.0536	755.2954	-8348.4839	303.7822	0.088	124.05	<.0001
3	**CNBL**	0.1052	0.1045	-4921.8691	-4921.8517	-8368.3539	752.4876	-8344.2904	303.693	0.088	3.95	0.047
4	**Mdepth**	0.1451	0.1441	-5076.8022	-5076.7778	-8523.2348	567.9219	-8493.0788	290.352	0.084	160.33	<.0001
5	**CEF**	0.1596	0.1583	-5133.6463	-5133.6137	-8580.1719	502.1285	-8543.7782	285.7025	0.083	59.25	<.0001
6	**usda_SILT LOAM**	0.1772	0.1758	-5204.7195	-5204.6776	-8651.2229	421.538	-8608.7068	279.8788	0.081	73.7	<.0001
7	**usda_SAND**	0.215	0.2134	-5364.8893	-5364.8369	-8810.9001	246.3861	-8762.7319	267.3133	0.077	165.66	<.0001
8	**Temp**	0.2219	0.2201*	-5393.3117*	-5393.2476*	-8839.2963*	216.0721*	-8785.0096*	265.2222*	0.076	30.48	<.0001

CORRECTED DATABASE – **DID50** - LASSO selection summary

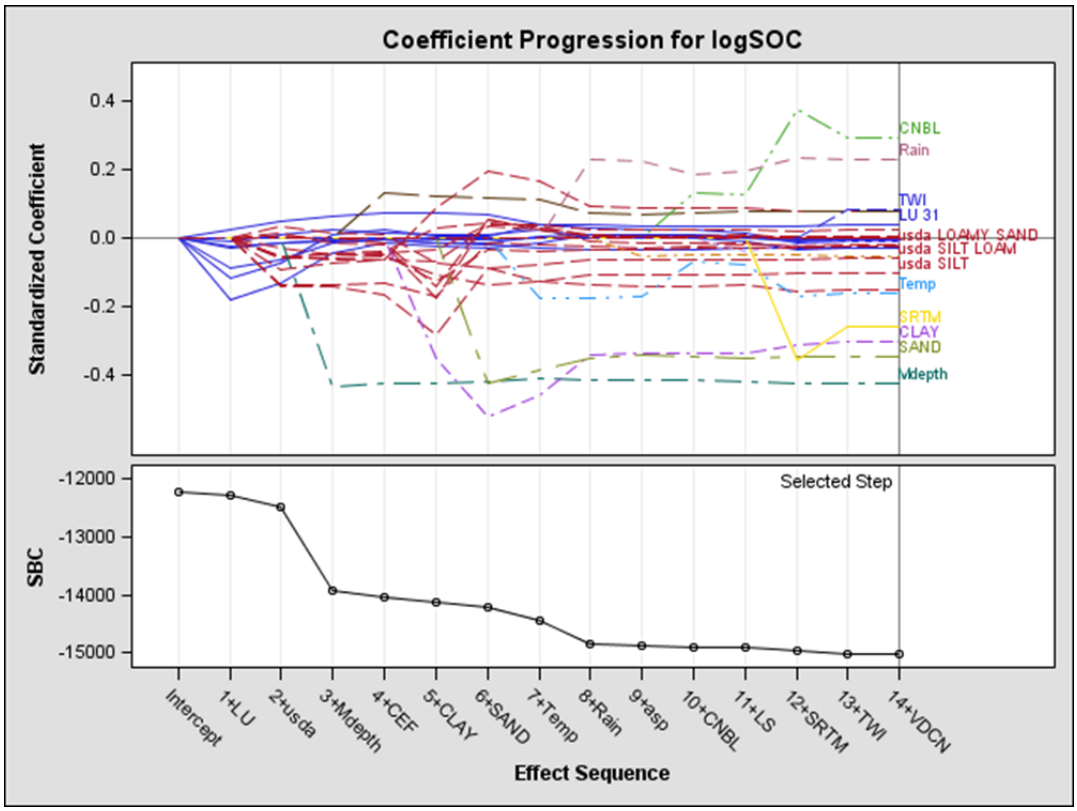
Step	Effect	Model R-Square	Adjusted R-Square	AIC	AICC	BIC	CP	SBC	PRESS	ASE	F Value	Pr > F
0	**Intercept**	0	0	-2703.9866	-2703.983	-6078.6562	1216.8362	-6073.8627	556.5978	0.165	0	1
1	**CNBL**	0.0733	0.073	-2958.8432	-2958.8361	-6333.7549	882.4624	-6322.5955	516.1482	0.153	266.73	<.0001
2	**Mdepth**	0.1315	0.131	-3175.7369	-3175.725	-6550.7043	617.336	-6533.3653	484.0371	0.143	225.95	<.0001
3	**Rain**	0.1596	0.1588	-3284.4547	-3284.4369	-6659.4872	490.6796	-6635.9593	468.7408	0.139	112.42	<.0001
4	**usda_SAND**	0.1901	0.1891	-3407.2332	-3407.2083	-6782.1722	352.6556	-6752.614	452.0882	0.134	126.93	<.0001
5	**TWI**	0.2117	0.2105	-3496.5118	-3496.4785	-6871.3284	255.4549	-6835.7686	440.2827	0.13	92.36	<.0001
6	**LU_32(SU_NAT)**	0.224	0.2226	-3547.5556	-3547.5128	-6922.2964	201.0284	-6880.6886	433.6958	0.128	53.35	<.0001
7	**usda_SANDY CLAY LOAM**	0.2332	0.2316	-3586.0274	-3585.9739	-6960.6882	160.5683	-6913.0366	428.7776	0.126	40.62	<.0001
8	**SAND**	0.2413	0.2395	-3619.6546	-3619.5892	-6994.2129	125.61	-6940.5399	424.6354	0.125	35.72	<.0001
9	**VDCN**	0.2459	0.2439	-3638.35	-3638.271	-7012.85	106.3202	-6953.111	422.301	0.124	20.7	<.0001
10	**CEF**	0.2493	0.2471	-3651.5166	-3651.4238	-7025.9696	92.8002	-6960.1542	420.7304	0.124	15.15	0.0001
11	**Temp**	0.2579	0.2555	-3688.4537	-3688.3454	-7062.6743	55.2745	-6990.9675	415.9328	0.122	39.02	<.0001
12	**usda_SILT**	0.2604	0.2577	-3697.6033	-3697.4783	-7071.7569	46.0402	-6993.9932	414.819	0.122	11.12	0.0009
13	**usda_SILT LOAM**	0.2625	0.2596	-3705.3387	-3705.1957	-7079.4236	38.2611	-6995.6047*	413.9026	0.122	9.71	0.0018
14	**LU_31(SU_NAT)**	0.2634	0.2603*	-3707.4144*	-3707.2523*	-7081.4692*	36.1755*	-6991.5565	413.6399*	0.121	4.06	0.044

124 UNCORRECTED DATABASE – **DID50** - LASSO selection summary

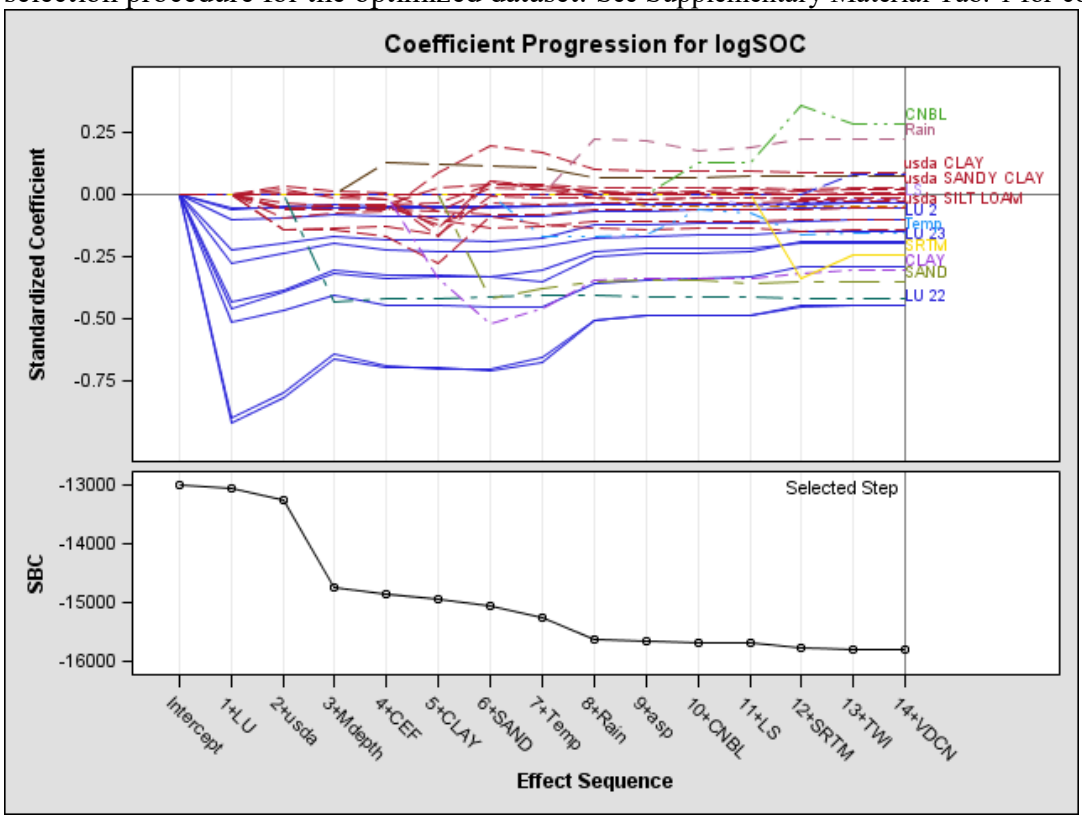
Step	Effect	Model R-Square	Adjusted R-Square	AIC	AICC	BIC	CP	SBC	PRESS	ASE	F Value	Pr > F
0	**Intercept**	0	0	-2944.8177	-2944.8143	-6485.4656	1230.2704	-6480.6458	566.4834	0.16	0	1
1	**CNBL**	0.0728	0.0726	-3210.5647	-3210.5579	-6751.4412	884.9247	-6740.2209	525.5656	0.148	277.98	<.0001
2	**Mdepth**	0.1282	0.1277	-3426.3377	-3426.3264	-6967.2725	623.1556	-6949.8221	494.5144	0.139	224.42	<.0001
3	**Rain**	0.1508	0.1501	-3517.4901	-3517.4731	-7058.5277	517.19	-7034.8026	481.989	0.136	94.28	<.0001
4	**LU_2(SU_AGR)**	0.1521	0.1511	-3520.8515	-3520.8277	-7062.1267	513.0621	-7031.9921	481.5181	0.136	5.36	0.0207
5	**usda_SAND**	0.1826	0.1815	-3648.6224	-3648.5907	-7189.7401	369.5334	-7153.5911	464.574	0.131	131.95	<.0001
6	**LU_32(SU_NAT)**	0.1968	0.1954	-3708.5388	-3708.498	-7249.6102	303.9563	-7207.3356	456.8246	0.129	62.34	<.0001
7	**TWI**	0.2141	0.2126	-3783.8233	-3783.7723	-7324.7143	223.2476	-7276.4483	447.2117	0.126	77.96	<.0001
8	**SAND**	0.2278	0.226	-3843.726	-3843.6637	-7384.4245	160.2904	-7330.1791	439.8175	0.124	62.29	<.0001
9	**usda_SANDY CLAY LOAM**	0.2303	0.2284	-3853.625	-3853.55	-7394.335	149.9343	-7333.906	438.594	0.123	11.89	6E-04
10	**LU_31(SU_NAT)**	0.236	0.2338	-3877.6594	-3877.5709	-7418.2833	125.043	-7351.7687	435.5983	0.122	26.05	<.0001
11	**VDCN**	0.2412	0.2388	-3899.86	-3899.7568	-7440.3798	102.2229	-7367.7974	432.8857	0.121	24.2	<.0001
12	**CEF**	0.2442	0.2417	-3912.0549	-3911.9358	-7452.5154	89.7433	-7373.8204	431.4712	0.121	14.17	0.0002
13	**usda_SILT**	0.2467	0.244	-3921.848	-3921.7118	-7462.2519	79.7578	-7377.4417	430.2842	0.121	11.77	0.0006
14	**Temp**	0.2541	0.2511	-3954.6666	-3954.5122	-7494.8067	46.6033	-7404.0884	426.0997	0.119	34.84	<.0001
15	**usda_SILT LOAM**	0.2563	0.2531	-3962.8418	-3962.668	-7502.8995	38.3951	-7406.0917	425.1494	0.119	10.14	0.0015
16	**LS**	0.2581	0.2548	-3969.8009	-3969.6066	-7509.7762	31.4312	-7406.8789*	424.4394	0.119	8.93	0.0028

17	**usda_SANDY LOAM**	0.2588	0.2553*	-3971.2317*	-3971.0158*	-7511.1707*	30.0045*	-7402.1378	424.3775*	0.119	3.42	0.0647

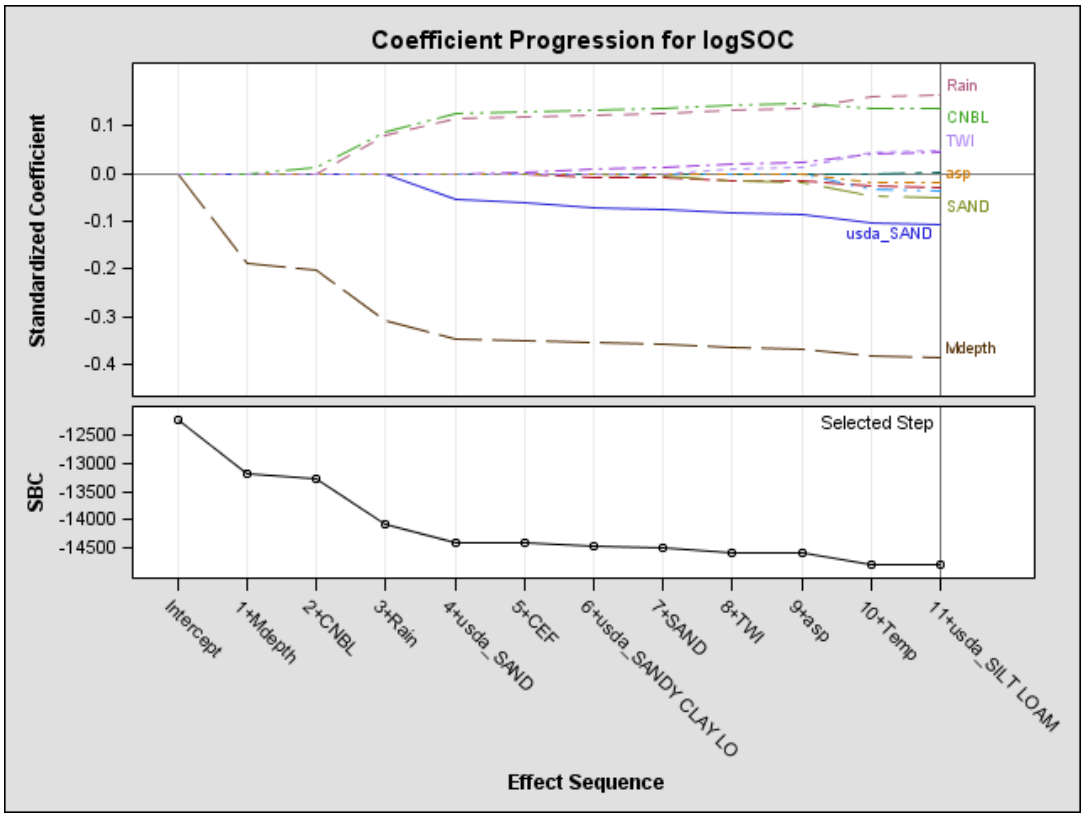
125



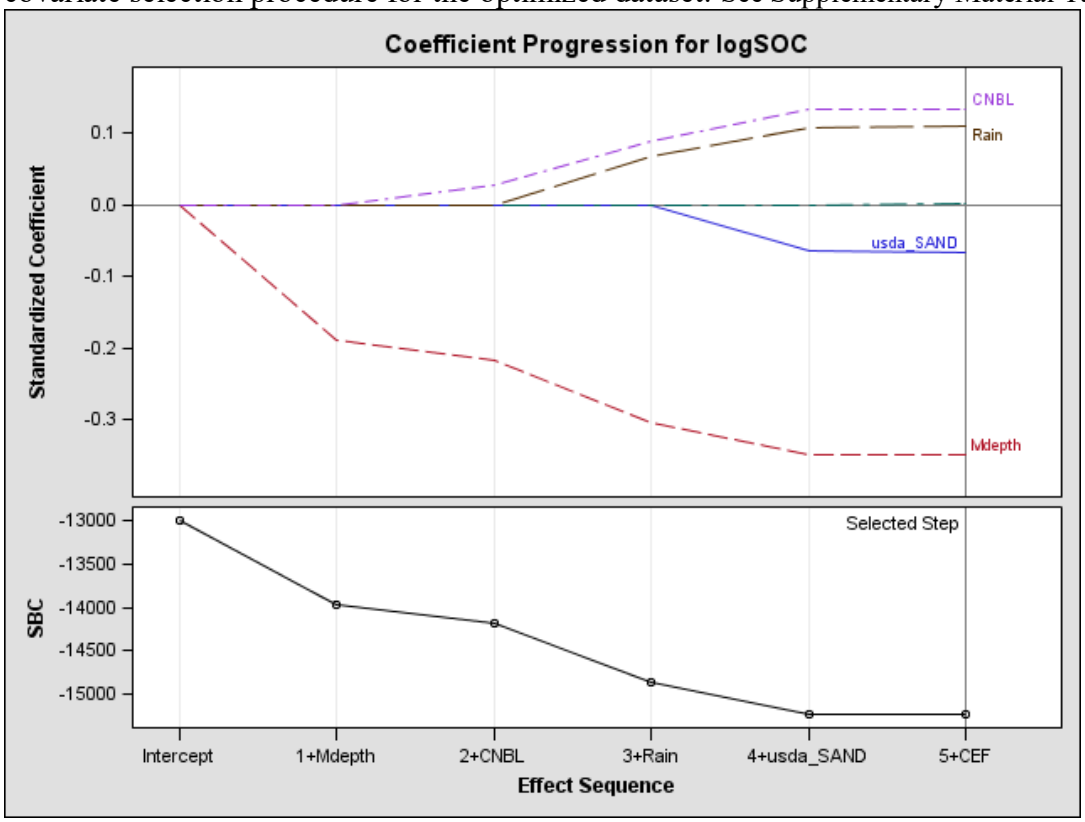
Supplementary Fig. 4. Coefficient of progression of each Standardised coefficient of the logSOC modelling at increasing the step of model building (i.e. inclusion of new variables) with no covariate selection procedure for the optimized dataset. See Supplementary Material Tab. 1 for codes.



Supplementary Fig. 5. Coefficient of progression of each Standardised coefficient of the logSOC modelling at increasing the step of model building (i.e. inclusion of new variables) with no covariate selection procedure for the non-optimized dataset. See Supplementary Material Tab. 1 for codes.



Supplementary Fig. 6. Coefficient of progression of each Standardised coefficient of the logSOC modelling at increasing the step of model building (i.e. inclusion of new variables) with LASSO covariate selection procedure for the optimized dataset. See Supplementary Material Tab. 1 for codes.



Supplementary Fig. 7. Coefficient of progression of each Standardised coefficient of the logSOC modelling at increasing the step of model building (i.e. inclusion of new variables) with LASSO covariate selection procedure for the non-optimized dataset. See Supplementary Material Tab. 1 for codes.

Chapter 4- Modelling the topsoil carbon stock of agricultural lands with the Stochastic Gradient Treeboost in a semi-arid Mediterranean region

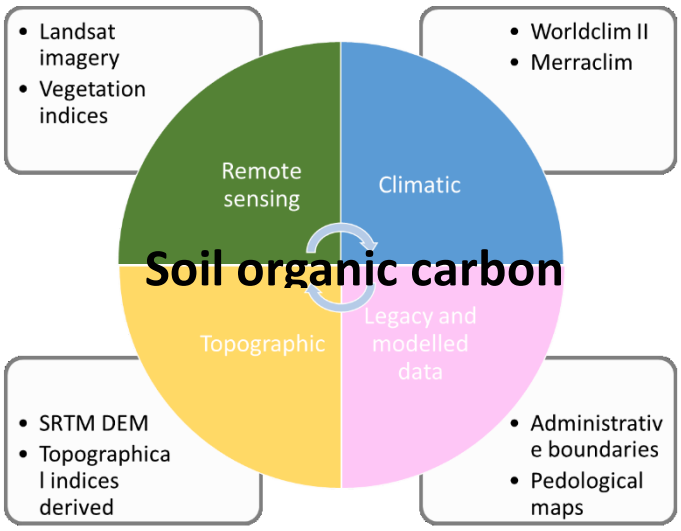
Slightly modified from a paper Published in Geoderma 2017 Schillaci, C., Lombardo, L., Saia, S., Fantappiè, M., Märker, M., Acutis, M., 2017b. Modelling the topsoil carbon stock of agricultural lands with the Stochastic Gradient Treeboost in a semi-arid Mediterranean region. Geoderma 286, 35–45. doi:10.1016/j.geoderma.2016.10.019

Keywords: Soil Organic Carbon; Stochastic modelling; Terrain analysis; Remote Sensing; Climate data

Abstract

Efficient modelling methods to assess soil organic carbon (SOC) stocks have a pivotal importance as inputs for global carbon cycle studies and decision-making processes. However, laboratory analyses of SOC field samples are costly and time consuming. Global-scale estimates of SOC were recently made according to categorical variables, including land use and soil texture. Remote sensing (RS) data can contribute to the better modelling of the spatial distribution of SOC stock at a regional scale. In the present study, we used Stochastic Gradient Treeboost (SGT) to estimate the topsoil (0–30 cm) SOC stock of a Mediterranean semiarid area (Sicily, Italy, 25,286 km^2). In particular, our study examined agricultural lands, which represent approximately 64% of the entire region. An extensive soil dataset (2202 samples, 1 profile/7.31 km^2 on average) was acquired from the soil database of Sicily. The georeferenced field observations were intersected with remotely sensed environmental data and other spatial data, including climatic data from WORLDCLIM, land cover from CORINE, soil texture, topography and derived indices. Finally, the SGT was compared to published global estimates (GSOC) and data from the International Soil Reference and Information Centre (ISRIC) Soil Grids by comparing the pseudo-regressions of the SGT, GSOC and ISRIC with soil observations. The mean SOC stock across the entire region that was estimated by GSOC and ISRIC was 3.9% lower and 46.2% higher compared to the SGT. The SGT efficiently predicted SOC stocks that were 70 t ha^{-1} (corresponding to the 90th percentile of the observed values). On average, the coefficient of variation of the SGT model was 3.6% when computed on the whole dataset and remained lower than 23% when computed on a distribution basis. The SGT mean absolute error was 14.84 t ha−1, 18.4% and 36.3% lower than GSOC and ISRIC, respectively. The mean annual rainfall, soil texture, land use, mean annual temperature and Landsat 7 ETM+ panchromatic Band 8 were the most important predictors of SOC stock. Finally, SOC stocks were estimated for each land cover class. SGT predicted SOC stock better than GSOC and ISRIC for most data. This resulted in a percentage of data in the prediction confidence interval ± 50% compared to the observed values of 71.4%, 65.8%, and 50.7%

for SGT, GSOC, and SGT, respectively. This consisted of a higher R2 and a slope (β) that was closer to 1 for the pseudo-regression constructed with SGT compared to GSOC and ISRIC. In conclusion, the results of the present study showed that the integration of RS with climatic and soil texture spatial data could strongly improve SOC prediction in a semi-arid Mediterranean region. In addition, the panchromatic band of Landsat 7 ETM+ was more predictive compared to the conventionally used NDVI. This information is crucial to guiding decision-making processes, especially at a regional scale and/or in semi-arid Mediterranean areas. The model performance of the SGT could be further improved by adopting predictors with greater spatial resolutions. The results of the present experiment yield valuable information, especially for assessing climate change or land use change scenarios for SOC stocks and their spatial distribution.



Graphical abstract

4.1 Introduction

Agricultural land plays a pivotal role in terms of carbon sequestration ability due to soil organic carbon (SOC) being in both topsoil and subsoil. SOC is recognized as the most important indicator of soil quality and determines plant productivity (Lal, 2004) through a wide range of mechanisms, including its activity as a main source of energy for microbial processes (Hudson, 1994); soil cation exchange capacity (Chan et al., 1992; Riffaldi et al., 1994); the effect on water holding capacity and infiltration rate through the soil profile (Macrae and Mehuys, 1985); and the reduction of the soil bulk density and cohesion (Soane, 1990). In addition, SOC is one of the most important CO_2 sequestration sources (Post et al., 1982). Rules and regulations were established by a number of countries to change the existing carbon balance, including the reduction of CO_2 emissions in the atmosphere from both soil and other sources. To achieve this, the CO_2 sequestration abilities of soils also need to be increased. The creation of a CO_2 accounting system, in which CO_2 levels are counted as C credits and

debits, is presently a key agenda point in Europe and around the world. However, the creation of such accounting systems relies on correct knowledge of the C content of soils at the regional (Martin et al., 2010) or farm level (de Gruijter et al., 2016). In turn, such knowledge is fundamental to monitoring changes in the C stocks and a better understanding of the global C cycle (Martin et al., 2014). Different global or regional estimates were produced (Batjes, 2009, 1996; Hiederer and Köchy, 2012; Nachtergaele et al., 2008; Viscarra Rossel et al., 2016) based on various modelling and data-mining algorithms (see Minasny et al., 2013 and reference therein). Based on these efforts, various global estimates of SOC stock in the soil were provided, including the International Soil Reference and Information Centre (ISRIC) Soil Grids (Batjes, 2016, 2009; Hengl et al., 2014) and the Global Soil Organic Carbon Estimates (GSOC) (Hiederer and Köchy, 2012) by the Joint Research Centre (JRC) of the European Commission.

De Brogniez et al. (2015) created a topsoil organic carbon map using Generalized Additive Models (GAM) for the entire European Union (EU). In this study, topography and land use were recognized as key indicators to assess SOC stock and its distribution. In addition, Novara et al. (2013, 2014) also highlighted that some Mediterranean soils see increases in SOC levels when they are no longer cultivated. This suggests that soil cultivation can play a major role in affecting SOC under Mediterranean conditions. However, SOC dynamics also depend on other factors, including climate, soil type and texture, soil moisture, temperature regimes, lithology, morphology, land use history and management (Fantappiè et al., 2010, 2011b; Pisante, M.et al, 2015). The knowledge of soil quality is a priority to support agricultural productivity and environmental quality. However, field sampling and laboratory analyses of SOC are costly and time consuming. Remote sensing (RS) data can contribute to modelling SOC information on a large scale (Gomez et al., 2008). In addition, RS predictors can also reduce uncertainties in SOC mapping through geographical soil unit classification (Köchy et al., 2015b). However, in complex terrain, the high number of ecological determinants of the topsoil organic carbon can reduce the outcomes of the prediction (Yao et al., 2013), especially if samples from some areas are lacking. Nonetheless, the use of a high number of topographic and other ecological indices as predictors can increase the ability of the models to explain large parts of the amount of plant residues that were returned to the soil as well as SOC variation (Ferrara et al., 2009; Grimm et al., 2008).

Strategies for modelling SOC for large areas, e.g., at a regional scale, often rely on data mining approaches. However, these methodologies require a correction for spatial heterogeneity, outliers or correct sampling design to achieve a highly precise estimation of SOC stocks (Brus, 2015; Friedman et al., 2000; Schapire and Freund, 2012; Viscarra Rossel et al., 2016). The Stochastic Gradient

Treeboost (SGT; Friedman, 2002), which is also referred to as Boosted Regression Trees (BRT; Elith et al., 2008), is an improvement of the classification and regression trees (CART; Breiman et al., 1984). The SGT aims at identifying group membership to classes (the SOC stock value at a given cell or pixel) by sequentially partitioning the predictors' hyperspace into random trees (Lombardo et al., 2015). In particular, the SGT binary-splits the observations in homogeneous groups of the target variable as a function of combined explanatory variables and afterward combines several additive regression models in a forward stepwise procedure (Elith et al., 2008). The SGT was previously applied to an SOC model at a regional scale in organic soils (Bou Kheir et al., 2010) and in temperate environments (Martin et al., 2014). However, little information is available concerning SOC modelling in Mediterranean areas.

The Italian peninsula is characterized by a complex terrain with a high incidence of hilly and mountainous areas. As mentioned above, this can limit the accuracy of interpolations if the availability of data in the region is low. Conversely, the use of machine learning approaches allows for the recognition of causative relationships between SOC, topographic attributes and RS indices (e.g., McBratney et al., 2003). Lugato et al. (2014) provided an SOC stock estimation on a European scale using modelling techniques and validated them using the European Environment Information and Observation Network for Soil (EIONET-SOIL data by Panagos et al., 2013a, b). However, no specific regional modelling examples were provided for semi-arid Mediterranean regions.

The aim of the present work was to estimate the SOC stock through the SGT using a set of topographical and environmental covariates. Sicily was chosen for the model application since more than half of its surface is extensively cultivated and extensively sampled. In addition, across the island, there is a strong heterogeneity of agro-ecosystems in terms of soil type, texture, land use and microclimates. In the present study, we also compared the SGT results to those obtained from the GSOC estimate (Hiederer and Köchy, 2012; Panagos et al., 2012) and to those obtained from the International Soil Reference and Information Centre (ISRIC) Soil Grids (Batjes, 2016, 2009; Hengl et al., 2014). In particular, the SGT implemented in the present study was constructed using a 3-arcsec spatial resolution, whereas GSOC and ISRIC are freely available for scientific purposes as a spatial layer with a 30-arcsec spatial resolution and offer a benchmark regarding the overall SOC stock pattern in the agro-ecosystems.

4.2 Material and methods

4.2.1 Study area

Sicily is an Italian island in the middle of the Mediterranean Sea and has an area of 25,286 km² (36.64° to 38.30° N; 12.42° to 15.66° E), excluding its 37 ancillary islands. According to the CORINE land cover 2000 (CLC2000; Bossard et al., 2000), 64.1% of its territory is cropped. The remaining 35.9% encompasses non-agricultural ecosystems, including urban areas, Mediterranean maquis, dunes, coastal systems, forests, and industrial complexes. Sicily has several sub-climatic zones, all of which are included in the temperate Mediterranean belt, with mean annual temperatures usually higher than 15.8 °C, where summer is the driest period of the year. According to the climate classification of the Italian territory (E.A.C. Costantini et al., 2013), most of Sicily has a Mediterranean to subtropical climate that is partly semi-arid and is characterized by low rainfall, high air temperatures and high evapo-transpiration. Mediterranean subcontinental to continental climates, which are partly semi-arid to arid, typify the hinterland of the island. The mountain areas (Madonie, Sicani, Nebrodi and Peloritani ridges) are barely cultivated and are characterized by Mediterranean sub-oceanic to Mediterranean subtropical climates that are influenced by mountains. The continentality index, which is determined by the difference between the mean air temperature of summer and winter, is similar in all climatic regions. According to the World Reference Base for soils (IUSS Working Group WRB, 2014), the dominant soils in Sicily are Calcaric Regosols, Haplic Calcisols, Calcic Vertisols, Vitric or Silandic Andosols, Calcaric and/or Mollic Leptosols, Calcaric Phaeozems, and Fluvic Cambisols (Fantappiè et al., 2011a).

4.2.2 SOC stock analysis and database

The soil database of Sicily was the source of information for SOC (dag kg^{-1}, Fig. 1) and bulk density (BD; g cm^{-3}). It stores information for approximately 5,658 georeferenced observations (soil profiles and minipits), of which 2,891 are analysed for SOC following Walkley-Black (1934) and on fine-earth fraction (FEF); 1,049 of them were also analysed for BD. Missing BDs were estimated with the following pedotransfer function (Pellegrini et al., 2007):

$$BD = 1.677601 - 0.0000116 * clay^2 - 0.000448 * sand - 0.235085 * \sqrt{SOC} \quad (1)$$

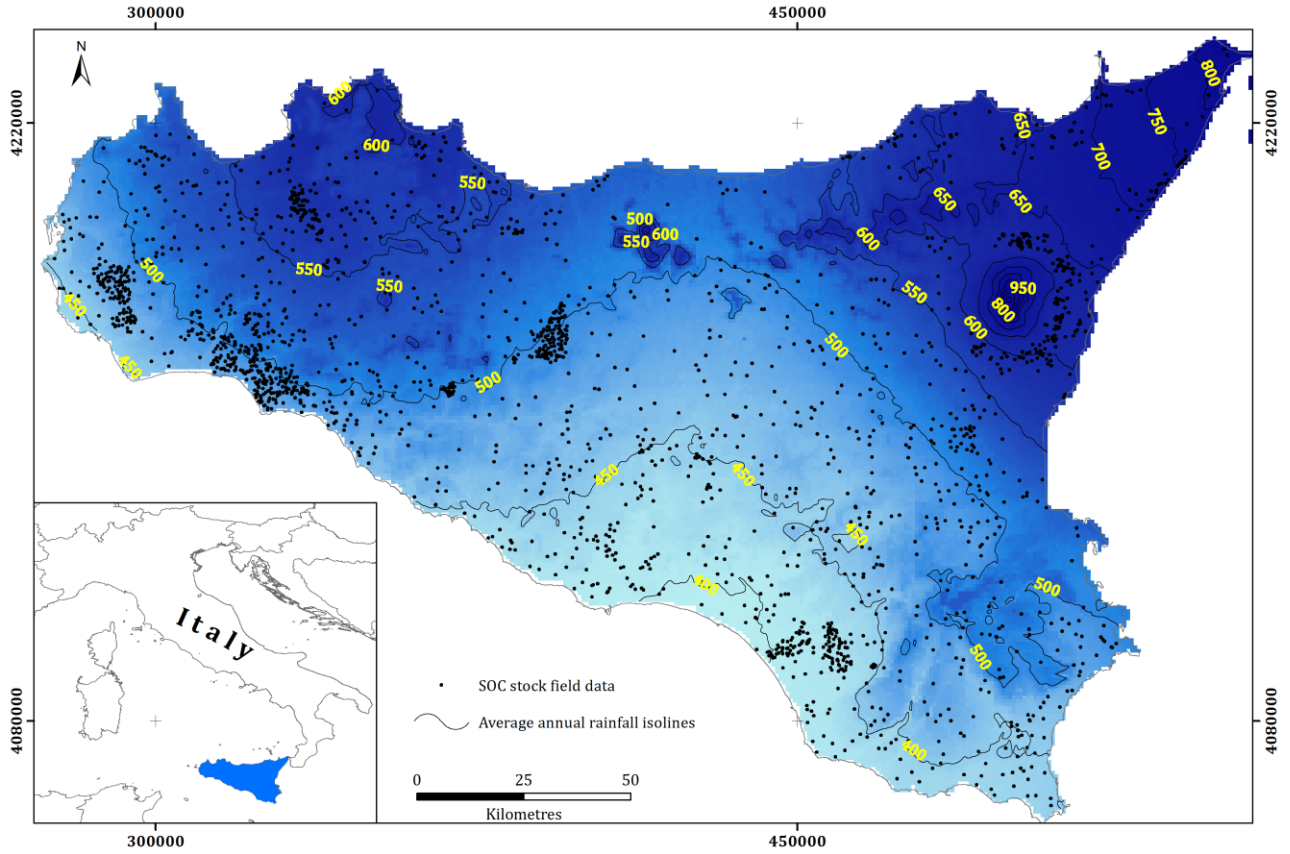


Fig. 1. Locations of the sampling sites in the area under study (Sicily). The mean annual rainfall is shown in mm/year.

where the clay and sand content is expressed as dag kg^{-1}, respectively. This formula was computed from data from the same and similar environments and presently represents one of the most accurate BD estimation formulas (Pellegrini et al., 2007).

A total of 2,891 sites were sampled and analysed for more than one horizon. Therefore, we selected all A and Ap horizons up to 0.3 m depth from the mineral soil surface and from any other type of soil horizon except of O, Oh, Of, Oi and C, with a lower boundary within 0.3 m from the mineral soil surface. The SOC stock (t ha^{-1}) for each of the resulting 3,674 horizons was calculated with this formula:

$$CS(\text{t ha}^{-1}) = T * SOC * FEF * BD \qquad (2)$$

where CS is the SOC stock (t ha^{-1}), T is the horizon thickness in meters, SOC is the soil organic carbon content (dag kg^{-1}), FEF is the fine-earth fraction in volumetric percentage (daL m^{-3}) and BD is the bulk density (g cm^{-3}) analysed when the data were available and estimated by the pedotransfer function when the data were missing. The SOC stock values of the 3,674 horizons were summarized to obtain a unique value for each of the 2,891 observations.

The dataset used in the modelling procedure initially included all the aforementioned 2,891 sampled locations. To constrain the modelling procedure to the locations that mostly represent agricultural land cover, the SOC stock field measurement vector layer was intersected with the available CORINE land cover 2000 map (Fig. 2). Only the points falling into the CORINE agricultural areas were included in the modelling to train and validate the model. Based on this procedure, 2202 locations were selected after excluding 689 samples from semi-natural areas, meadows, and woods. The whole dataset and the model are considered representative for the year 2000.

4.2.3 SGT for SOC stock estimation

The SGT combines CART together with the Stochastic Gradient Boosting algorithm (Friedman, 2002). The algorithm was described as an additive regression approach in which many weak learners are added to the basic tree structure to minimize the negative gradient of a Huber-M loss function (Friedman, 2001). The first step of this procedure consists of a CART analysis that recursively screens the observations in matched datasets that are composed of a dependent variable, either categorical (classification) or continuous (regression), and one or many explanatory variables. The SGT iteratively generates trees of a fixed dimension. Each tree is based upon the previous one, minimizing the loss function to improve the predictive performance. The procedure ceases when the creation of trees produces overfitting effects, which are evaluated by scoring the prediction residuals over a random independent sub-sample. In the present research, 10 replicates were randomly generated and modelled from the original SOC stock dataset. Each replicate was randomly built, extracting 75% of the SOC stock data for calibration purposes. The remaining 25% was kept for model validation. The model was actually built with a *3-arcsec spatial resolution; however, since both Wordclim-derived predictors and the reference models (GSOC and ISRIC) were available at a 30-arcsec spatial resolution, the SGT was re-sampled at a 30-arcsec spatial resolution to compare its performance with that of GSOC and ISRIC.*

SGT modelling was performed by means of TreeNet® (Salford Systems), with a maximum of 200 trees and six nodes. The set of independent covariates comprised DEM-derived topographic attributes, RS indices, and thematic layers, including that from CORINE land cover 2000. The predictive skills were evaluated by the difference measured against predicted SOC stock (as both absolute values and percentage misfits) and the coefficient of variation of the modelling procedure. The role of the predictors was assessed through the predictor importance (PI) and response curve (RC) plots *(Friedman, 2002)*. PI represents the strength with which each predictor affects the outcome, which is normalised to the greatest contributor. RCs link the domain of each predictor to the deviation from the mean SOC stock value that the given covariate is able to produce, keeping all

the other explanatory variables constant. The SGT was constructed with a spatial resolution of 3 arcsec to exploit the maximum resolution of the predictors. In a subsequent phase, we downscaled the SGT predictive map to 1-km^2 spatial resolution to compare it to the GSOC and ISRIC maps. Finally, an uncertainty map (as prediction confidence map) displaying the 'SGT-predicted to observed values' ratio was built.

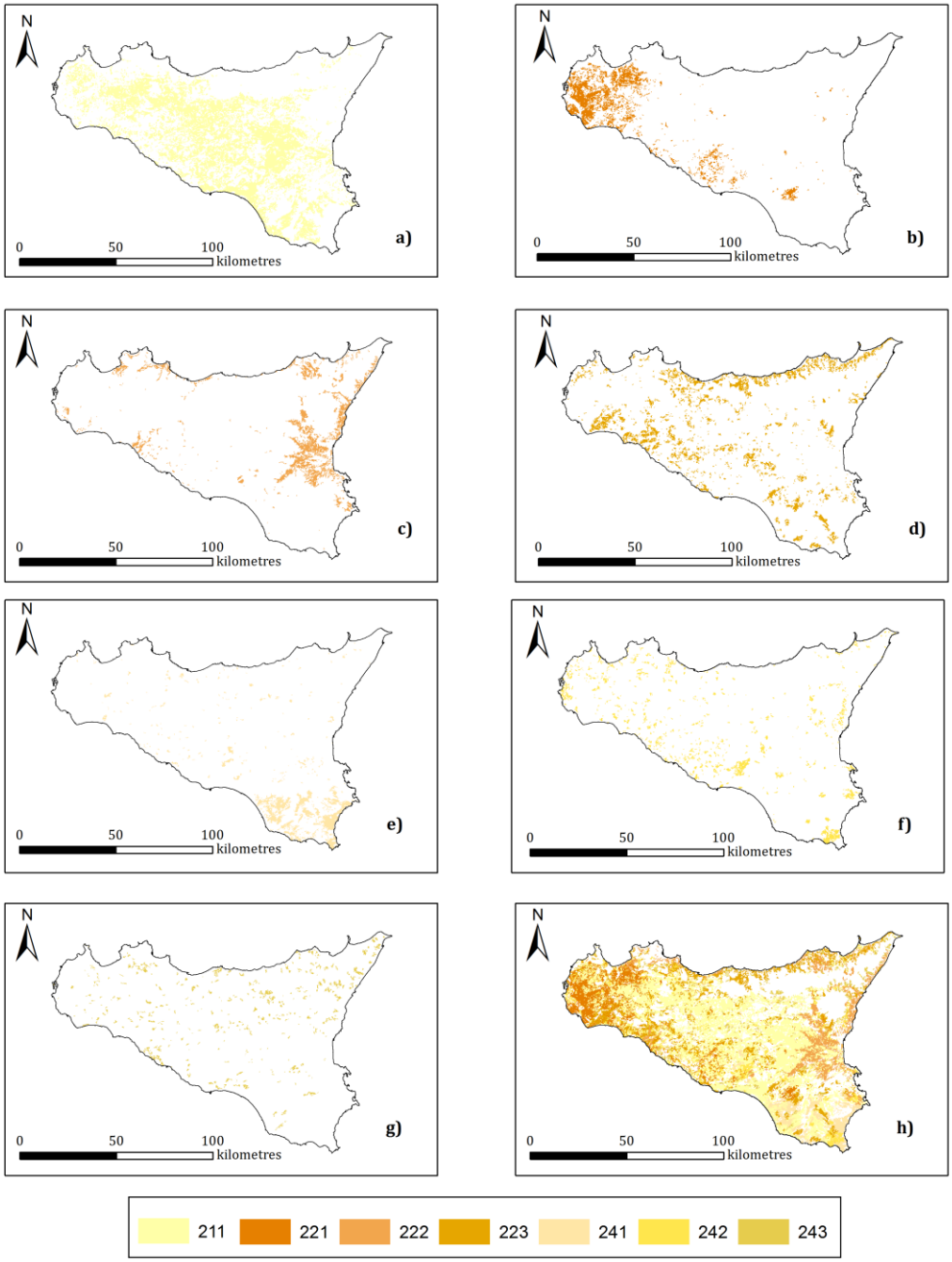


Fig. 2. CORINE land cover 2000 coverage of samples in the area under study (Sicily). The codes are drawn from CORINE 2000. a) 211, non-irrigated arable; b) 221, vineyards; c) 222, fruit and berry, 223; d) olive groves; e) 241, annual with permanent crops; f) 242, complex cultivation patterns; g) 243, land principally occupied by agriculture, with significant areas of natural vegetation; h) all land classes previously showed.

4.2.4 Covariates used in the modelling procedures.

Texture: The soil texture after the USDA classification (Schoeneberger et al., 2012) was derived from the soil map of Sicily (Fantappiè et al., 2011a). Texture was used as a categorical covariate that was classified following the USDA system. The texture codes are provided in Table 1.

Bioclimatic data: Worldclim bioclimatic data (Hijmans et al., 2005) were used as climatic covariates. This climatic dataset is currently the most detailed data on a global scale (delivered at 1 km^2 cell size). The global dataset was derived with the thin-plate smoothing spline algorithm, where latitude, longitude, and elevation were used as independent variables. This interpolative methodology uses different data sources. These climatic records belong to the time window between the years 1950 and 2000 and ensure the depiction of environmental variability that is otherwise lost at lower resolutions. In the present study, monthly average temperature (layer BIO_1) and precipitation (BIO_{12}) were used.

Topographic covariates: morphometric independent variables were constructed within an open source GIS environment (SAGA GIS). We derived all the topographic attributes from the Shuttle Radar Topography Mission SRTM-C DEM with a 3-arcsec (85 meter) spatial resolution. Eight terrain attributes were calculated and included 1) slope (Zevenbergen and Thorne, 1987); 2) aspect (Wilson, J.P. , Gallant, 2000); 3) plan curvature; 4) profile curvature (Shary et al., 2002); 5) Topographic Wetness Index (Moore et al., 1993, 1991); 6) length-slope factor (Behrens et al., 2010); 7) catchment area (Zevenbergen and Thorne, 1987); and 8) landform classification (Weiss, 2001). Categorical predictor codes are provided in Table 1.

RS-derived covariates: The Landsat ETM7+ imagery was used to derive vegetation and soil-specific indexes, which were also included as explanatory variables in the modelling phase. We used geometrically corrected images L1G. The multi-temporal mosaic required normalization to adjust for inconsistencies between images because of the proximity of the sun, earth and zenith angle. The procedure involved the conversion of the digital number to radiance in the sensor and then to reflectance. The calibration coefficient was provided in the imagery metadata (Guyot and Gu, 1994). The images were obtained by mosaicking four Landsat 7 ETM+ scenes (east: p188r0347k19990926z33nn1/8; centre: p189r034_7k20010501z33nn 1/8; west: LE71900342001160EDC00; and south-eastern: LE71880352000256SGS00). Two RS datasets were calculated and imported as predictors: Normalized Difference Vegetation Index (NDVI; Rouse Jr. et al., 1974) and the panchromatic band 8 (0.5 – 0.9 μm). This latter band has an original spatial resolution of 14.25 meters and was subsequently downscaled to an 85 m resolution. This predictor was chosen due to its strong link to the soil colour and, thus, the C content. In particular, the panchromatic channel is sensitive to reflected light energy across a broad range of wavelengths,

including blue, green, red and near. The Landsat imagery was freely acquired from the United States Geological Survey catalogue (http://earthexplorer.usgs.gov).

Land cover: The CORINE land cover from the year 2000 was used to both address the analysis only for the agricultural lands and as a predictor. The land cover for a small (<1,000 ha) area was merged with the most similar land cover types used in the analysis. According to the CORINE level 3, the land cover types used in the modelling stage were 1) non-irrigated arable land, 2) vineyards, 3) fruit trees and berry plantations, 4) olive groves, 5) annual crops associated with permanent crops, 6) complex cultivation patterns, 7) land principally occupied by agriculture with significant areas of natural vegetation. The CORINE codes are provided in Table 1.

Landforms*		**Aspects**		**Land Use*/CORINE code**		**Texture USDA**	
Internal Code	**Class**	**Internal Code**	**Class**	**Internal Code**	**Class**	**Internal Code**	**Class**
0	Canyons	**0**		**0**		**0**	No soil
1	Midslope drainage	**1**	5.5-0.5 = N	**1**	Non-irrigated arable / 211	**1**	Clay
2	Upland drainage	**2**	0.5-1=NE	**2**	Vineyards / 221	**2**	Silty-Clay
3	U-shape valleys	**3**	1-2=E	**3**	Fruit and berry / 222	**3**	Loam
4	Plains	**4**	2-2.5=SE	**4**	Olive groves/ 223	**4**	Clay-Loam
5	Open slopes	**5**	2.5-3.5= S	**5**	Annual with permanent crops / 241	**5**	Silty-Loam
6	Upper slopes	**6**	3.5-4= SW	**6**	Complex cultivation patterns / 242	**6**	Silty-Clay-Loam
7	Local ridges-hills in valleys	**7**	4-5 =W	**7**	Land principally occupied by agriculture, with significant areas of natural vegetation / 243	**7**	Sandy-Loam
8	Midslope ridges, small hills in plains	**8**	5-5.5= NW			**8**	Sandy Clay Loam
9	Mountain tops, high ridges					**9**	Sand
						10	Sandy-Loam

*Landforms from Weiss (2001); land cover from CORINE 2000.

Table 1. Landform, aspect, land use and texture classes used in the present work. Internal codes of each variable are referred to those used in Fig. 4.

4.2.5 Global Soil Organic Carbon (GSOC) and International Soil Reference and Information Centre (ISRIC) Soil Grids Estimates

Two global SOC stock estimates, GSOC and ISRIC, were drawn up by the JRC-IES (Hiederer and Köchy, 2012) and ISRIC (http://soilgrids1km.isric.org/; Batjes, 2016, 2009; Hengl et al., 2014), respectively. The aim of these estimates was to provide scientific-technical support for the protection and sustainable development of the environment at regional and global scales. The generation of the data layer in GSOC relied on the use of different Pedo-Transfer Functions (PTF). The structure of the PTF has been revised from the PTF Pedo Transfer Rule Database of the European Soil Database (distribution version v2.0) (http://eusoils.jrc.ec.europa.eu/esdb_archive/esdbv2/fr_intro.htm) to accommodate special environmental conditions such as peatland and other organic soils. Some features, such as temperature differences, were used in these procedures for the PTF to produce a SOC content response. The GSOC model was validated using field sample data from soil profiles across Europe. The PTF conditions and the derived GSOC estimates are applicable to the topsoil layer. We used the GSOC of the agricultural land cover types (CORINE mask) by resampling the original spatial resolution (from 30 arcsec to 3 arcsec) without modifying the original information.

The generation of the data layer of the ISRIC soil grids was made by 2D or 3D regression with splines to model soil properties and to create a multinomial logistic regression for the soil classes (see also http://www.isric.org/content/faq-soilgrids#How_were_the_spatial_predictions_generated for further information).

The SGT estimate of SOC stock was compared to those of the GSOC and IRSIC after extracting the estimates with each methodology on the same location of the samples analysed and thus computing the mean absolute error (n=2202), as suggested by Bennett et al. (2013). This comparison was also carried out at a land cover class level to show a general pattern of SOC stock by SGT, GSOC and ISRIC.

4.3 Results

4.3.1 SGT modelling

We predicted the SOC stock and its distribution across Sicily, including an assessment of the prediction error. The latter was performed using a multi-fold technique by randomly resampling the original dataset into ten replicates. We also estimated the SOC stock for different land covers. The mean SOC with SGT for the whole area under study was 37.44 t ha^{-1} using the total pixel estimates (n=2228754, 7225 m^2 per pixel) and 38.88 t ha^{-1} by means of the pixel estimates at the sampling points (n=2202, 1 profile/7.31 km^2 on average).

The five most relevant contributors from the initial 14 predictors (Fig. 3) were 1) annual rainfall, 2) soil texture, 3) land cover (CORINE), 4) mean annual temperature, and 5) Band 8. In particular, soil texture, land use, temperature and band 8 showed 28.2%, 32.0%, 49.3% and 49.6% less importance compared to annual rainfall in the prediction step. The mean RCs in decreasing importance (from left to right and from up to down) is shown in Fig. 4. The influence of rainfall on mean SOC stock estimates were maximised when the rainfall was in the range of 600-650 mm y^{-1} (up to +26%) and was negative in this range. In particular, the dependence of the model upon mean rainfall decreased due to the decreasing rainfall from 600 to 400 mm y^{-1} (up to −5%). Similarly, the positive and relatively strong influence of soil texture on SOC stock estimate was found for the clay, silty clay loam and sandy loam textures. These textures represented 13.3%, 4.0% and 16.2% of the total SOC dataset, respectively. The CORINE land cover classes with the highest positive influence on SOC stock were the complex cultivated pattern (CORINE code 242) and lands principally occupied by agriculture with significant areas of natural vegetation (CORINE code 243). Non-irrigated lands (CORINE code 211, mostly field crops including legumes, durum wheat and other cereals) negatively correlated with the SOC stock content. The CORINE land covers 242, 243, 211, accounting for 8.5%, 4.4% and 39.6%, respectively, of the total SOC stock in the observed value dataset.

The effect of the mean annual temperature on SOC stock estimate was high and positive up to 15 °C, whereas its contribution strongly decreased from 15 °C to 16 °C and was mildly negative for the higher temperatures (+18 °C, −2% RC). Similarly, the Landsat ETM+7 panchromatic Band 8 positively affected the SOC stock estimate at values lower than 50 W $(m^2\ sr\ \mu m)^{-1}$ and negatively affected the estimate at values higher than 60 W $(m^2\ sr\ \mu m)^{-1}$.

Fig. 5 shows each realization of the model (Fig. 5A, upper panel), the coefficient of the variation of the 10 models constructed from the 2202 data for the SGT on the sampling location (Fig. 5B) and the

standard deviation of the models on all pixel estimates (n=2228754, Fig. 5C). On average, the whole suite robustly fitted the range between 20 and 70 t ha^{-1}. Values lower than 20 t ha^{-1} and higher than 70 t ha^{-1} (19.1% and 10.0% of the observed values, respectively) were poorly estimated, which is likely due to the low number of data points in these ranges. The model robustness was also evaluated by plotting the coefficient of variation of the 10 realizations of SGT that was extracted on the 2202 sampling locations (Fig. 5B). The coefficient of the variation of the SGT model was lower than 11.2% in the 0-0.1 percentile, lower than 13.6% in the 0.1-0.5 percentile, lower than 23.0% in the 0.5-0.9 percentile, and lower than 12.2% in the 0.9-1.0 percentile. In addition, the standard deviation of the model realizations against the predicted SOC (Fig. 5C) was always lower than 11 t ha^{-1} and was, on average, lower than 5 t ha^{-1}.

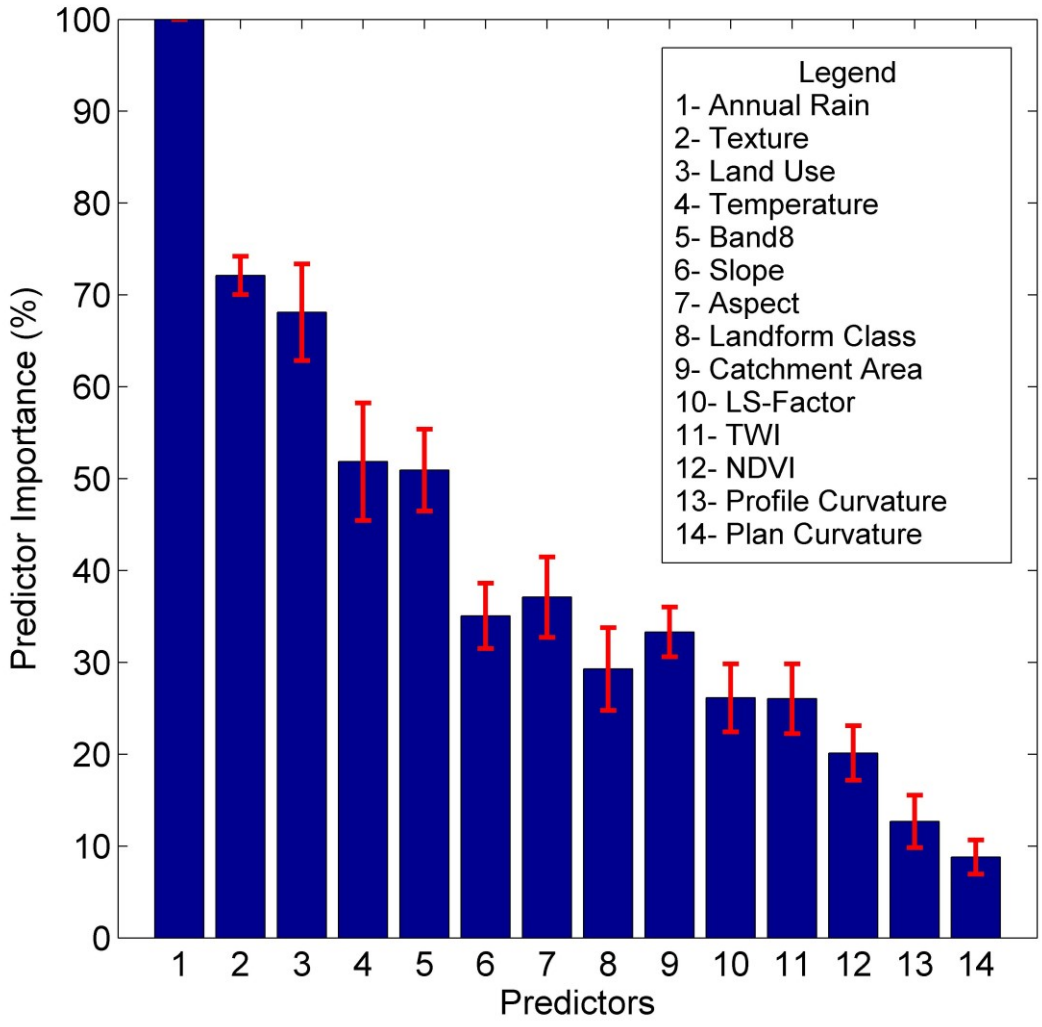


Fig. 3. The i*mportance of each* of the 14 predictors *within* the *stochastic gradient treeboost model to estimate the soil organic carbon stock of the agricultural land of Sicily, Italy.*

4.3.2 SGT output and comparison with GSOC and ISRIC

The SOC stock predicted by SGT, GSOC and ISRIC were plotted against the observed SOC stock (Fig. 6; see supplementary material for a map of the prediction of the SOC stock in the area under study by means of SGT, GSOC, and ISRIC). Overall, SGT R^2 was 0.470, GSOC R^2 was 0.034 and ISRIC R^2 was 0.127 (Table 2). This mostly depended on the better prediction of the samples by SGT compared to GSOC and ISRIC (Table 2). The SGT resulted in better prediction compared to GSOC and ISRIC for most of the data, which resulted in SGT-predicted SOC stock values in the range of ±50% higher than GSOC and SGT (65.8% and 50.7%, respectively) compared to the observed value (71.4%). The mean absolute error of SGT was 14.84 t ha^{-1}, whereas those of the GSOC and ISRC were 18.19 t ha^{-1} and 23.28 t ha^{-1}. The intercept of SGT was 21.6% and 45.5% lower than GSOC and ISRIC (Table 2), respectively. On the contrary, the slope (β) of the pseudo-regression for SGT was closer to 1 and, in particular, were 459% higher than GSOC and 64% higher than ISRIC. On average, the SOC stock per unit area estimated by SGT was 2.10 t ha^{-1} (+5.9%) more than GSOC (Table 3) and 17.31 t ha^{-1} (–31.6%) less than ISRIC. The standard deviation of the SOC stock estimate per unit area (n=2228754) by means of SGT was 29.0% and 91.5% less than GSOC and ISRIC, respectively. The differences between SGT and GSOC or SGT and ISRIC in terms of the prediction within each land use class varied between –39.0% (corresponding to –31.3 t SOC stock ha^{-1} in land, principally agriculture with natural vegetation in ISRIC, which is 4.6% of the area under study) to +42.4% (corresponding to +14.8 t SOC stock ha^{-1} in land principally occupied by agriculture with significant areas of natural vegetation) in SGT compared to GSOC. Notably, ISRIC strongly overestimated the C stock on all land uses (on average, +46.2%). Such differences lead to a 5.9% higher SOC stock estimation in SGT compared to GSOC (+3401 Mt) and a 30.6% lower SOC estimation in SGT compared to ISRIC (–26973 Mt). In general, the standard error within each land use class obtained through SGT modelling was lower than those obtained by GSOC or ISRIC and ranged from −15.9 t ha^{-1} (land, principally agriculture with natural vegetation in SGT, as compared to ISRIC) to +3.4 t ha^{-1} (vineyards, 9.6% of the area under study, in SGT compared to GSOC). This also resulted in 31.8% and 87.5% lower coefficients of variation, on average, in SGT compared to GSOC and ISRIC, respectively, when estimating the mean SOC of each land use class.

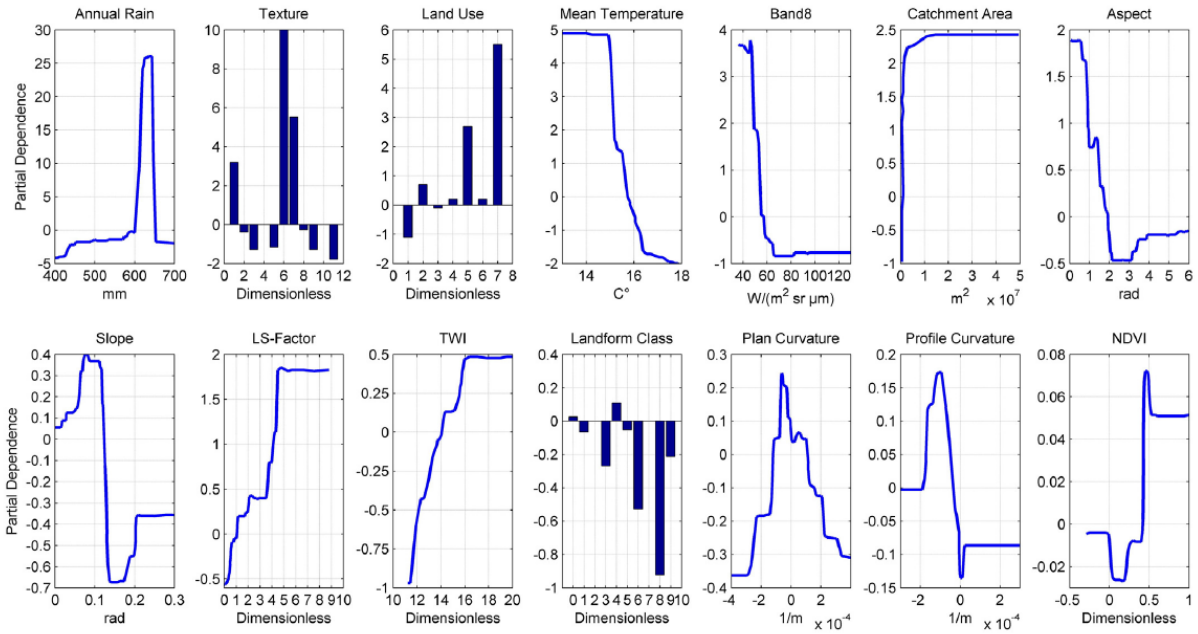


Fig. 4. *Partial dependence (response curve plots) of the model on the 14 predictors included in the stochastic gradient treeboost model of soil organic carbon stock of the agricultural land of Sicily, Italy. See Table 1 for Texture Class, Land Use and Landform codes.*

4.4 Discussion

The mean prediction of SOC stock by SGT amounted to 37.44 t ha^{-1}. Values below 20 t ha^{-1} were predicted in a limited number across the ten replicates but similarly occurred in each replicate. On the contrary, SOC stock values higher than 70 t ha^{-1} frequently occurred among replicates, and their variance was high. This resulted in R^2=0.47 of the SGT model. Similar or lower fit statistics were found when other BRT algorithms were applied at a regional scale in France (Martin et al., 2014), Indiana (Mishra, U., et al., 2009), Nigeria (Akpa et al., 2016), or Western Ghats (India) (Seen et al., 2010).

The general pattern of SOC stock obtained by SGT was similar to those of the GSOC and ISRIC estimates. However, both GSOC and ISRIC pseudo-regressions resulted in lower βs and higher intercepts than STG. In addition, the SGT classification of samples in the range comprised ±50% compared to those observed, which occurred +5.7% and +20.7% more than GSOC and ISRIC. This suggest that these databases may be unsuitable for structural decision-making measures to increase SOC stock in soils with low SOC stock when applying rules and decisions using a Mediterranean scale. The SGT individuated annual average rainfall and temperature as fundamental factors. It has been argued that the importance of temperature on SOC stock is high when the mean rainfall is adequate to sustain the microbial activity dealing with C turnover and when the variation of temperature is high (Ma et al., 2014). In such conditions, lower temperatures could constrain the SOC

turnover (Davidson and Janssens, 2006; Phachomphon et al., 2010). However, in the present study area, such conditions mostly occurred in mountainous areas, which were scarcely sampled due to the lack of cultivation. In addition, Davidson and Janssens (2006) also showed that other environmental constraints, including the SOC fixing in clay and nitrogen and the lignin content of both plant-derived and soil organic carbon, can strongly impair the temperature sensitivity of the soil organic carbon. This can explain why we found high partial dependence of SGT on areas with more than 650 mm annual rainfall. Areas with less than 650 mm are mostly non-irrigated lands and account for 51.2% of the entire area included in the model. Indeed, soils in these areas frequently have high clay content and are cultivated with durum wheat, whose plant residue has a very high C:N ratio. This agrees with the relatively high partial relationship of SOC with clay and loamy textures (Clay, Silty-Clay-Loam, Sandy-Loam). Moreover, it also agrees with other studies highlighting the strong SOC protection ability of clays (Six and Paustian, 2014; Velthof et al., 2002; Wei et al., 2014). In addition, Martin et al. (2011, 2014) suggested that the clay content is a crucial trait for the outcome of the BRT modelling. This explains why the SGT strongly took into account the texture information to draw the model.

Despite the importance of topographic indexes in determining soil erosion and thus the removal of the high C layers, we did not find a high partial dependence of the model on slope, aspect, landform, or the catchment area (with PIs <35%). Other experiments showed that topographic indexes are likely to explain most of the variation of SOC stock in the topsoil (0-10 cm depth) layer, whereas they are less important in the subsoil (10-50 cm depth) (Grimm et al., 2008). It was also suggested that SOC variation mostly depends on climatic variations, which in turn vary in large rather than short-scale covariates, including soil depth and texture (Vaysse and Lagacherie, 2015). In the present experiment, the 'Landsat panchromatic band 8' was more predictive compared to NDVI, which was also included as explanatory variables in the model. This could be due to both a higher resolution and range across the visible band of the band 8 compared to NDVI. Indeed, the band 8 includes data from the red, green, and blue portions of the electromagnetic spectrum, whereas NDVI only takes into account of the ratio of band 3 and 4, which span across the red and near IR portions, respectively. In other experiments, the contribution of the NDVI in the variable importance was lower than other Landsat bands (Akpa et al., 2016), where the NDVI explained less than 5% of the SOC stock variation. The influence of the land use on SOC stock could be attributed to the different disturbance level associated with the agricultural management (Söderström et al., 2014) and the potential to have high erosion rates due to the climate or management (Nadeu et al., 2015).

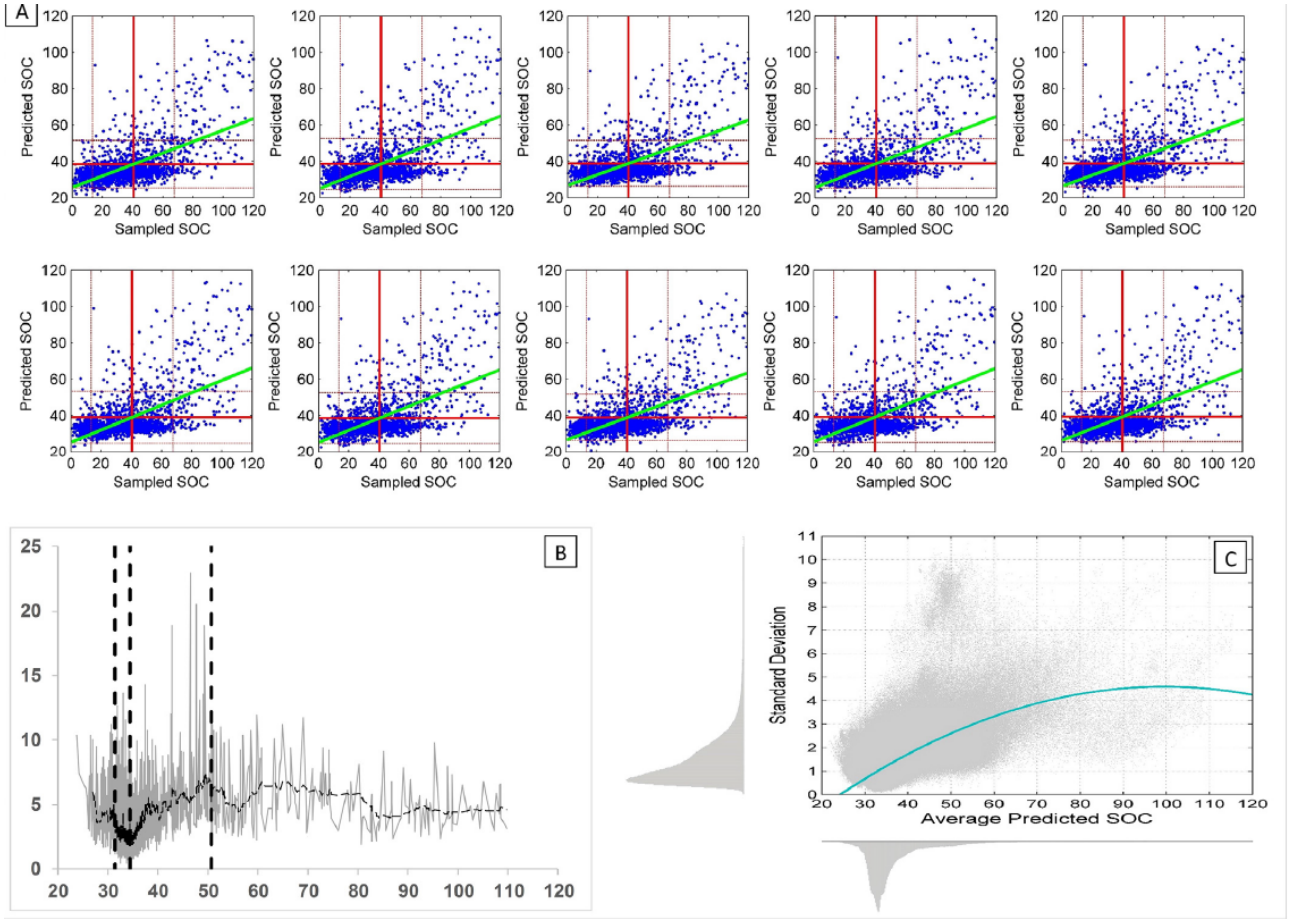


Fig. 5. *Replications of the model and uncertainty. Panel A: each scatter represents the outcome of each model replicated. Red vertical and horizontal thick lines represents means that are observed and predicted by* the *stochastic gradient treeboost, respectively. Dashed lines are means ± standard deviations. Panel B:* Scatter plot of the *coefficient of variation (grey line) of the models (%) against the SGT-predicted mean. Dashed vertical bold lines represent the 10th, 50th and 90th percentiles of the distribution of the SGT-predicted values; the dashed black line interpolating the CV represents the mobile media at n=22. Panel C: Scatter plot of the mean SOC stock prediction against its own standard deviation among pixel replicates (n=2.3 x10^{6} pixels).*

4.5 Conclusions

Monitoring and modelling spatial distribution of SOC, as input for global carbon cycles studies and guiding decision-making processes, is presently a global challenge that needs to involve environmental scientists, socio-economists and policy makers for the coming decades (Galati et al., 2016). The Mediterranean climatic context plays an important role in the decay processes of organic residue, and climate change in the Mediterranean area is likely to affect SOC distribution and, consequently, both the potential yield of crops and the biodiversity of natural and cultivated areas. Thus, a fine estimation of the SOC stock is crucial to structure efficient public supporting measures. We showed that such an estimation relies on accessible data (including rainfall, temperature and RS data) and detailed information on soil texture and land cover. These features should be coupled with a geo-referenced model that is able to correct for spatial traits that, if ignored, can mislead the

estimation (including Catchment Area, Slope and LS-factor) (Phachomphon et al., 2010). A limit to developed well-fitting models is represented by the lack of sufficient soil data, their non-Gaussian distribution among the predictors, the spatial resolution of the predictors and a correction for outliers. Nonetheless, when correctly integrated, such as in the SGT model of the present study, such data can yield a reliable estimation of SOC stock.

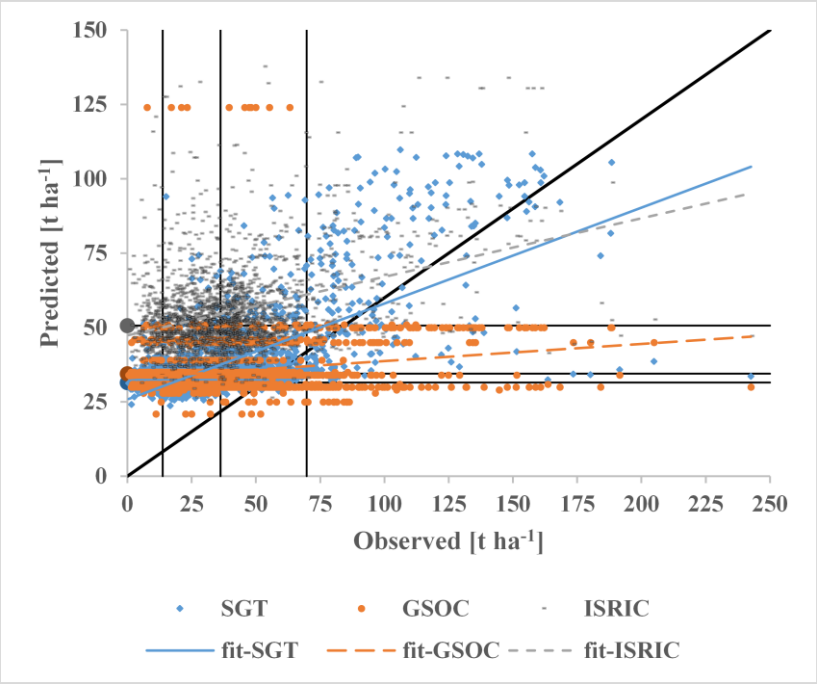


Fig. 6. *Scatter plot of the predicted against observed SOC stock values by means of a stochastic gradient treeboost (SGT, black circles) and Global Soil Organic Carbon Estimates (GSOC, grey crosses) (Hiederer and Köchy, 2012). The R^2 for SGT, GSOC, and ISRIC were 0.47 (y = 0.32x + 25.81), 0.034, and 0.127, respectively. Vertical dashed lines represent the 10^{th}, 50^{th}, and 90^{th} percentiles of the distribution of the observed values (13.7 t ha^{-1}, 36.2 t ha^{-1}, and 69.7 t ha^{-1}, respectively). Horizontal dashed lines represent the 10^{th}, 50^{th}, and 90^{th} percentile of the distribution of the SGT-predicted SOC stock (31.5 t ha^{-1}, 34.5 t ha^{-1}, and 50.6 t ha^{-1}, respectively). The bold dashed line represents the y=x function. Please note that observed unit in abscissa is double that in the ordinate axis.*

Further improvements are expected when adopting predictors with greater spatial resolution. The results of the present experiment also yield valuable information for assessing the effect of a climate change scenario on SOC stocks and their spatial distribution.

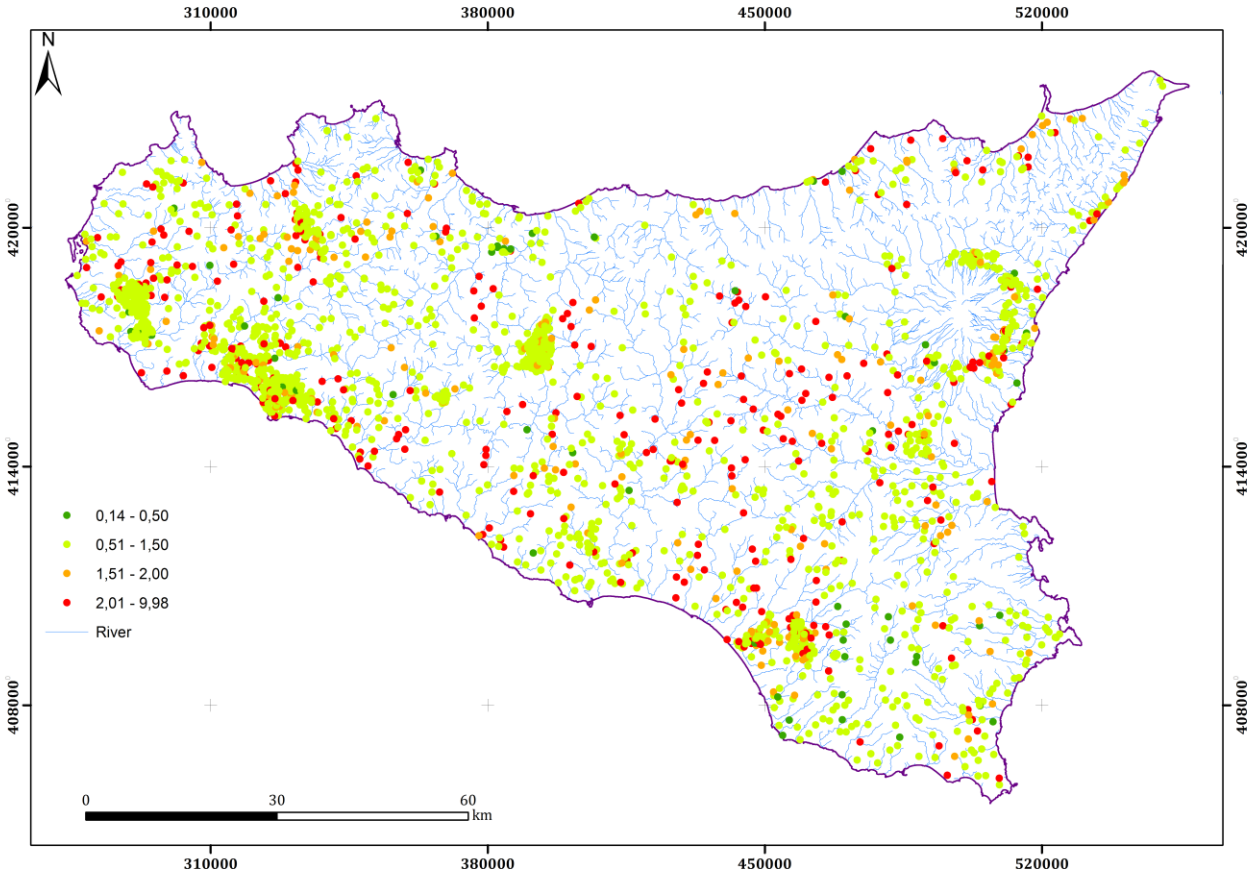


Fig. 7. *Prediction confidence map. Each point represents the ratio between SGT-predicted and observed values. The closer the ratio is to 1, the higher its representation of the observed value is. The number of points (compared to the total point, n=2202) enclosed in the 0.5-1.5 range (predicted SOC corresponding to ±50% compared to the observed value) were 71.4%, 65.8%, and 50.7% for SGT, GSOC, and ISRIC, respectively.*

Acknowledgements

The authors would like to thank Maria Gabriella Matranga, Vito Ferraro from the Regional Bureau for Agriculture, Rural Development and Mediterranean Fishery, the Department of Agriculture, Service 7 UOS7.03 Geographical Information Systems, Cartography and Broadband Connection in Agriculture, Palermo. The authors also thank the Joint E-i-C and two anonymous reviewers for their constructive comments, which helped to improve the present manuscript.

	Obs	***SGT***	***GSOC***	***ISRIC***
intercept	-	25.811	32.917	47.318
β	-	0.323	0.058	0.197
R^2	-	0.470	0.034	0.127
% of points classified in the following 'predicted to observed' ratios				
<0.5	-	2.6	9.7	1.2
0.5-1.0	-	46.0	43.5	20.3
1.0-1.5	-	25.4	22.3	30.4
1.5-2.0	-	10.3	10.0	17.4
2.0-2.5	-	4.6	4.4	8.8
2.5-10	-	9.3	9.4	20.4
>10	-	0.0	0.8	1.5
percentiles				
0.05	9.50	30.11	30.00	37.71
0.10	13.68	31.48	30.00	41.65
0.25	23.28	32.95	31.00	46.57
0.50	36.33	34.57	34.00	52.21
0.75	49.54	38.00	35.00	59.70
0.90	70.01	50.67	45.00	73.85
0.95	91.51	69.62	50.00	84.39
1.00	242.57	109.77	124.00	137.72

Table 2. Intercept, *β, and* R^2 *of the pseudo-regression of the modelled (by SGT, GSOC, and ISRIC) data against the observed (Obs) data, % classification of sampled points according to the 'predicted to observed' ratio of each model and main percentiles of the observed and modelled distributions.*

Land Use Classes	Corine Code	Area		SGT SOC	GSOC	ISRIC	Total SOC Stock			Difference in mean SOC				Difference in SOC Standard Deviation				Difference in total SOC Stock			
				Mean ± SD	Mean ± SD	Mean ± SD	SGT	GSOC	ISRIC	SGT-GSOC		SGT-ISRIC		SGT-GSOC		SGT-ISRIC		SGT-GSOC		SGT-ISRIC	
		[ha × 1000]	[%]	[t ha^{-1}]	[t ha^{-1}]	[t ha^{-1}]	[Mt]			[t ha^{-1}]	%	[t ha^{-1}]	%	[t ha^{-1}]	%	[t ha^{-1}]	%	[Kt]	%	[Kt]	%
Non-irrigated arable	**211**	791.97	48.9	34.9 ± 5.90	35.6 ± 10.90	49.6 ± 9.42	27.62	28.22	39.28	-0.8	-2.1	-14.7	-29.7	-5.0	-45.9	-3.5	-37.4	-602	-2.1	-11665	-29.7
Vineyards	**221**	155.54	9.6	39.7 ± 4.20	35.3 ± 0.80	52.7 ± 8.94	6.17	5.49	8.20	4.4	12.5	-13.0	-24.7	3.4	425.0	-4.7	-53.0	687	12.5	-2029	-24.7
Fruit and berry	**222**	154.58	9.5	39.1 ± 11.10	36.2 ± 15.50	57.8 ± 17.81	6.04	5.59	8.94	2.9	8.0	-18.8	-32.4	-4.4	-28.4	-6.7	-37.7	448	8.0	-2901	-32.4
Olive groves	**223**	219.35	13.6	39.1 ± 6.50	34.3 ± 6.90	58.1 ± 12.77	8.58	7.52	12.74	4.8	14.0	-19.0	-32.7	-0.4	-5.8	-6.3	-49.1	1055	14.0	-4166	-32.7
Annual with permanent crops	**241**	108.24	6.7	41.1 ± 7.90	40.1 ± 11.37	57.3 ± 11.90	4.45	4.34	6.20	1.0	2.6	-16.2	-28.2	-3.5	-30.5	-4.0	-33.6	113	2.6	-1749	-28.2
Complex cultivation patterns	**242**	114.33	7.1	41.1 ± 9.40	35.8 ± 7.90	59.7 ± 15.87	4.70	4.09	6.83	5.3	14.9	-18.6	-31.2	1.5	19.0	-6.5	-40.8	611	14.9	-2126	-31.2
Land principally occupied byagriculture with natural vegetation	**243**	74.68	4.6	49.0 ± 11.10	34.4 ± 8.80	80.3 ± 27.01	3.66	2.57	6.00	14.6	42.4	-31.3	-39.0	2.3	26.1	-15.9	-58.9	1089	42.4	-2337	-39.0
	total	**1618.69**		**37.44 ± 1.24**	**35.97 ± 1.74**	**54.75 ± 14.54**	**61.21**	**57.81**	**88.19**	**1.47**	**4.1**	**-17.31**	**-31.6**	**-0.51**	**-29.0**	**-13.31**	**-91.5**	**3401**	**5.88**	**-26973**	**-30.59**

Table 3. *SOC stock per unit*

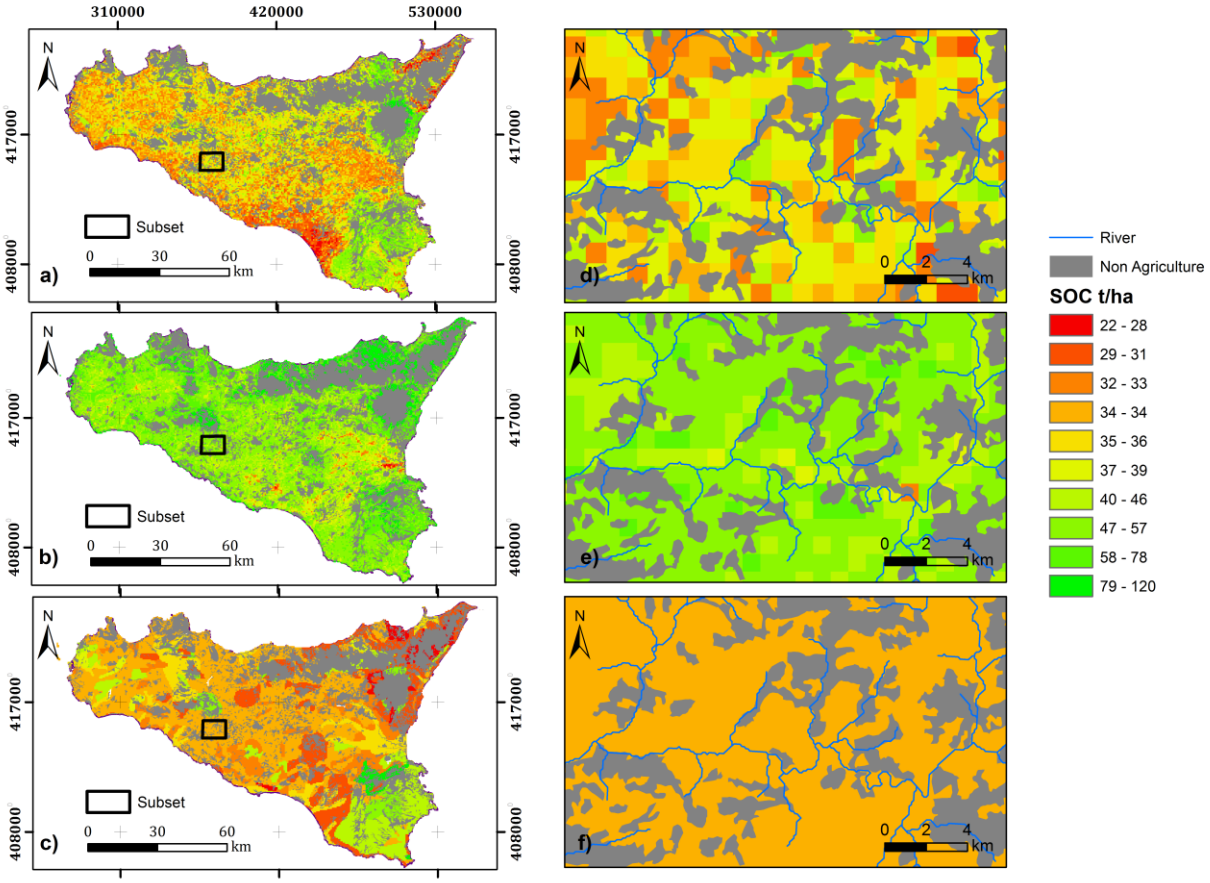


Fig.1 *Suppmat, Prediction map of the soil organic carbon Stock by means of the SGT (a, d), GSOC (b, e), and ISRIC (c, f) in the agricultural areas of the whole area under study, Sicily, Italy (a, b, c) and on a sub-area within the area under study (d, e, f). Outputs of each model are shown at a 1-km^2 spatial resolution.*

Chapter 5-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

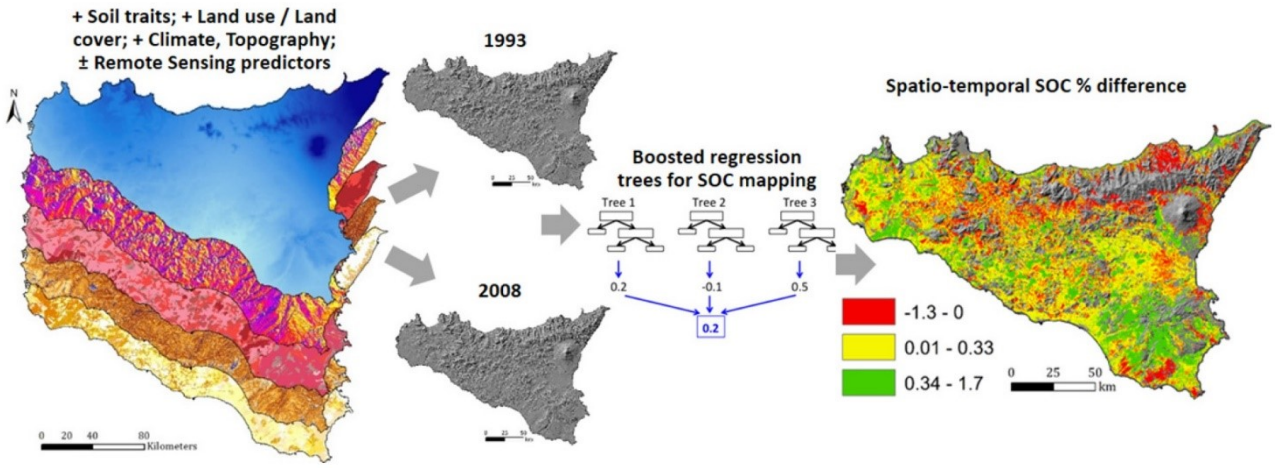
From: Schillaci, C., Acutis, M., Lombardo, L., Lipani, A., Fantappiè, M., Märker, M., Saia, S., 2017a. Spatio-temporal topsoil organic carbon mapping of a semi-arid Mediterranean region: The role of land use, soil texture, topographic indices and the influence of remote sensing data to modelling. Sci. Total Environ. 601–602, 821–832. doi:10.1016/j.scitotenv.2017.05.239

Keywords: SOC mapping, Space-time SOC variation, Agro-ecosystems, R programming

Abstract

SOC is the most important indicator of soil fertility and monitoring its space-time changes is a prerequisite to establish strategies to reduce soil loss and preserve its quality. Here we modelled the topsoil (0-0.3 m) SOC concentration of the cultivated area of Sicily in 1993 and 2008. Sicily is an extremely variable region with a high number of ecosystems, soils, and microclimates. We studied the role of time and land use in the modelling of SOC, and assessed the role of remote sensing (RS) covariates in the boosted regression trees modelling. The models obtained showed a high pseudo-R^2 (0.63-0.69) and low uncertainty (s.d. < 0.76 g C kg^{-1} with RS, and < 1.25 g C kg^{-1} without RS). These outputs allowed depicting a time variation of SOC at 1 arcsec. SOC estimation strongly depended on the soil texture, land use, rainfall and topographic indices related to erosion and deposition. RS indices captured one fifth of the total variance explained, slightly changed the ranking of variance explained by the non-RS predictors, and reduced the variability of the model replicates. During the study period, SOC decreased in the areas with relatively high initial SOC, and increased in the area with high temperature and low rainfall, dominated by arables. This was likely due to the compulsory application of some Good Agricultural and Environmental practices. These results confirm that the importance of texture and land use in short-term SOC variation is comparable to climate. The present results call for agronomic and policy intervention at the district level to maintain fertility and yield potential. In addition, the present results suggest that the application of RS covariates enhanced the modeling performance.

Keywords: SOC mapping, Space-time SOC variation, Agro-ecosystems, R programming, Digital soil mapping, Legacy dataset.



graphical abstract

5.1 Introduction

Agricultural lands play a major role in the storage of soil organic carbon (SOC) and sequestration/release of atmospheric CO_2 (Bradford et al., 2016; Filippi et al., 2016; W M Post, 2000). SOC is directly linked with a number of ecosystem services and agronomical benefits and is the main driver of soil fertility. However, agricultural soils have been depleted from their original SOC stock due to cultivation, which also negatively affected soil aggregation status, water infiltration rate, soil fertility and biota (Bruun et al., 2015; Parras-Alcántara et al., 2016; Saia et al., 2014). The preservation of soil quality is a priority to maintain agricultural productivity and environmental quality. In this framework, monitoring SOC concentration and stock changes through space and time is important to establish strategies to reduce soil loss and preserve its quality. SOC monitoring at regional scale relies on sparse sampling and application of an estimation process. Such a process should take into account the spatial interdependence of samples and abundance of predictors (Martin et al., 2014); and the distribution heterogeneity in space and among determinants (predictors) of SOC accumulation (Lacoste et al., 2014). With regards to the latter, the relationship in the domain of each predictor with SOC and the resolution of the predictors is particularly relevant for any spatial estimation (Miller et al., 2016; Miller et al., 2015a,b). The spatial estimation of SOC concentration and stocks is commonly performed by statistical approaches (Meersmans et al., 2009; Orton et al., 2014) with different interpolation methods and machine learning predictive models (Henderson et al., 2005; Yang et al., 2015). The former is better suited to areas with dense SOC measurements, whereas the second is more appropriate for non-regularly sampled regions, since its outcome does not rely on the sample proximity to extract functional (ecological) relationships between dependent and independent variables.

SOC dynamics under different land uses are still poorly understood (Francaviglia et al., 2017b; Meersmans et al., 2008; Purton et al., 2015), especially when deriving data from wide areas and with

different climates. In Mediterranean environment, lack of knowledge on SOC dynamic is further due to variable climatic and erratic meteorological conditions. It has been shown that cultivation exerts a negative role on SOC accumulation in various environments (Francaviglia et al., 2017b; Kämpf et al., 2016; Novara et al., 2013) and this likely depends on both soil tillage and reduction of biomass return to the soil. In particular, a reduction of the tillage intensity can favor SOC accumulation irrespective of aridity (from semi-arid to humid) and can be up to 1 t SOC ha^{-1} yr^{-1} (Conant et al., 2001; Kämpf et al., 2016; Kurganova et al., 2014; W M Post, 2000). The SOC dynamic also depends on other factors such as soil genesis and type, land use history and management and useful information could be gained from SOC spatial models (Badagliacca et al., 2017; Martin et al., 2014;; Schillaci et al., 2017b, 2015; Vereecken et al., 2016).

In the last two decades the integration of physical, chemical, and biological information derived from different covariates in the models has boosted the studies on soil properties (Bui et al., 2009; Henderson et al., 2005) and also for SOC mapping from global or continental (Hengl et al., 2014; Lugato et al., 2014a) to regional and plot scales (Akpa et al., 2016; de Gruijter et al., 2016; Martin et al., 2014; Schillaci et al., 2017b). SOC mapping attempts at giving an image of the spatial distribution despite it is costly (Minasny et al., 2013 and reference therein).

The most recent developments in the digital soil mapping include machine learning (Forkuor et al., 2017; Gasch et al., 2015; Hengl et al., 2017) to study space-time variation of soil properties and use of remote sensing (RS) covariates (Castaldi et al., 2016a). Thanks to their high accessibility, resolution and availability for wide areas, RS data gained importance for spatial prediction of the topsoil organic C (Bou Kheir et al., 2010; Poggio et al., 2013). For example, Bou Kheir et al. (2010) found that the construction of SOC maps with a classification-tree analysis by the sole RS parameters gave the same accuracy of a model built with sole digital elevation model (DEM) parameters, and both of them had sole ca. 10% less accuracy that a full RS+DEM+soil parameters model built. Poggio et al. (2013) found that integration of RS with terrain attribute data increased the predictive ability comparing to the model built with only terrain parameters. However, some of the SOC estimates lack uncertainty analysis and this compromises the reliability of predictions for decision making (Maia et al., 2010; Minasny et al., 2013; Ogle et al., 2010). In addition, Conant et al. (2011) highlighted the limitation to document time changes in SOC because of the spatial variability in the factors that influence SOC distribution.

In a regularly-spaced data collection, SOC samples are taken from representative or random sampling sites in a given study area. Legacy data comes from a mixture of sampling campaign resulting in data collected for different aims (Chartin et al., 2017), which frequently allow to make predictions for areas with sampling limitations (Rial et al., 2017b). Depending on the scope of each survey (e.g.

regional soil characterization or precision agriculture) sample density can change abruptly. This can consist in drawbacks including their non-regular distribution in space, which call for the use of particular modelling method and predictors. Due to these difficulties, only few examples on mapping at regional extent with legacy data are available. For example, Ross et al. (2013) and Grinand et al. (2017) carried out a space-time assessment of SOC in subtropical regions of south-eastern United States and Madagascar, respectively.

Little information is available on SOC dynamics in semi-arid Mediterranean areas due to the unavailability of consistent databases. Nonetheless, time dynamic of SOC storage in the soil is highly dependent to the climatic zone of the area under study (Doetterl et al., 2015b). In addition, spatial and time change of SOC can respond to different determinants at varying the climate of area under study. The present work fits within the big picture of spatial SOC mapping and time change. This was made by means of a legacy dataset and use of remotely sensed data. In particular, we used legacy data of two sampling campaigns 15 years apart (1993-2008,), coupled with climate (from Worldclim data Bio1,12), and land use information (from CORINE 1990-2006) to map the topsoil SOC variation across time in the agricultural area of a semi-arid Mediterranean region (Fig. 1). Such aim was achieved by applying a machine learning method, namely boosted regression trees (BRT), to each sampling campaign dataset using land use, soil texture, topographic and remote sensing predictors. We also tested the role of remote sensing covariates in the spatial SOC prediction and predictors' importance by running each model either with or without the implementation of the RS covariates. In the area under study, i.e. cropped field in which plants (mostly field crops) have limited or no growth during summer and early fall, the inclusion of remote sensed variables could capture part of the SOC variation due to biomass return to the soil.

5.2 Material and methods

5.2.1 Study area

The study area, Sicily (Italy), is a semiarid region located in middle of the Mediterranean Sea (Fig. 1). Its area is about 25,286 km^2. Approximately 60% of the Sicilian territory is cultivated. The macroclimate of the region is Mediterranean with three main bioclimatic areas: thermo-, meso-, and supra-Mediterranean. Mean annual temperatures in the cropped area range from 7 °C to 15 °C and mean annual precipitation from 350 to 1000 mm, whereas mean annual temperatures and rainfall in the natural, uncropped area can be 1.8 °C and up to 1300 mm (Cannarozzo et al., 2006; Viola et al., 2014). The main annual crops are durum wheat, winter-seeded barley, pulses and forage legumes and a wide range of horticultural crops; the main perennial crops are olive groves, vineyards and fruit trees such as citrus, almonds, and stone fruits. Woodlands and secondary forests are not targeted by

the SOC concentration mapping in the present work, except those areas in which agriculture abandonment occurred.

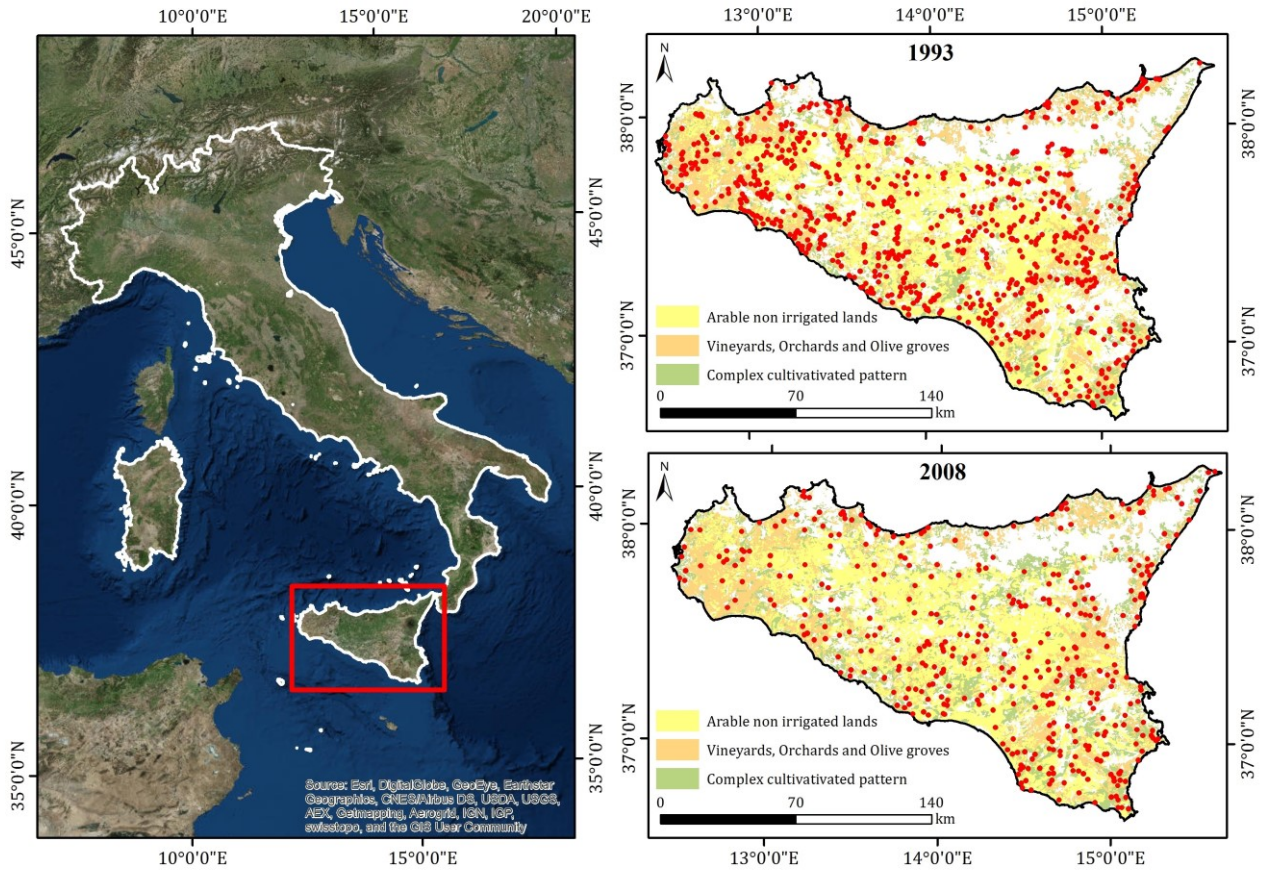


Fig. 1. *Locations of the sampling sites in the 1993 and 2008 in the area under study (Sicily). Land use groups used in the study are displayed.*

Adoption of conservation soil management techniques is almost absent (Ruisi et al., 2014). In the region, different soil survey campaigns were undertaken between 1968 and 2008. The criteria for the selection of the locations of the soil sampling are explained in the next section. The island has a great pedoclimatic variability: dominant soils according to the World Reference Base for soils are Calcaric Regosols, Haplic Calcisols, Calcic Vertisols, Vitric or Silandic Andosols, Calcaric and/or Mollic Leptosols, Calcaric Phaeozems, and Fluvic Cambisols. Hence it can be considered quite representative of most of the Mediterranean countries. A number of ecological and anthropic traits make Sicily unique for ecological studies. These traits include a relatively high population density and degree of cultivation, an ancient environmental history, climatic variability, land uses and several dominations from different populations, which introduced various plant species and management techniques. All these factors made Sicily an open and extremely variable laboratory for the study of the impact of anthropic pressure and environmental variation at microscale, land cultivation and management on other environmental traits, including SOC distribution and dynamics. Such

characteristics strongly help in the exportation of the results of environmental studies to other similar and different environments and scale, such as also suggested by others (Legendre and Legendre, 1998; Novara et al., 2017; Schmolke et al., 2010).

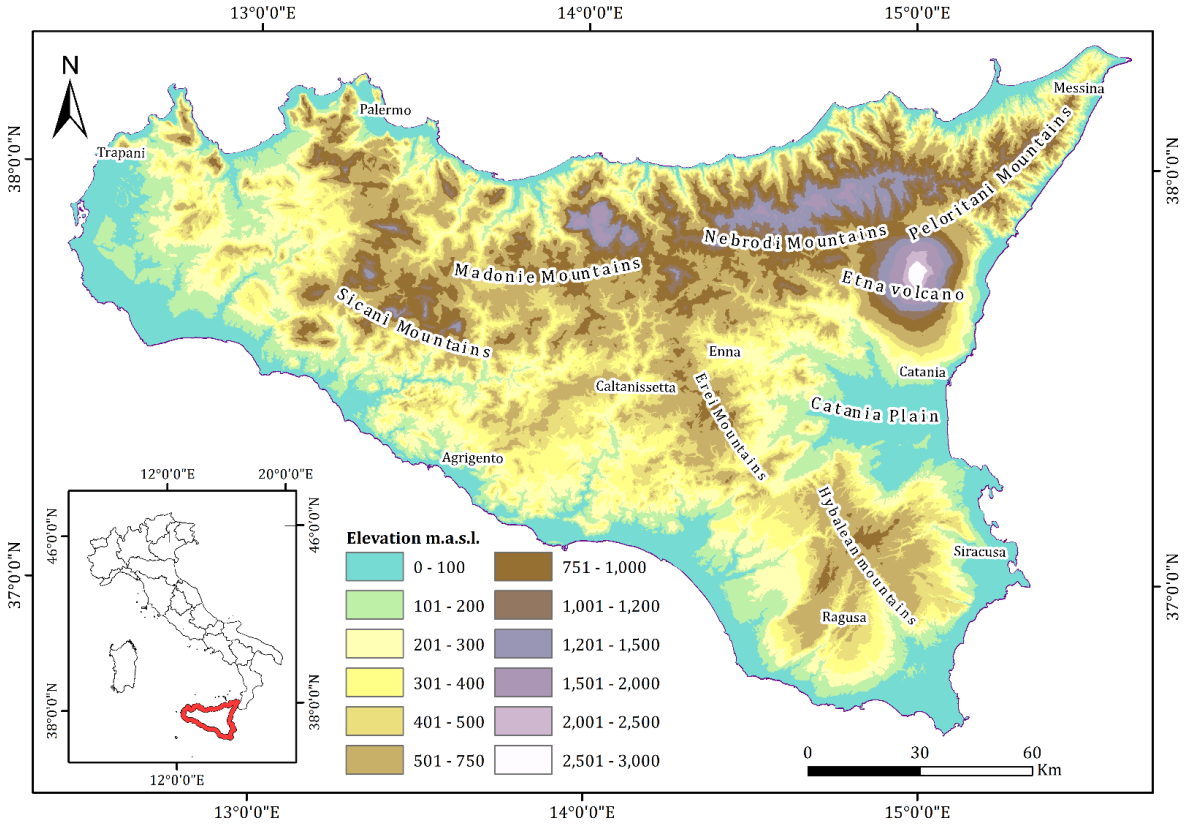


Fig.1 **Supplementary material.** *Physiographic map of the area under study.*

The region under study, Sicily (see Supplementary material Fig. 1 for a physiographic map of the area with orography and toponymy information used), is a setting of different agro-ecosystems and natural environments though it is mainly semi-arid and with few incidence of forestlands. The island has three main, almost continuous, mountain chains: Peloritani from the north-eastern corner moving to west few km down the northern coast, followed by the Nebrodi and then by the Madonie. In the western/central part of the island there is an irregular mountain area: the Sicani, somehow continuing the ridge formed by the previous mountain chains. Mean height of the mountain chains decreases from east to west. These chains were formed as part of the Apennines, which span across the island as a geological bridge between peninsular Italy (on the east end) and Tunisia (on the west end). The highest mountain of Sicily is the Etna Volcano (about 3600 m above sea level [a.s.l.]), located in the northeastern part of the region, south of the Peloritani. To the south of the Etna Volcano, a wide plain (the Catania plain) is formed by the alluvium of the Simeto River, south of which there is the expansion of a hilly to mountainous area: the Hyblaean mountains/plateau. The rest of the core of the island, from the plain of Catania to the Erei Mountains and cities of Enna, Caltanissetta and Agrigento

is a mostly hilly area with clayey, high pH, seldom gipsic saline soils. Such as for the main mountain chains, mean height of this latter ridge decreases from east to west. Other minor plains can be retrieved all along the coasts. All the rivers, with the exception above-mentioned Simeto, have a strong seasonal flow. This is due to the low rainfall south of the Apennines ridge, or low basin extent north of it.

5.2.2 SOC dataset

The Regional Bureau for Agriculture, Rural Development and Mediterranean Fishery, the Department of Agriculture, and Service 7 UOS7.03 provided the legacy dataset used in this study. The surveys that produced the legacy dataset had different aims (such as redaction of suitability or pedological maps). SOC, soil texture, actual land use, GPS positioning and relative metadata were measured in every survey and provided for the present work. From the complete record of observation (about 2700 different locations in a timespan of 30 years), we selected the years with the most of samplings, which were 1993 (685 points) and 2008 (337 points) (Fig.1). The 1993 database is a regional subset of the national soil survey performed in the framework of the AGRIT project of the Italian Ministry of Agriculture and Forestry (MIPAAF), all over Italy in the years 1993 to 1994. The 2008 campaign (undertaken in the frame of the project "Soil Map of Sicily at 1:250,000 scale") was aimed at closing the gap of previous campaigns basing on a GIS oriented pedo-landscape sampling design (Fantappiè et al., 2011b). Only SOC data sampled in agricultural fields were taken into account for further modelling procedures.

In both the 1993 and 2008 campaigns, soil-sampling scheme was designed to collect samples from various pedo-landscape (combinations of physiographies, lithologies and land uses) delineations as representative at a 1:250,000 scale. Samples of the 1993 campaign were taken following a specific guide for soil sampling and description, and consisted of minipits excavated up to a 50 cm depth to represent the top-soil, and sampled with the auger for the subsoil. The 2008 campaign consisted of soil profiles described according to the official methods of Italian Ministry of Agriculture (Paolanti et al., 2010). Soils from each campaign were sampled at various depths (maximum depth sampled up to 2.80 m). For the present study, the topsoil layer (up to 0.3-m depth) was taken into account. As stated above, soil layers were sampled according to the pedological description and thus upper and lower limit of each depth sampled varied among sampling points. Thus, to standardize the SOC concentration value, SOC was considered to decrease linearly with depth within each layers. In particular, soil layer in the depth 0-0.3 m were selected and those deeper than 50 cm were not used for the present experiment. The soil samples were passed through a 2 mm sieve, air dried, then analyzed for organic C content following Walkley-Black procedure.

5.2.3 Predictors

Climatic data were drawn from Worldclim (Hijmans et al., 2005). The original resolution of the Climatic data is about 1 km and were resampled to the desired 100 m mapping unit for the modelling process. Worldclim offers different datasets including bioclimatic data. Mean yearly rainfall and temperature of the 1950-2010 period were used.

Soil texture was obtained by the sedimentation method of the samples and reported according to the USDA classification. Soil texture for the whole area was provided by the Regional Bureau for Agriculture, Rural Development and Mediterranean Fishery, the Department of Agriculture, Service 7 UOS7.03 Geographical Information Systems, Cartography and Broadband Connection in Agriculture, Palermo.

The **CORINE land cover maps** of the years 1990 and 2006 at 100-m spatial resolution were used in order to identify the agricultural land uses for the model built for the year 1993 and 2008, respectively (http://land.copernicus.eu/pan-european/corine-land-cover).

The analysis was carried out according to the CORINE level 3, the Land cover type used in the modelling stage were: i) non-irrigated arable land (CORINE code 2.1.1, grid code 12, hereafter referred as ARA), ii) vineyards (CORINE code 2.2.1, grid code 15), fruit trees and berry plantations (CORINE code 2.2.2, grid code 16), and olive groves (CORINE code 2.2.3, grid code 17) (hereafter grouped in VFO), iii) annual crops associated with permanent crops (CORINE code 2.4.1, grid code 19), complex cultivation patterns (CORINE code 2.4.2, grid code 20), land principally occupied by agriculture, with significant areas of natural vegetation (CORINE code 2.4.3, grid code 21) (hereafter grouped in CCP). The land uses within the groups VFO and CCP were grouped since the SOC stock and relationship between SOC-predictors and SOC stock in these land uses is very similar due to similarities in plant density and soil management, as observed in Schillaci et al. (2017). CORINE codes are provided in Supplementary material Table 1.

Remote sensing-derived predictors consisted of the LANDSAT 5 spectral bands. The imagery was also used to derive the Normalized Difference Vegetation Index (NDVI), which was included as explanatory variables in the modelling phase. We used geometrical corrected images L1G. Multi-temporal mosaic required normalization to adjust for inconsistencies between images because of the proximity of the sun, earth and zenith angle. The procedure involved the conversion of the digital number to radiance at sensor. Calibration coefficient were provided in the imagery metadata (Guyot and Gu, 1994). The images used for the study were obtained by mosaicking the following five LANDSAT 5 scenes using the only cloud free scenes belonging to the path 188 row 33 (East), path 198 row 33 and 34 (middle) and path 190 row 33 and 34 (West) from the 1987 and 2003 for modelling data of 1993 and 2008, respectively. This time differences (1987 for the 1993 and 2003 for the 2008)

were needed since the regional extent of the study area requires at least 3 LANDSAT path to make a complete mosaicking of the region and these years were the closer to those of the sampling periods, in which the satellites scenes close each other in time had no or very few clouds, thus allowing a homogeneous dataset. LANDSAT imagery was freely acquired from the United States Geological Survey catalogue (USGS, http://earthexplorer.usgs.gov) and coincided with summer period (Rouse Jr et al., 1974), when most of the field crops have stubble or bare soil and very few or no crop growth occurs in other crops due to extremely high temperature and low water availability . All the RS predictors had an original spatial resolution of 30 meters and they have been subsequently resampled to the desired 100 m mapping unit. The choice of such predictor is due to their strong linkage to vegetation and other soil traits, and thus, to SOC.

Topographical indices

Shuttle Radar Topography Mission (SRTM-C) digital elevation model (DEM) released in September 2014 with a 1-arcsec (30 meter) spatial resolution (resampled to 100 meter to fit the land use classification) was used for the calculation of the morphometric spatial predictors by means of SAGA GIS (Conrad et al., 2015). DEM was downloaded from the earthexplorer.com website, then pre-processing such as mosaicking and fill sink was applied to the 10 SRTM DEM tiles covering the regional extent. Eleven terrain attributes were calculated: 1) slope 2) catchment area, 3) aspect, 4) plan curvature; 5) profile curvature, 6) length-slope factor, 7) channel network base level, 8) convergence index, 9) valley depth, 10) topographic wetness index, 11) landform classification. See http://www.saga-gis.org/saga_tool_doc/2.1.3/a2z.html for details on the computation of these covariates. Categorical predictor codes are provided in Supplementary material Table 1.

5.3.4 Boosted regression trees and map comparison

Boosted Regression Trees (BRT, Elith et al., 2008) was used to identify the relationships between SOC and its predictors and to regionalize the SOC prediction. This method and other decision trees-based models have already been used as DSM techniques to deal with SOC concentration and stock mapping (Bou Kheir et al., 2010; Grimm et al., 2008; Martin et al., 2011; Schillaci et al., 2017b). BRT is based on the integration of weak learners (or tree-based rules). In a data mining context, a weak learner is defined as a models that performs just slightly better than random guessing (Freund and Schapire, 1997). In this sense, the BRT algorithm combines multiple weak learners into a single strong learner (Lombardo et al., 2015). This allow the algorithm to progressively increases the accuracy of the prediction by reducing the chance of obtaining outliers since weak learners also produces weak outliers. This additive structure allows for capturing the variance of a dependent variable in a way where the deeper the tree is grown, the more fitting segments are obtained and added

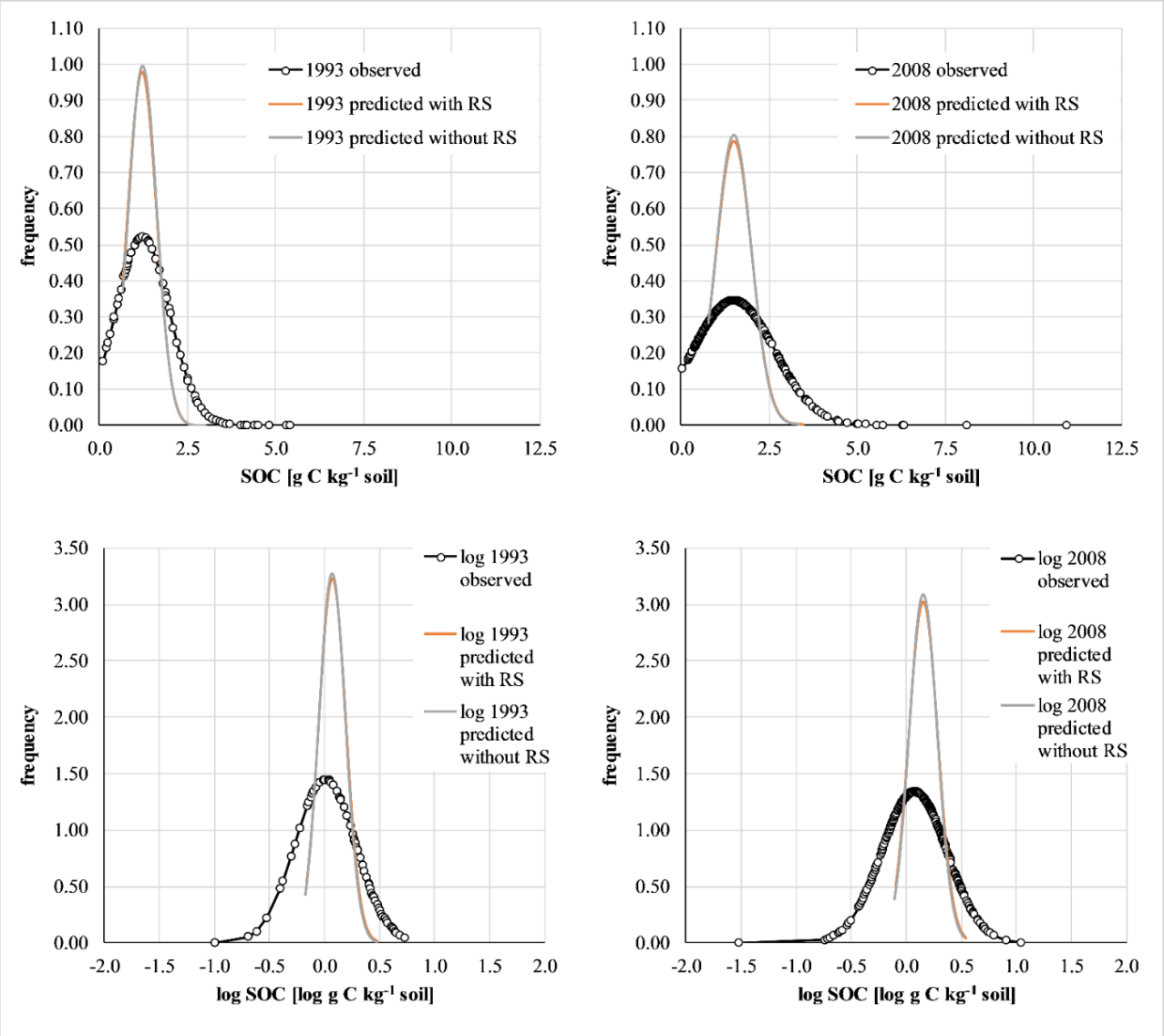
to the initial tree, to accommodate the SOC concentration at each mapping unit. The first step of this procedure consist of a Classification And Regression Trees (CART) analysis which recursively screens the observations in matched datasets made up by a dependent variable, either categorical (classification) or continuous (regression), and one or many explanatory variables. Explanatory variables can be either categorical or continuous. Differently from a classic CART approach, where a single tree can grow only to be finally pruned to get a readable model, the application of the BRT (second step) iteratively generates trees of a fixed dimension. Each tree is based upon the previous, and BRT gradually minimizes a loss function in order to improve the predictive performance. The adoption of the Huber-M loss function instead of a more common square loss function reduces the noise when iteratively measuring the difference between estimated and actual values for SOC concentration data. The procedure ceases when the creation of trees produces overfitting effects. The evaluation of the overfitting is performed by measuring the prediction residuals or deviance for each of the consecutive trees over a random independent sample that was kept separate from the calibration phase. Typically, the testing error quickly decreases the more trees are generated and subsequently slows down reaching an inflection point from where it starts to increase. This behavior is recognized as overfitting, determining the choice of the best model before the tree starts fitting the noise of the training data instead of revealing ecological relationships.

In the present research, 100 replicates were randomly generated and modelled from each of the original SOC concentration dataset. Relationships between variables are explained through response curves (Lombardo et al., 2015). We used R (R Development Core Team, 2008), with the 'dismo' package developed by Elith et al. (2008). The package allows for the customization of: i) learning rate (lr), which is set to determine the contribution of each tree to the final tree architecture; ii) tree complexity (tc), which controls the number of splits; iii) bag of fraction (bg), the proportion of data selected at each step of the modelling procedure. Following Hashimoto et al. (2016) we performed the 10-fold cross-validation procedure to determine the optimal number of trees (maximum numbers of trees 10,000) and a tc value of 20. Regarding each single run, model performances was assessed using the coefficient of determination of the scatter plot of the predicted against the observed values (pseudo-R^2) and root mean square error (RMSE). Standard deviation maps of the 100 runs were also constructed.

The maps of organic carbon generated for the 1993 and 2008 were compared and a difference (SOC_{08}–SOC_{93}) in which an increase of SOC was displayed as positive and a decrease as negative. An error map of the difference was built by adding the standard error of the 1993 and 2008 maps and highlighting those pixel which SOC difference (as absolute value) was higher than the sum of the standard errors. In such pixels, SOC difference was considered as reliable.

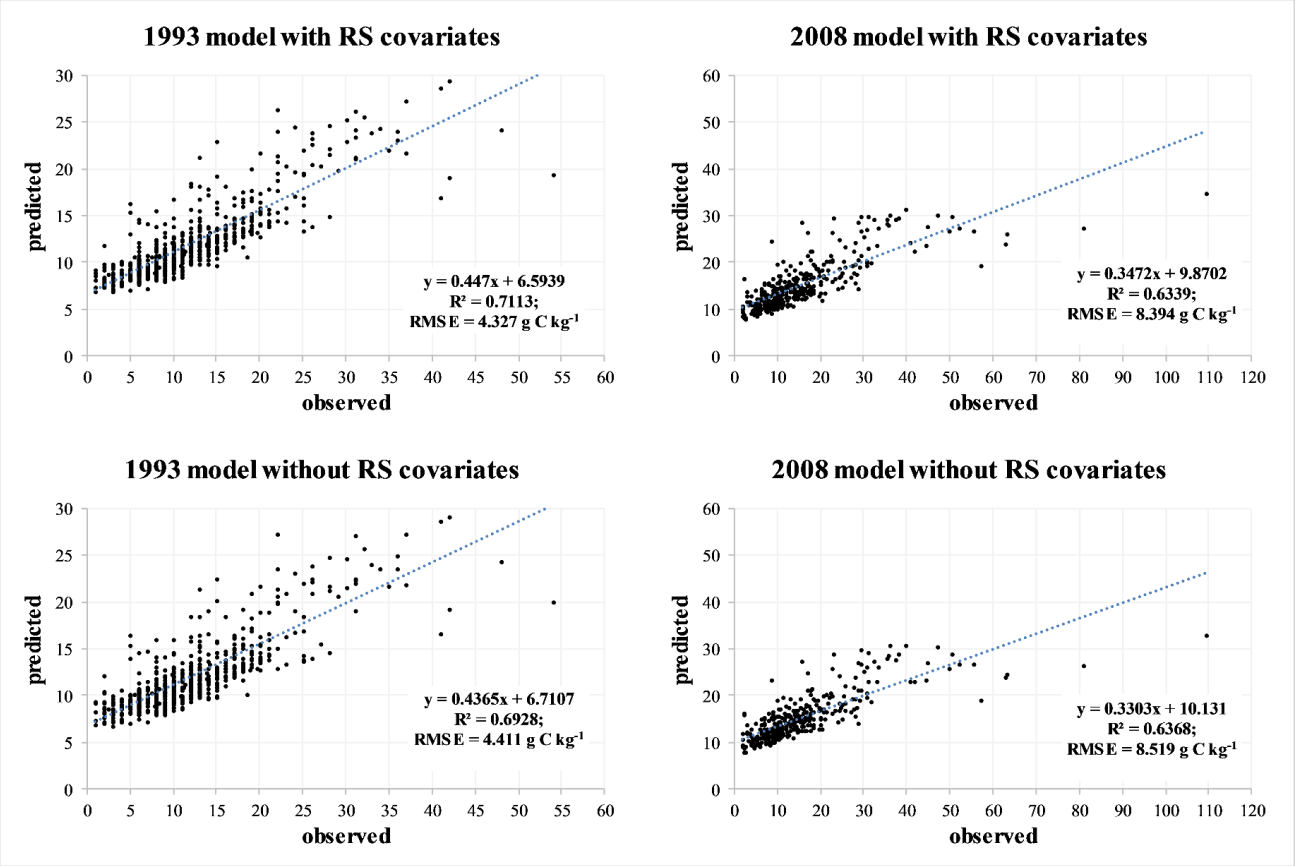
5.3 Results

Distributions of observed and predicted data with and without remote sensing (RS) predictors were log shaped (Table 1 and Supplementary material Fig. 2).



Supplementary material Fig. 2. *Frequency distributions of observed and predicted data with and without remote sensing (RS) predictors in 1993 and 2008 as both raw data (upper panels) and log-transformed data (lower panels). See table 1 for the descriptive statistics and main percentiles of each distribution.*

Distribution of predicted data showed similar skewness than observed data in 1993 and lower, but always positive, kurtosis in 1993 and kurtosis and skewness than observed data in 2008, which suggest that this method better estimates SOC in the central values of the distribution. All models had pseudo-R^2 higher than 0.693 for the 1993 model and 0.634 for 2008 model. The accuracy of the models with and without RS predictors was similar (Supplementary material Fig. 3).



Supplementary material Fig. 3. *Scatter plots of the boosted regression trees models of SOC performed with (upper panels) and without (lower panels) RS covariates by means of the data of 1993 (left) and 2008 (right).The function, pseudo-R2 and root mean square error (RMSE) of each pseudo regression is also shown.*

The removal of the RS predictors had a negligible effect on both the variation of the pseudo-R^2 and angular coefficient of the pseudo regression lines of both models, which was 0.43-0.45 in the 1993 and 0.33-0.34 in the 2008. Similarly, the intercepts were from 6.59 to 10.13 g organic C kg^{-1}, thus the predictions overestimated the observed value when SOC is low and down-estimated it when SOC is high.

	1993			2008		
	with RS	**without RS**	***fold variation***	**with RS**	**without RS**	***fold variation***
Non-remote sensed (RS) predictors						
Soil Texture	16.18	16.17	*1.00*	22.64	24.14	*1.07*
Land use	12.02	14.37	*1.20*	6.79	8.56	*1.26*
Valley depth	9.24	10.21	*1.10*	2.38	3.24	*1.36*
Rainfall	5.91	9.27	*1.57*	4.21	5.93	*1.41*
Channel network base level	4.97	6.96	*1.40*	9.05	10.35	*1.14*
LS factor	4.61	5.65	*1.23*	3.35	4.27	*1.28*
Landforms	4.19	5.04	*1.20*	4.44	5.34	*1.20*
Aspect	3.88	4.89	*1.26*	4.54	5.84	*1.29*
Elevation	3.38	4.65	*1.38*	3.12	3.90	*1.25*
Temperature	3.07	4.00	*1.30*	4.63	5.57	*1.20*
Cross sectional curvature	2.55	3.25	*1.27*	2.40	3.33	*1.39*
Slope	2.24	2.84	*1.27*	2.64	3.65	*1.38*
Vertical distance to channel network	2.00	2.62	*1.31*	2.78	3.74	*1.35*
Relative slope position	1.97	2.42	*1.23*	2.02	2.58	*1.28*
Catchment area	1.93	2.63	*1.36*	2.33	2.87	*1.23*
Convergence index	1.88	2.42	*1.29*	3.70	4.59	*1.24*
Topographic wetness index	1.85	2.60	*1.40*	1.60	2.09	*1.31*
RS predictors						
NDVI	7.11	-	n.a.*	2.45	-	n.a.
Landsat 1	1.98	-	n.a.	2.33	-	n.a.
Landsat 2	1.45	-	n.a.	1.45	-	n.a.
Landsat 3	1.80	-	n.a.	1.18	-	n.a.
Landsat 4	2.31	-	n.a.	2.73	-	n.a.
Landsat 5	1.91	-	n.a.	1.28	-	n.a.
Landsat 6	0.00	-	n.a.	3.93	-	n.a.
Landsat 7	1.57	-	n.a.	2.04	-	n.a.

* remote sensing; ** non applicable

Table 2. *The importance of each of the 25 predictors used in the boosted regression trees model to estimate the soil organic carbon performed on the 1993 and 2008 samples in Sicily, Italy. The role of the remote sensed (RS) predictors on the contribution to the total variance explained by the non-RS predictors and fold variation after removal of the RS predictors is shown*.

Removal of the RS predictors slightly changed the ranking of the predictors in terms of contribution to the total variance explained (Table 2). Among the RS predictors, only NDVI in the 1993 model showed a relatively high contribution to the variability explained (7.11%, the 4th strongest predictor), whereas its importance was negligible in the 2008 model (2.45%, the 15th predictor).

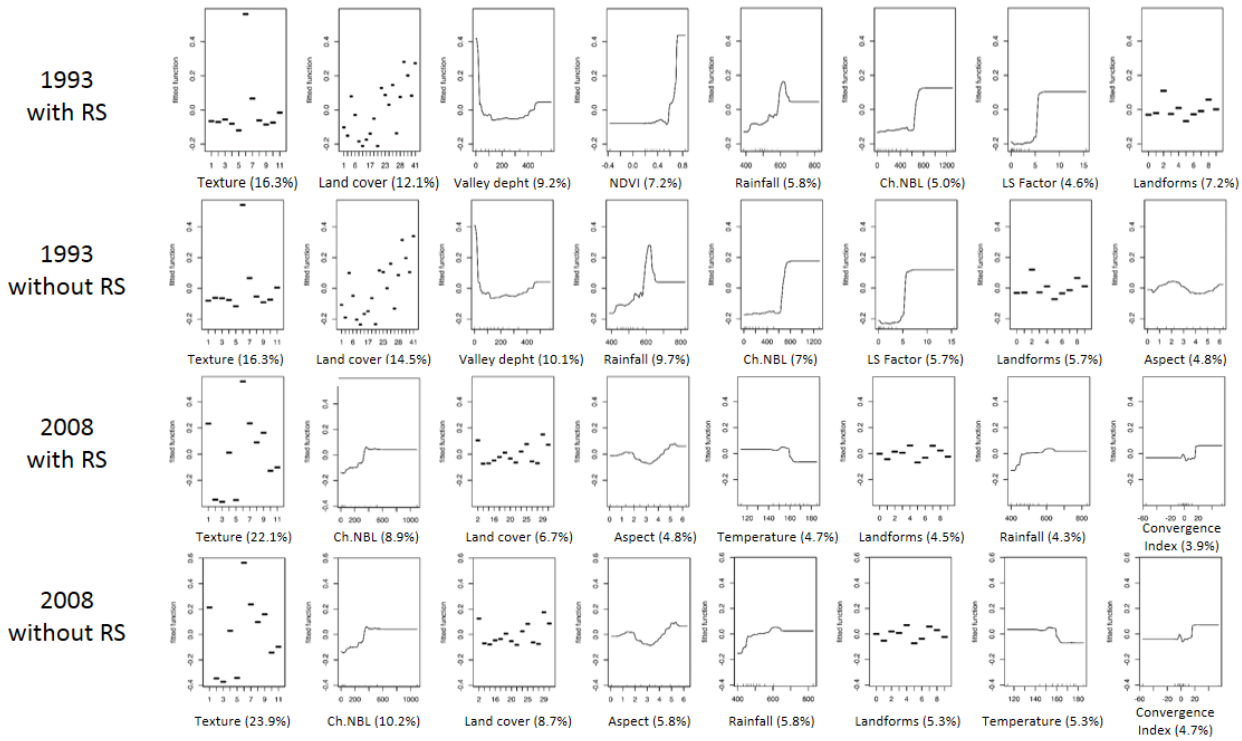
In general, the removal of the RS predictors resulted in an increase of the contribution to the total variance of the lowest contributing predictors (Table 1), with the exception of rainfall (5.91% in the 1993 model and 4.21% in the 2008 model). Rainfall contribution to the total variance explained was

1.57 and 1.41 fold after removal of the RS predictors. In total, the removal of the RS predictors from the modelling procedure increased the total contribution to the total variance explained of the six most important non-RS predictors by 9.71% in 1993 and 8.08% in 2008. The most important predictor of SOC content in both the 1993 and 2008 models was texture (19.18% and 22.64%, respectively, in the models with RS predictors). The six most important non-RS predictors across all 4 models were soil texture, land use, valley depth, rainfall, channel network base level (that is correlated with the height above the see level [a.s.l.] of the basin upon each pixel and thus to the chance of receiving SOC by erosion) and LS factor.

Table 1. Descriptive statistics of the observed soil organic carbon (SOC) concentration values and that of the distributions of the predicted SOC values modelled extracted on the same locations of the observed values. RS if for remote sensing covariates. Descriptive statistics were produced for both row and log-transformed data. Unit of measure for row data is % SOC.

	raw data						***log-transformed data***					
	1993			***2008***			***1993***			***2008***		
	observed	***predicted with RS***	***predicted without RS***	***observed***	***predicted with RS***	***predicted without RS***	***observed***	***predicted with RS***	***predicted without RS***	***observed***	***predicted with RS***	***predicted without RS***
Mean	1.2219	1.2246	1.2246	1.4881	1.4959	1.4965	0.0080	0.0687	0.0693	0.0743	0.1536	0.1546
Standard error	0.0273	0.0146	0.0143	0.0567	0.0249	0.0244	0.0098	0.0044	0.0044	0.0146	0.0065	0.0064
Minimum	0.1000	0.6821	0.6665	0.0300	0.8027	0.7774	-1.0000	-0.1661	-0.1762	-1.5229	-0.0955	-0.1093
Percentile 1%	0.2000	0.7322	0.7231	0.2000	0.8523	0.8889	-0.6990	-0.1354	-0.1408	-0.6990	-0.0694	-0.0512
Percentile 2.5%	0.2000	0.7779	0.7811	0.2533	0.9137	0.9222	-0.6990	-0.1091	-0.1073	-0.5965	-0.0392	-0.0352
Percentile 25%	0.8000	0.9599	0.9611	0.8325	1.1294	1.1416	-0.0969	-0.0178	-0.0172	-0.0796	0.0529	0.0575
Median	1.0000	1.1125	1.1148	1.1450	1.3573	1.3480	0.0000	0.0463	0.0472	0.0588	0.1327	0.1297
Percentile 75%	1.5000	1.3453	1.3392	1.7575	1.6973	1.7033	0.1761	0.1288	0.1268	0.2449	0.2298	0.2313
Percentile 97.5%	3.2475	2.4201	2.3855	4.4638	2.9182	2.8322	0.5115	0.3838	0.3776	0.6497	0.4651	0.4521
Percentile 99%	4.2000	2.7196	2.7162	5.6966	2.9813	3.0149	0.6232	0.4345	0.4340	0.7556	0.4744	0.4793
Maximum	5.4000	2.9830	3.0140	10.9500	3.4762	3.3565	0.7324	0.4746	0.4791	1.0394	0.5411	0.5259
Mode	1.0000	1.0554	1.0151	0.9900	0.8205	0.7774	0.0000	0.0234	0.0065	-0.0044	-0.0859	-0.1093
Standard deviation	0.7648	0.4074	0.4002	1.1530	0.5074	0.4968	0.2751	0.1237	0.1218	0.2972	0.1318	0.1294
Kurtosis	5.1596	3.4557	3.4879	14.9722	1.5478	1.5422	1.3897	0.7781	0.8042	2.1546	-0.0729	-0.0682
Skewness	1.8570	1.7964	1.7953	2.9695	1.3679	1.3563	-0.6215	0.9741	0.9742	-0.3757	0.6848	0.6774

In the models both with and without RS predictors, a discrepancy in the association between soil texture levels and relative importance for SOC prediction was found between the 1993 and 2008 models (Supplementary material Fig. 4).



Supplementary material Fig. 4. *Partial dependence (response curve plots) of the model on the 8 most important predictors included in the boosted regression trees. Models were built for 1993 and 2008 data and with or without remote sensed (RS) predictors. The importance of each predictor is shown in parentheses.*

In the 1993 model, only Silty-Clay-Loam (texture 6) and Sandy-Loam (texture 7) showed a positive association to the SOC, whereas in the 2008 model, such a positive association was also found for Clay (texture 1), Sandy Clay Loam (texture 8), and Sandy soils (texture 9). In both models, CCP contributed more than VFO to SOC estimation and VFO more than ARA. Channel network base level negatively correlated with SOC estimation in the first half of its range in both the 1993 and 2008 models (up to 660 and 330 m a.s.l., respectively), after which its contribution to the function of SOC estimation was always positive and constant. Similar trends were observed for the SOC to rainfall relationship. The role played by valley depth was strong in the 1993 model, only. Valley depth, which is inversely correlated with the deposition process, positively associated with SOC only in the lowest SOC concentration samples.

As expected, the highest SOC concentrations were mostly found in sites with relatively low mean temperature and high rainfall, which, in this area, are conducive for C accumulation in soil (see Cannarozzo et al., 2006; and Viola et al., 2014 for maps of rainfall and temperatures). In our study area, these sites are mainly located at the boundaries of the mountain chains (Fig. 2 and 3): the northern mountain chains (Madonie, Nebrodi and Peloritani), the Volcano Etna in the eastern part of the island, the Sicani Mountains in the western part of the island, the Hyblaean area in the south-eastern corner. In general, the higher the SOC concentration, the higher the standard error of the model. The models with RS showed a lower standard error than the models without RS, especially in 1993.

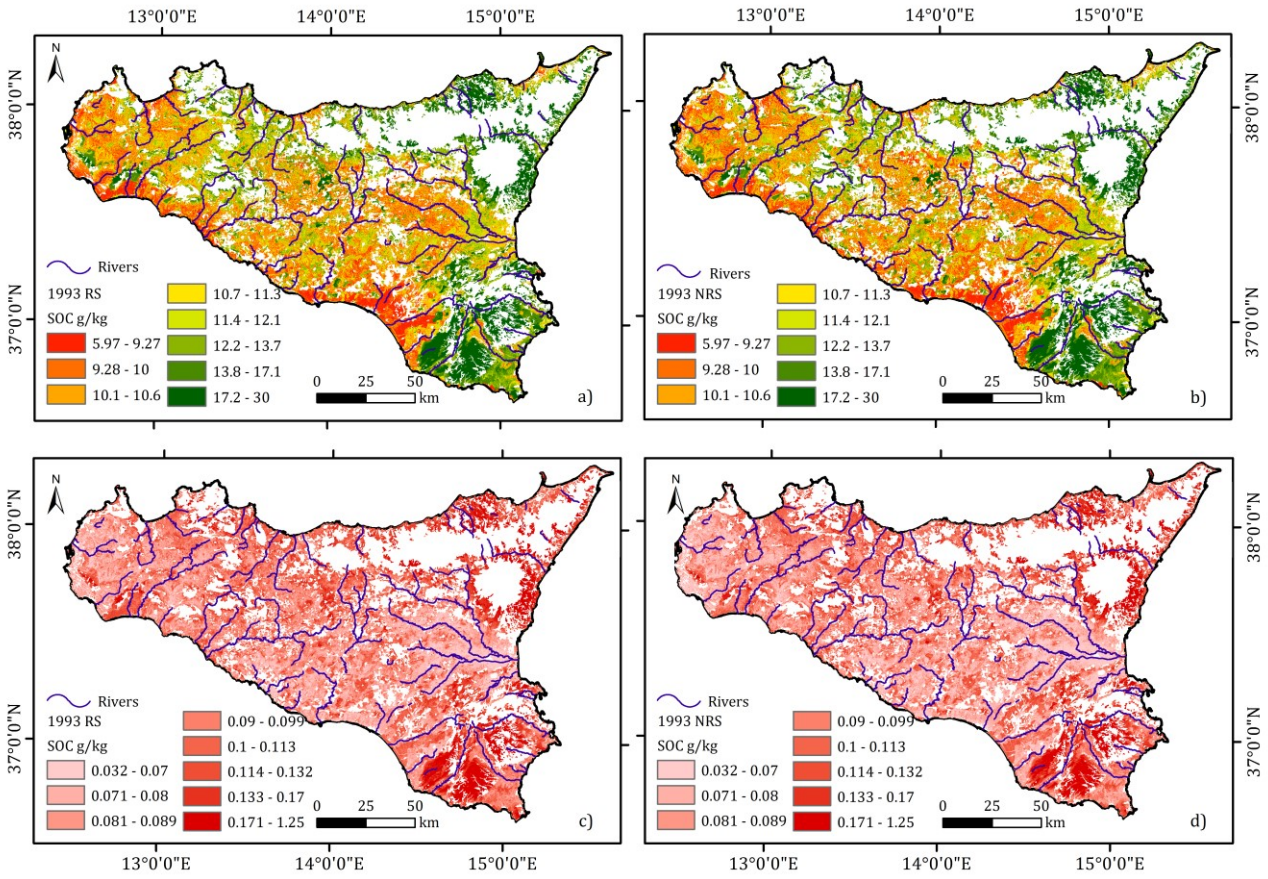
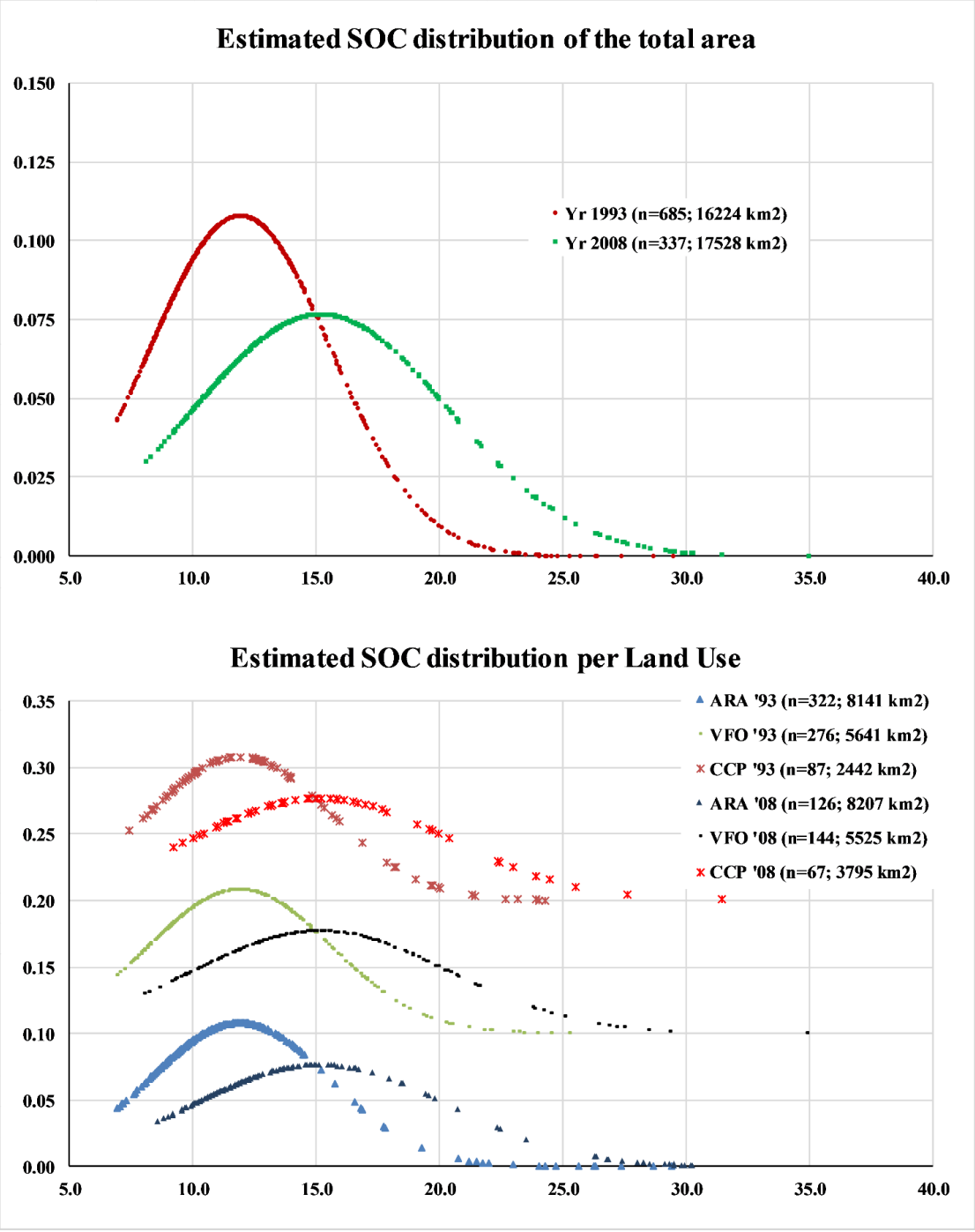


Fig. 2. *One-hundred meters resolution maps of the SOC (expressed in g C kg–1, a, b) and uncertainty maps (c, d) of the boosted regression trees models built with data from 1993 with (a, c) or without (b, d) remote sensed covariates. Please note that range vary among classes.*

Classification of the predicted samples in the range ±50% than the observed was high for both the 1993 and 2008 models (81% and 72% of the estimated data extracted on the same location of the entry data; Fig. 4) and well distributed in the area. Samples classified in the ranges < or >50% than the observed were also well distributed.

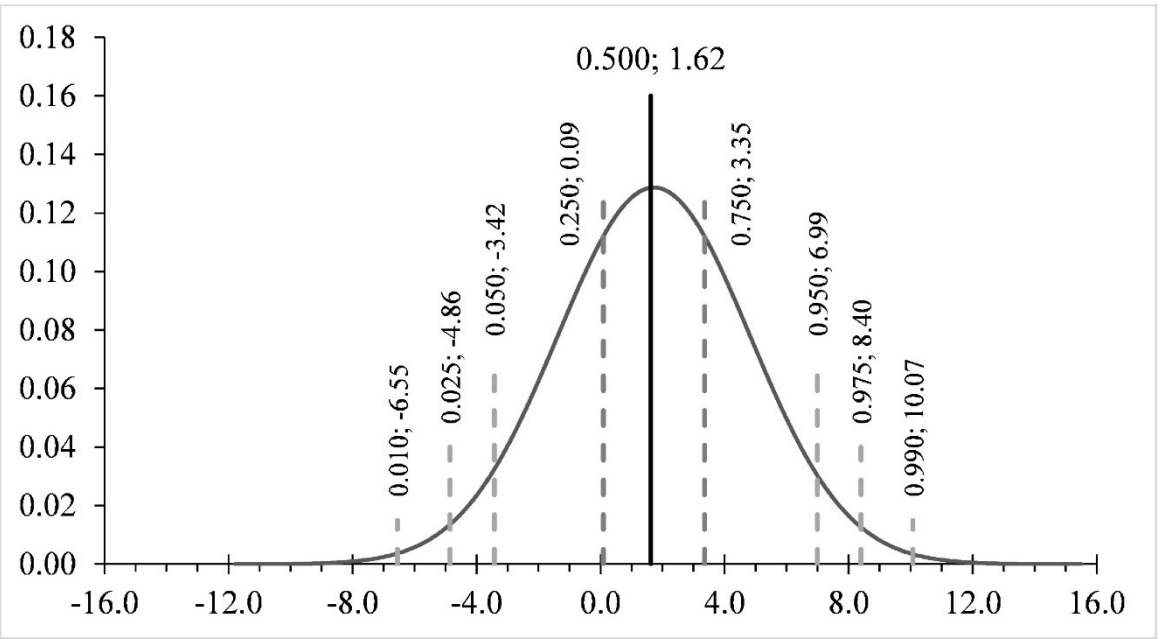
The removal of the RS predictors did not exert an effect on the SOC prediction (Fig. 5), which was on average 11.9 g organic C kg^{-1} in ARA, 12.6 g organic C kg^{-1} in VFO, and 14.4 g organic C kg^{-1} in CCP. Irrespective of the presence of the RS covariates in the model, such amount increased by 1.9%, 1.9% and 0.9% in ARA, VFO, and CCP, respectively, from 1993 to 2008 and such increase occurred in all land use groups considered in a similar extent (Supplementary material Fig. 5).



Supplementary material Fig. 5. *Distribution function of the SOC (expressed in g C kg^{-1}) from data estimated by boosted regression trees extracted on the same location of the observed values. In the lower panel, the same distribution of the upper panel is shown with data divided per land use for*

readability purposes. ARA is for non-irrigated arable land; VFO is for vineyards, fruit trees and berry plantations, and olive groves; CCP is for annual crops associated with permanent crops, complex cultivation patterns, land principally occupied by agriculture, with significant areas of natural vegetation. Data of frequency of VFO were added arbitrarily +0.1 and data of CCP were arbitrarily added +0.2 to help in the comparison among land use groups and year of sampling. The number of data in each sub-dataset and area (expressed in km^2) covered by each group is shown in parentheses.

The variation of the SOC in the area under study strongly depended on the subarea within the region and did not match the SOC map at the baseline (1993) (Fig. 6) In contrast, the reliability of this difference [measured as $|SOC_{08-93}|-(STDEV_{08}+STDEV_{93})$] did not depend on the area and was positive in almost all pixels. An increase of SOC concentration (up to +17.0 g SOC kg^{-1} in the right end of the difference distribution, +10.1 g SOC kg^{-1} in the 99th percentile, i.e +0.67 g SOC kg^{-1} yr^{-1}, Supplementary material Fig. 6) was frequently found in the Hyblaean area, especially in the mountains and hilly environments, in the western hilly to plains areas, and, unexpectedly, on the central area located on the south of the northern mountain ridge.



Supplementary material Fig.6. *Distribution function of the SOC differences (expressed in g C kg^{-1}) resampled from a 1-km resolution map. Vertical lines indicate, from the left to the right, the quantile 0.01, 0.025, 0.05, 0.25, 0.50 (median, as a continuous line), 0.75, 0.95, 0.975, 0.99 of the distribution.*

A loss of SOC (up to –13.0 g C kg^{-1} in the left end of the difference distribution, –6.6 g SOC kg^{-1} in the 1st percentile, i.e –0.44 g SOC kg^{-1} yr^{-1}) was observed in the areas surrounding the other

mountains ridge, the areas between the eastern slope of Etna Volcano and the sea and the Catania plain to the south of Etna, the Hyblaean plains on the south of the Hyblaean Mountains, and in part of the far-western plains, near the western corner of the island.

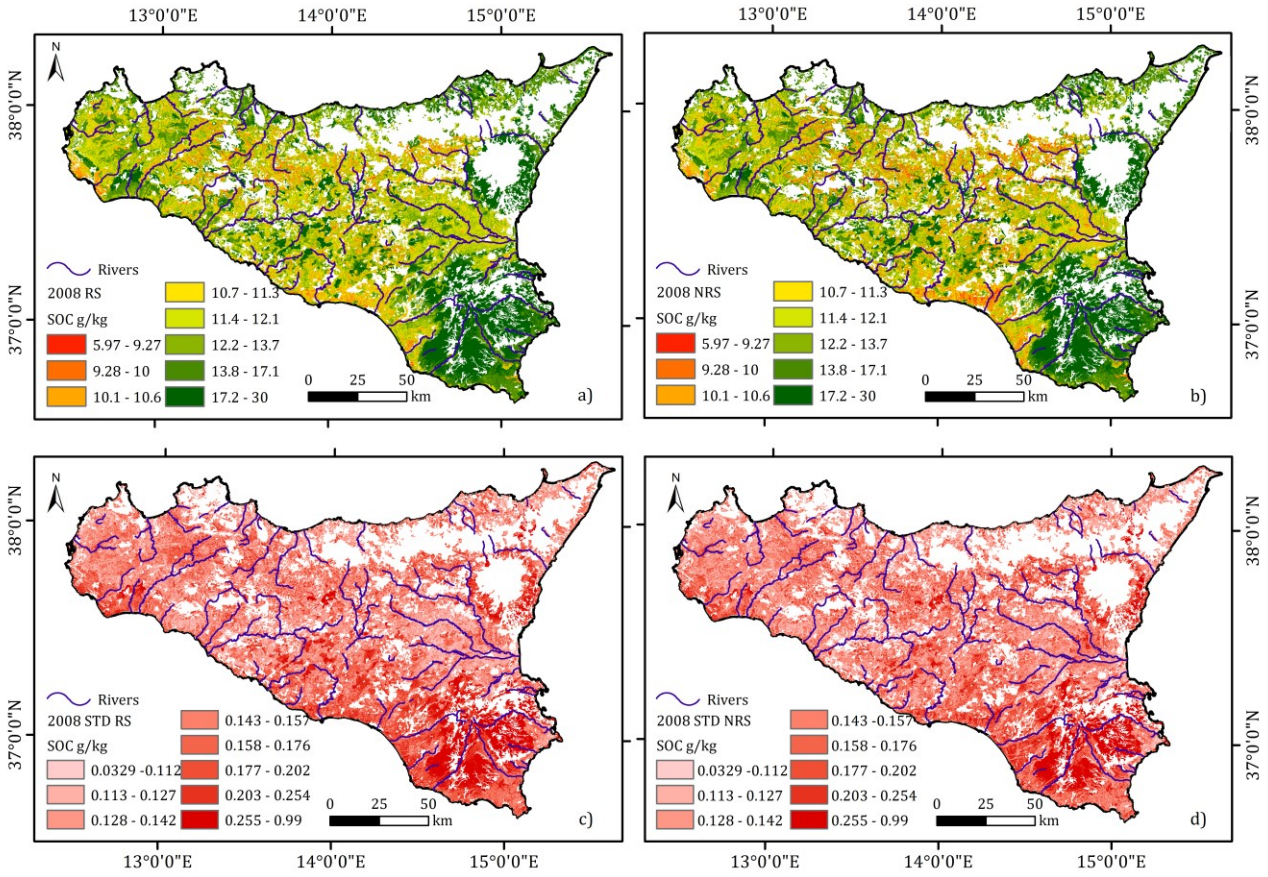


Fig. 3. *One-hundred meters resolution maps of the SOC (expressed in g C kg^{-1}, a, b) and uncertainty maps (c, d) of the boosted regression trees models built with data from 2008 with (a, c) or without (b, d) remote sensed covariates. Please note that range vary among classes.*

5.4 Discussion

The understanding of the space-time variation of SOC is a prerequisite to hypothesize future scenarios and the outcome of any policy on crop yield, yield potential and ecosystem service (Dono et al., 2016; Elith et al., 2008; Luo et al., 2015; Novara et al., 2017). Thus SOC should be primarily managed to increase (agro)-ecosystem resilience to anthropic pressure and climate change. However, the mutual relationship of SOC and climate change depends on several variables (e.g. soil texture or tillage) and have wide variation (Kirschbaum, 1995; Stockmann et al., 2013). In this framework, the integration of short and long term comparisons (Conant et al., 2001; Kämpf et al., 2016; Kurganova et al., 2014; W M Post, 2000) can strongly boost the accuracy of SOC prediction (Luo et al., 2015). However, single-point comparisons, even when analyzed for a wide timespan, have the drawback of being uncorrected for position in the stochastic population of the data and are thus not representative of wide areas.

In the present study, the integration of DSM and BRT modelling allowed us produce maps of probable agricultural topsoil SOC distribution (along with reliability and error maps) for two sampling campaigns performed 15 years apart (1993 and 2008). This allowed us to estimate how SOC varied through space and time at each land use group (arables [ARA], tree-like crops [VFO], and cropped areas with semi-natural vegetation [CCP]) and the importance of some ecological characteristics on space-time SOC variation.

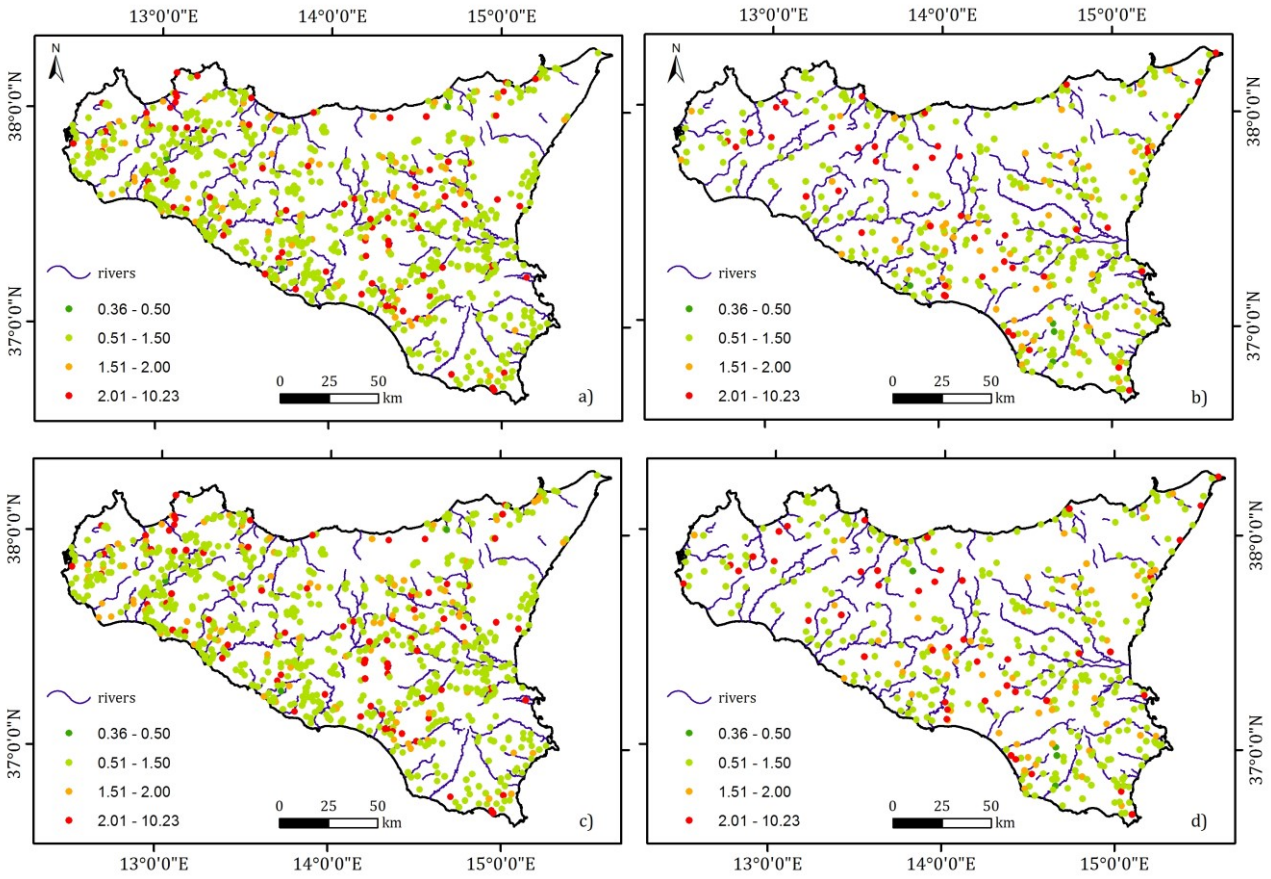


Fig. 4. *Prediction confidence map of the boosted regression trees (BRT) models of 1993 (a, c) and 2008 (b, d) built with (a, b) or without (c, d) remote sensed predictors. Each point represents the ratio between BRT-predicted and observed values. The closer the ratio is to 1, the better its representation of the observed value is.*

The study period was selected according to the highest availability of data within each campaign and its timespan (15 years) allowed us to depict a short-term variation of SOC within a well-characterized period. Its beginning (1993) luckily fell soon before a number of European and worldwide policy measures which profoundly impacted agriculture, including the Regulation EEC 1272/88 on set-aside (compulsory from the 1992); the United Nations Framework Convention on Climate Change of 1993 (into force from 1994); and the World Trade Organization Marrakesh Agreement of 1994. Similarly, its end (2008 campaign) fell soon after the abolishment of the compulsory set-aside in the EU

(Common Agricultural Policy [CAP] health check 2008) and the decoupled CAP EU payments to agriculture in 2005 (Regulation EEC 1782/2003). This collocates our research study in a period of low agricultural dynamic in term of land use change and management techniques, the latter of which were dominated by deep plowing.

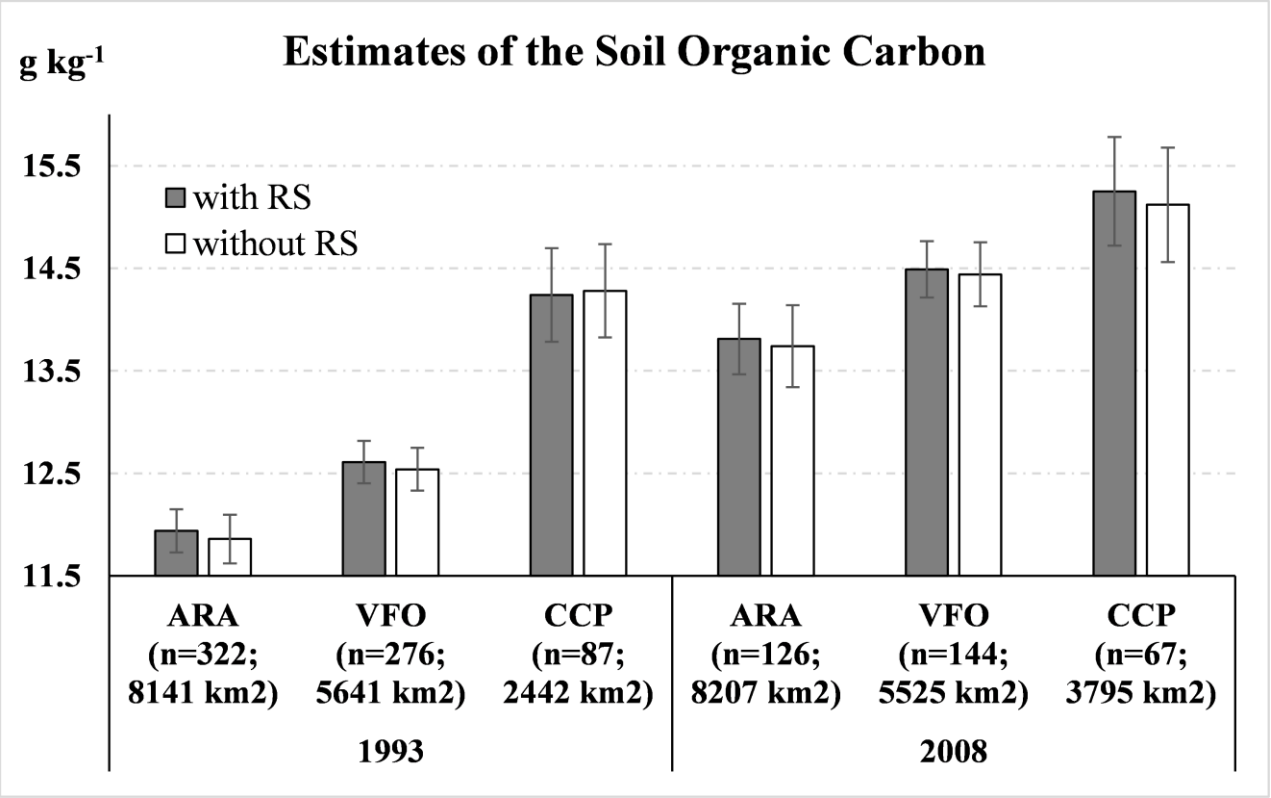


Fig. 5. *Estimates of the soil organic carbon in each of the land use groups used in the present study as affected by the presence of the remote sensed (RS) covariates in the model. ARA is for non-irrigated arable land; VFO is for vineyards, fruit trees and berry plantations, and olive groves; CCP is for annual crops associated with permanent crops, complex cultivation patterns, land principally occupied by agriculture, with significant areas of natural vegetation. Data are means ± standard error. Number of sampling points falling into an area of each land use group is shown.*

Indeed, we found that the area covered by ARA and that by VFO were almost constant during the study period (1993 to 2008), whereas the area covered by CCP increased by 55%, which was likely due to the temporarily conversion of grassland to pastures. As expected, we found that SOC of ARA was predicted as lower than VFO and that of VFO lower than CCP. The increase in the SOC stock during the study period was however partly unexpected. From the one hand, we expected to find an increase in the ARA and VFO due to many conditions. These include the application of Good Agricultural and Environmental Conditions (Borrelli et al., 2016), which effects on ARA were directly elucidated in similar environments (Ventrella et al., 2011); the high clay content in the soils cropped with these species, as directly addressed by Zinn et al. (2005); massive recourse to the set-aside (partly compulsory); the minor role of climate change in agricultural areas (Cannarozzo et al.,

2006; Fantappiè et al., 2011b); and ease of SOC increase in low-SOC soils (Kämpf et al., 2016), such as those in the present study (<12.6 g kg^{-1} ± 0.21 g kg^{-1}). From the other hand, such an increase was expected to occur in the northern, rainy, part of Sicily thanks to the presence of conditions conducive to a SOC accumulation, rather than in the southern, more arid parts, whereas we found an opposite pattern. Nonetheless, these results agree with those of other lower resolution studies in the same area (Chiti et al., 2012; Fantappiè et al., 2011b; Freibauer et al., 2004; Hashimoto et al., 2016; Lugato et al., 2014a) or studies conducted in similar environments (Farina et al., 2016; Rodríguez Martín et al., 2016), where soil management exerted an important role in the percentage or reduction of SOC in relatively humid areas.

Climate change effect on Sicily are under debate: no change in the rainfall in most of ARA and VFO-dominated areas is expected (Cannarozzo et al., 2006), and a temperature increase is likely to occur (Viola et al., 2014). However, the interaction between water availability and temperature with the effect of soil traits and land use on potential and actual mineralization and C inputs are yet to be clarified (Badagliacca et al., 2017; Bogunović et al., 2017b; Davidson and Janssens, 2006; Purton et al., 2015). For example, in a high organic C area (Galapagos), Rial et al. (2017) suggested that the increase in the amount of rainfall and in general water availability (through occult precipitations, too) will likely consist in an increase of the SOC stock.

During this 15-years study (1993-2008), mean increase in SOC in the agricultural area of the region (median = + 1.62 g C kg^{-1} soil; lower confidence interval 95%: – 4.86 g C kg^{-1}; upper confidence interval 95%: + 8.40 g C kg^{-1}) appeared similar to the time trends in temperature and rainfall observed in the region (Cannarozzo et al., 2006; Viola et al., 2014) and the degree of lithological and soil diversity (Costantini and L'Abate, 2016; Fantappiè et al., 2015). This occurred despite the most important predictors of SOC at any pixel were soil texture, land use and topographic covariates, as also found elsewhere (Bogunović et al., 2017b), whereas rainfall and temperature only contributed by 8.98% and 8.94% of the total variability explained in the 1993 and 2008 model, respectively.

Grinand et al. (2017), by means of an algorithm similar to the one we used, found that SOC change modelled in a 20-years timespan was likely negative in humid and not different than zero in arid areas and that such variation strongly depended on both the climatic predictors and degree of deforestation. However, in contrast to Grinand et al. (2017), we found an increase of the CCP, which effect on SOC is more similar to that of forests compared to ARA and VFO.

A matching between SOC and climatic gradient was observed by Vaysse and Lagacherie (2015) in southern France, a colder and more rainy environment than Sicily. In addition, in the 'Vaysse and Lagacherie (2015)' modelling of soil traits, a similarity among maps of SOC, soil pH and soil clay content can be observed. It is likely that in our environment, the variability of some important traits

related to soil erosion and deposition (such as valley depth and channel network base level) and thus C movements by erosion and deposition across pixel was better related to trends in rainfall and temperature, than their long-term mean.

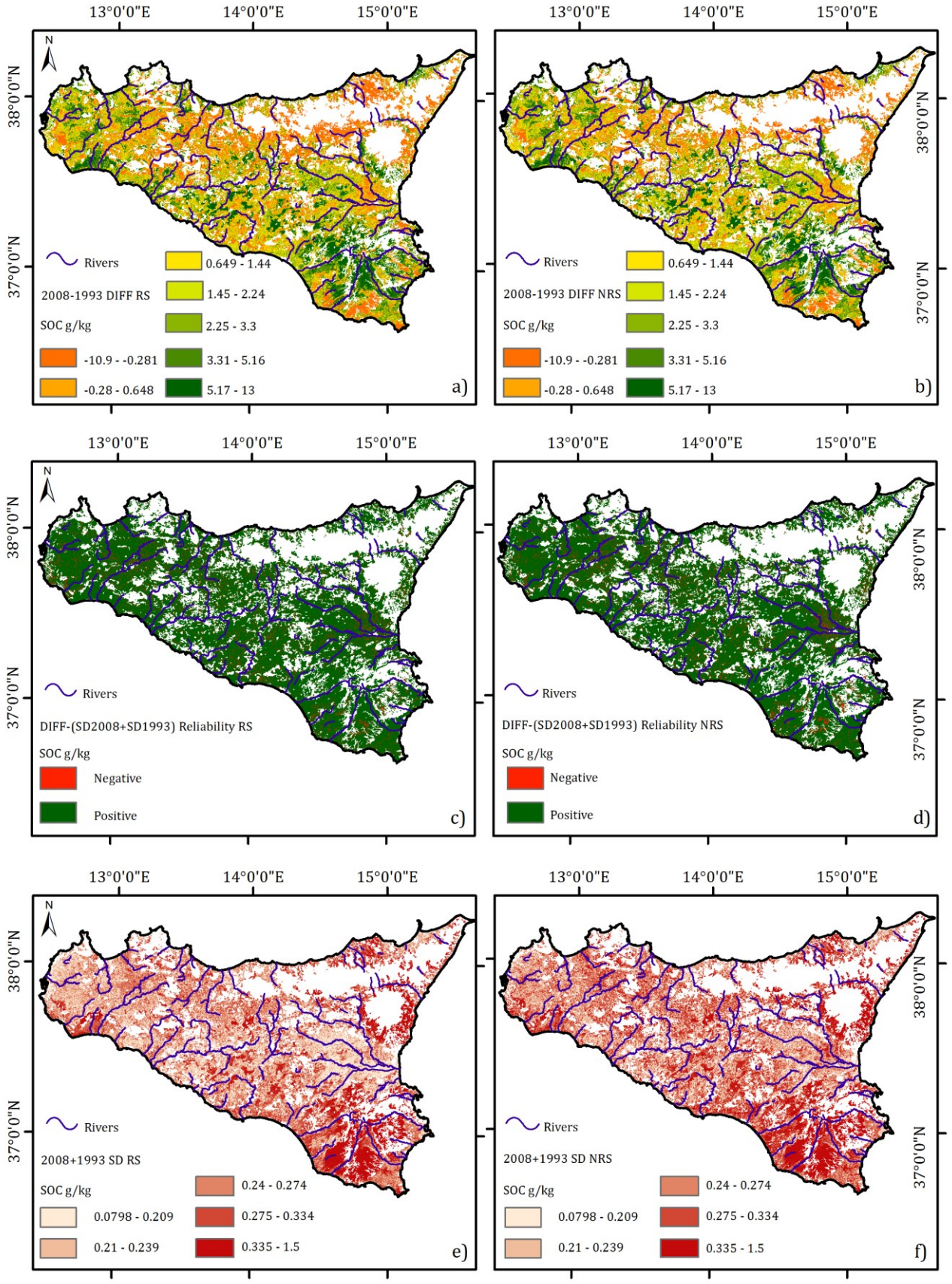


Fig. 6. *One-hundred meters resolution map of the difference in the SOC (expressed in g C kg^{-1}) during the study period (a, b). Reddish pixels indicates a loss and greenish pixels a gain in the SOC in 2008 compared to 1993. Please note that range vary among classes. Reliability (c, d) of the maps in A and B panels, respectively, computed as the difference between the SOC difference and the sum of the standard errors (in lower panels of Fig. 2 and 3). Green points indicate those pixel in which the difference of SOC is reliable. Maps of the sum of the standard deviations of the 'map of SOC' (e, f). Each computation and mapping was made for models built with (a, c, and e) and without (b, d, and f) remote sensing (RS) predictors.*

Nevertheless, the present results only partly fitted the erosion risk map published soon before the beginning (Ferro et al., 1991) or the end (Fantappiè et al., 2015) of the present experiment. This latter discrepancy can depend on both the difference in the spatial resolution between the present map and those of Ferro et al. (1991) and Fantappiè et al. (2015) and the lack in these of the information about the deposition of the eroded soil and C (Adhikari et al., 2014). Indeed, we found that catchment area, landforms, valley depth and channel network base level, which are related to soil deposition, contributed by 20.3% and 18.2% of the total SOC variability explained in 1993 and 2008, respectively. Topographic indices can strongly affect SOC concentration through erosion and deposition, whereas their role in SOC stock can be minimal (Grimm et al., 2008; Schillaci et al., 2017b). In the present work, we found that RS indices minimally increased the pseudo-R^2 of the fitting functions and mostly affected both the variance explained by each covariate and the variability among model replicates. In particular, the RS covariates captured on their whole 18.1% and 17.4% of the total variance explained in the 1993 and 2008, respectively. Bou Kheir et al. (2010) found that removal of RS indices can increase the total variance explained by the less important predictors and, in contrast to the present study, also the overall accuracy of the model.

Other studies indicated that the importance of RS indices in SOC mapping can depend on a range of factors, including the variable mapped, the resolution of the measured and ancillary variables, the extent of the study and the importance of the processes of SOC accumulation in relation to the study area (Castaldi et al., 2016b; Grinand et al., 2017; Poggio et al., 2013; Priori et al., 2016). It is thus likely that the high number of non-RS covariates in this work and their ability to explain a high degree of variability reduced the ability of the RS data to explain an additional amount of variability. In addition, the need of using more than one Landsat image (each of which took 13-32 days apart from each other) could have reduced the importance of RS indices for the whole area and impaired their contribution to the prediction. Similarly, some experiments with fewer input points and or coarser covariates than the present found a high percentage of variance explained by the RS indices in either SOC or other environmental traits (Stephen I.C. Akpa et al., 2016; Castaldi et al., 2016b; Wang et al., 2016).

5.5 Conclusions

In the present work, two legacy sub-datasets of SOC concentration were integrated in a DSM procedure to estimate the SOC variation along a 15-years period (1993-2008). This results was possible since the application of the covariates produced a pseudo-R^2 of SOC representation of 0.63-0.69, which allowed a time comparison of SOC at the pixel level. Texture and land use classes showed the highest predictor importance, around one third of the variance explained. Yigini and Panagos (2016) indicated these traits as capable of having a short-term impact on the SOC higher than climate-driven processes.

The integration of RS indices used in this study did not increase the pseudo-R^2, but captured about one fifth of the total variance explained by the covariates and strongly reduced the modelling variability. This suggests that their integration in the models can overcome problems related to erroneous attribution of some samples to the other covariate levels.

Finally, the present results can imply both agronomic and policy consequences at the district level and call for an intervention on soil fertility to maintain agriculture productivity (Dono et al., 2016). These results can help in calibrating models of SOC dynamic under various management or climate change scenarios, especially at regional extent, by removing the noise in the modelling phase by a correction with RS or other soil traits and geographical covariates, as already shown with other disturbing covariates in SOC modelling (Bogunović et al., 2017a, 2017b; Zinn et al., 2005a), which provide measures of covariates with a unique resolution in broad areas.

Acknowledgment

The authors would like to thank Maria Gabriella Matranga, Vito Ferraro and Fabio Guaitoli from the Regional Bureau for Agriculture, rural Development and Mediterranean Fishery, the Department of Agriculture, Service 7 UOS7.03 Geographical Information Systems, Cartography and Broadband Connection in Agriculture, Palermo. The authors also thank three anonymous reviewers for their constructive comments, which helped to improve the present manuscript.

Chapter 6- Comparison between geostatistical and machine learning models to predict topsoil organic carbon with a focus to local uncertainty estimation

From: Veronesi F., Schillaci C.- Comparison between geostatistical and machine learning models to predict topsoil organic carbon with a focus to local uncertainty estimation. Manuscript submitted to Ecological Indicators (Under review)

Keywords: machine learning; kriging; digital soil mapping; random forest; boosted regression trees; local uncertainty; regression kriging; agriculture; confidence intervals

Abstract

In recent years, the environmental modeling community has moved away from kriging as the main mapping algorithm and embraced machine learning (ML) as the go-to method for spatial prediction. The drawback of this shift has been a gradual decline in the number of papers in which uncertainty is presented and mapped alongside estimates of the target variables because in some ML algorithms, computing the local uncertainty can be challenging. This drawback has been recently identified in the literature as one of the key areas in DSM where progress is most needed. The main objective of this work is to compare geostatistical techniques, ML methods and hybrid methods, e.g., regression kriging, in terms of not only their overall accuracy but also their precision in providing useful confidence intervals at unsampled locations. We aim to provide clear application guidelines for future mapping exercises.

For this experiment, we used a legacy soil dataset (n=414) of topsoil observations from the semi-arid Mediterranean region of Sicily. This dataset was collected in a 2008 survey with a pedo-landscape sampling design; hence, it is ideal for comparing geostatistics and ML. In the comparison, we included algorithms that have been widely adopted in the literature: ordinary and universal kriging, linear regression, random forest (RF), quantile regression forest, boosted regression trees (BRT) and hybrid forms of kriging (e.g., regression kriging with RF and BRT used as regressors).

To evaluate the accuracy of each algorithm, a validation test that was based on the random exclusion of 25% of the samples was repeated 100 times. In addition, we performed a test of the transferability, in which the locations with the largest nearest-neighbor distances were excluded from training and re-predicted. The validation results demonstrate that ordinary and universal kriging are the best performers, followed closely by random forest (RF) and quantile regression forest (QRF). In terms of local uncertainty, RF and QRF provide confidence intervals that most often include the observed values of SOC. However, they both provide very wide confidence intervals, which may be problematic in some studies. Other algorithms, such as boosted regression trees and boosted regression kriging, performed slightly worse (on this dataset), but produced narrower ranges of

uncertainty. Hence, they may be more attractive since their estimates are very robust against changes and noise in the predictors.

Keywords: boosted regression trees, digital soil mapping, machine learning, kriging, local uncertainty, random forest, regression kriging.

6.1 Introduction

Spatial prediction is the process of estimating a target variable at unsampled locations and can be realized by applying a wide range of models, including the very popular kriging model, which is still heavily used in the digital soil mapping (DSM) and ecological modeling communities.

In recent years, with the spread of remote sensing and the open sharing by many public organizations of high-resolution weather and climate data, ecological modelers have become increasingly interested in using ancillary data, which has fundamentally changed the way in which spatial prediction is realized, with practitioners relying less on interpolation and more on machine learning (ML), which has become very popular.

From the literature, it is unclear which class of algorithms performs the best overall for soil organic carbon (SOC) mapping. Vermeulen and Van Niekerk (2017) tested several algorithms for predicting soil salinity in a region of South Africa and demonstrated that kriging with an external drift was outperformed all the ML algorithms that they tested**.** Rhee and Im (2017) compared ML and kriging for drought forecasting in South Korea and concluded that ML yielded more promising results. Recently, ecosystem service analyses of the soil carbon and its spatial distribution were conducted for various environments: i) forest environment (Ottoy et al., 2017). In Beguin et al., (2017), the authors compared eight algorithms, including random forest, boosted regression trees and kriging, and reported that kriging outperformed the others; ii) coastal ecosystems (Carranza et al., 2018; S. Wang et al., 2018); iii) agro-ecosystems (Chen et al., 2018; Schillaci et al., 2017; Song et al., 2017; B. Wang et al., 2018); iv) Tibetan plateau (Dai et al., 2014; Yang et al., 2016); and v) Afromontane ecosystems, for which (Were et al., 2015) tested ML algorithms and ordinary and regression kriging to estimate soil organic matter and concluded that regression kriging realized the highest level of accuracy. Nussbaum et al (2018) reported that RF outperformed the other algorithms used in their experiment. Performances are evaluated to identify the best algorithm and the best fitting statistics. Prediction performances are highly dependent on not only the algorithm but also the environment, data and predictors that are used in a study. Hence, it is unsurprising that no “one-size-fits-all algorithm” exists.

The essential concern for the scientific community is that in recent years, a shift from geostatistics to ML seems to have occurred, despite the lack of evidence that the level of accuracy differs significantly. To examine this, we performed a bibliographic search on SCOPUS (constraining the search to journal papers and conference proceedings in agricultural and biological sciences, earth and planetary sciences, and environmental sciences). According to the search results, in the last 10 years (from 2007 to 2017), an average of 500 documents that reported the use of kriging were published each year. Moreover, since 2000, there has been a rapid increase in the number of documents in which ML is specified as a topic, with kriging being overtaken in 2014.

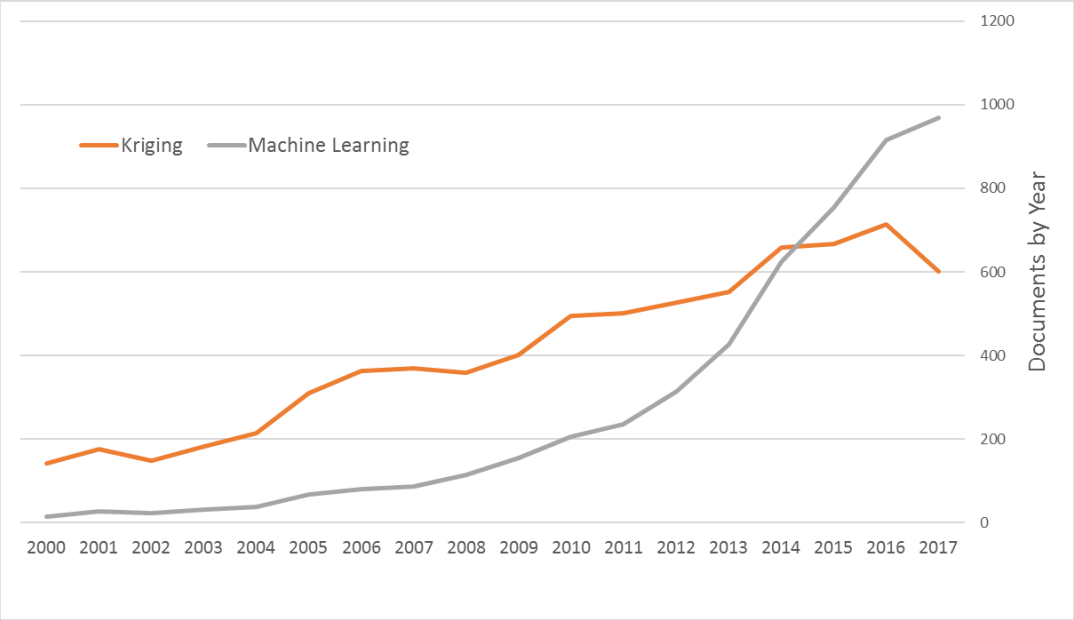


Figure 1: Results from a bibliographic search in SCOPUS. This figure shows the number of documents per year in which the word "kriging" or "machine learning" appeared as a topic.

These results clearly demonstrate that ML is quickly becoming the first and only choice for many researchers who are interested in ecological modeling, which poses a new and important challenge. Kriging has a "built-in" ability to provide researchers with not only a predicted value for each unsampled location but also a robust measure of uncertainty. This situation is often not the case for ML. Due to this inherent difficulty, many authors simply avoid discussing local uncertainty in their work (Verrelst et al., 2013). Arrouays et al. (2017) identifies the "evaluation of uncertainty of soil predictions" as "an aspect where most progress is needed" and laments that in many cases, "prediction intervals [...] have little practical use". However, this should be a priority for every mapping exercise,

where stakeholders must be sure that what is presented on the map is an accurate prediction that has a robust confidence interval.

The main objective of this work is to compare popular algorithms for SOC in terms of not only the overall accuracy but also the way in which they compute the local uncertainty. This comparison is of crucial importance for the scientific community because if we are only able to produce an average figure that is related to the overall uncertainty (i.e., cross-validation error), we may provide end-users with soil maps that are, in some areas, completely unreliable and unusable for practical purposes, such as designing remediation strategies.

To evaluate each method, we employed the same validation strategy as was implemented by Vaysse & Lagacherie (2017). We randomly excluded 25% of the samples from training and used them to test the accuracy of each algorithm. This operation was repeated 100 times.

Moreover, we evaluated the transferability. In this case, we split the dataset into calibration and test sets based on nearest-neighbor distances. This test attempts to mimic a realistic scenario in which practitioners must use a model that has been trained in one area to estimate locations that may be farther afield. We assessed each method in terms of the uncertainty and the confidence intervals it provides if it is used in areas that are far away from where it was trained.

After providing technical descriptions of the algorithms that are included in this comparison, we present a general overview of their advantages, disadvantages and limitations. We hope readers will better appreciate the differences among the methods and understand how and in what scenarios each method is more suitable than the others. Geostatistical interpolation, which includes ordinary and universal kriging, has the main advantages of being widely available across multiple software applications and relatively easy to perform. There are many applications that require users to have very little knowledge of the mathematical details of kriging because the complete interpolation process is partially or entirely automated; examples include the Geostatistical Analyst in ArcGIS (ESRI, 2011) and the automap package in R (Hiemstra et al., 2009). These applications enable users to automatically de-trend data, fit the variogram model and interpolate points to create the final map, which is a substantial advantage; however, it is also a concern because it may allow people to use a method they do not fully understand. Kriging, like many statistical models, requires data to be normally distributed; if this assumption is not satisfied, transformation and back-transformation are required (as in this work with SOC, which follows a log-normal distribution). Another extremely important assumption of kriging is that 100 points are necessary for modelling the variogram (Webster and Oliver, 2008). If this assumption is violated, kriging would still be possible, for example, by modelling the variogram using maximum likelihood (Kerry and Oliver, 2007); however, many commercial software packages do not include this advanced option. The main disadvantages

of kriging are that it is sometimes regarded as a method that does not perform as accurately as more complex methods and that it tends to produce smooth map surfaces (Veronesi et al., 2012) because kriging tends to reduce the variance of the estimates, in relation to the variance of the observed data, to reduce the impact of outliers. When dealing with a complex dataset, this may become an issue because it may hamper the ability of kriging to accurately estimate locations that are poorly represented in the dataset (e.g., when various land types are underrepresented because they are difficult to sample). Linear modelling or linear regression is another method that is relatively simple to apply and featured in many commercial software applications. Since it is one of the first statistical models to be taught at universities, it is also well accepted and understood. The main disadvantage of linear modelling is that it is a biased method. To facilitate understanding of this concept, we introduce the variance/bias trade-off, which is important for understanding the differences among models. Predictive algorithms model the dependent variable (i.e., SOC) as a function of several predictors. According to James et al. (2013), the error of the function that models the variable is the sum of two quantities: its bias and its variance. Bias refers to the approximation error of a function that can accurately fit only data that follow a strict pattern; for example, a linear model can only fit lines and will not chance shape to accommodate other data patterns. The linear model creates an error that is intrinsic to the inability of the function to change shape.

On the other end of the spectrum, we have algorithms that can change the way they fit data according to the pattern of the data, such as a spline that can adapt its curvature. These algorithms have more variance and include models such as random forest and boosted regression trees, which are used in this experiment. The main advantage of these algorithms is that they can model complex interactions between dependent and independent variables and non-linear patterns in the data, which can potentially increase the accuracy with which they can predict SOC at unsampled locations. Their main disadvantage is that they may also tend to overfit the data. Environmental data are affected by random variation or noise, which may decrease the accuracy of complex algorithms. These algorithms try to fit all points as closely as possible; with a noisy dataset, they may fit the noise along with the real signal. Overfitting can substantially decrease the accuracy of ML algorithms, which require robust validation to be successfully implemented for environmental modeling.

6.2 Materials and Methods

6.2.1 Study Area and Dataset

The study area, namely, Sicily, is a semiarid island that is located in the middle of the Mediterranean Sea. Sicily is the largest Italian region, with a surface area of 25.000 km^2, of which

approximately 60% is cultivated. The geology of Sicily reflects tertiary and quaternary modifications of parent materials such as clayey flysch, limestone, sandstone and gypsum, and coastal plains. Due to the complex geomorphology, there are a variety of soil families, which are mainly represented by Cambisol, Luvisol, Vertisol, Leptosols and Regosols (Costantini et al., 2013).

For this experiment, we used a dataset that was collected by the Sicilian Regional Bureau for Agriculture, Rural Development and Mediterranean Fishery. This dataset is part of the legacy soil data of Sicily (Lombardo et al., 2018, Schillaci et al., 2018a,2018b,2017b) and the selected sampling campaign is particularly suitable for this study because it is the most recent widespread sampling (2008). This dataset was collected within the project "Soil Map of Sicily at 1:250,000 scale", which employed a pedo-landscape sampling design (Fantappiè et al., 2011), which should be optimized for geostatistics. The 2008 campaign collected 414 topsoil samples from various land-uses and measured the SOC concentration and texture of each sample. In this work, we focus on SOC concentrations of cultivated soils, which are vulnerable to soil losses due to the erosion and consequent decrease in soil fertility, particularly in Mediterranean semi-arid regions such as Sicily.

The campaign, which was conducted in 2008, recorded mostly soil samples according to the horizon-based approach: surveyors dug small trenches, identified soil horizons and collected bulk samples for each of these. Since pedological processes may substantially vary in space, the depth of each horizon also changes spatially. However, the soil volume typically does not change under the effect of tillage; therefore, it is the volume that is most in danger of losing precious organic carbon. The first 30 cm of soil is typically referred to as top-soil. Since the purpose of this experiment is to compare various modeling algorithms, we homogenize the dataset by applying a power depth function, as suggested by Veronesi et al. (2014), and compute the average SOC content of the first 30 cm of soil. A mathematical function that is represented by a power curve was fitted to each soil profile. Via this approach, we estimated the SOC continuously with depth in each location. Then, the estimates from the first 30 cm were averaged to obtain SOC values for only the top-soil.

6.2.2 Validation and Transferability

As discussed in the introduction, this research has two main objectives; these objectives require different methodologies and validations. To accomplish the first objective, we relied on the robust validation technique that was suggested by Vaysse & Lagacherie (2017). A percentage of the samples (25%) were randomly selected from the dataset to form the validation set. Then, each algorithm was trained using the remaining 75% of the samples and used to predict the validation set. This operation was repeated 100 times to account for random variations. The root mean squared error (RMSE), mean absolute error (MAE) and the concordance correlation coefficient (CCC; Lin, 1989) were calculated.

The results are also presented in bar-charts, where the error bars represent 95% confidence intervals around the mean values. This graphical presentation allows readers to visually assess the differences among the models. Confidence intervals provide a robust approach to evaluating whether samples are significantly different without the need for a formal inferential test (Hector, 2015).

To fulfill the second objective, we conducted an experiment in which we tried to mimic a realistic scenario in which soil samples are used to estimate SOC in areas that may be distant from where we have direct observations. For this, we created a test set that is based on nearest neighbor (NN) distances. First, we calculated the NN distance for each point, thereby obtaining a distribution of distances. Then, we calculated the 95^{th} percentile of the NN distribution and included in the test set the points with NN distances that are greater than or equal to this value. In total, the calibration set includes 393 samples, while the test set includes 21 randomly selected samples, which have an average distance from their nearest neighbor of approximately 9.8 km (the average NN distance in the calibration set is 3.8 km). Although the sample size of 21 is not large, it is sufficient for the second part of the experiment, in which we simply compare algorithms in terms of the way in which they represent the local uncertainty.

The rationale behind this choice was to simulate a real scenario in which a professional is asked to provide SOC estimates in arable fields that are far away from locations where soil samples are available. In addition, we opted for this approach because it enables us to compare the local uncertainties that are obtained via several algorithms in areas that are as far away as possible from other observations, which simulates a realistic scenario. Although Sicily provides a complex case study, its wide range of soils and changes in elevation and gradient should provide satisfactory testing ground for this comparison. Our objective is to identify the differences among the algorithms in estimating SOC outside the training area (transferability) and to assess whether ML algorithms can provide the same robust confidence intervals as kriging.

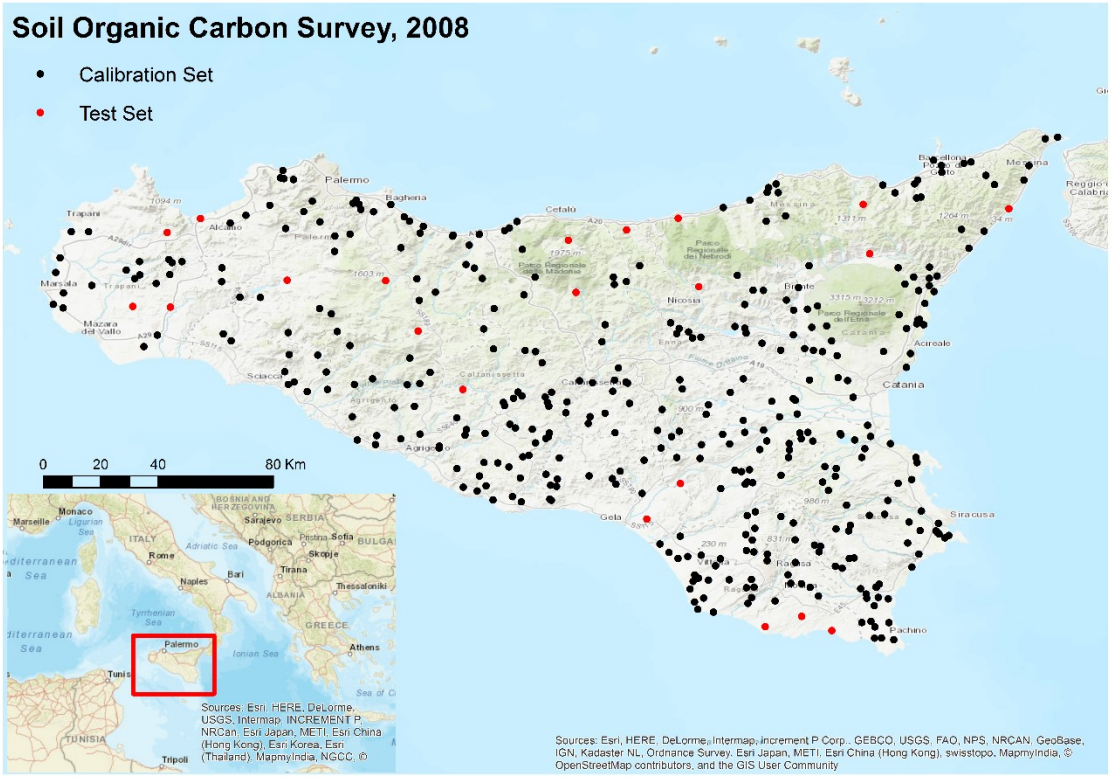


Figure 2: Locations of soil samples from the 2008 survey in Sicily, divided into calibration and test sets. The entire dataset was used for validation, while a set of 21 soil samples was used for the transferability test.

6.2.3Predictors

In the ML model, we included all the climatic variables that were produced for WoldClim v2 (Fick and Hijmans, 2017); Landsat optical satellite images (Chander and Markham, 2003); CORINE land-cover maps from 1990, 2000, 2006 and 2016 (Bossard et al., 2000); and all the MERRAclim climatic variables (Vega et al., 2017). In addition, we included the Shuttle Radar Topography Mission (SRTM) digital elevation model (Farr et al., 2007), from which we calculated additional geomorphometric derivatives: slope, relative slope, aspect, topographic position index, topographic wetness index, convergence index, cross-sectional curvature, diurnal anisotropic heating, flow accumulation, longitudinal curvature, length and steepness (LS) factor, valley depth, vertical distance to the channel network and channel network base level (using SAGA GIS, Conrad et al., 2015).

3.2.4 Algorithms

We tested the following algorithms because they are some of the most popular algorithms in the literature for soil mapping: ordinary kriging (OK), kriging with an external drift (KED), random forest

(RF), quantile regression forest (QRF), linear regression (LM), boosted regression trees (BRT), regression kriging based on random forest and boosted regression trees.

Ordinary Kriging (OK)

The ordinary kriging algorithm was selected because ordinary kriging is likely the most common form of kriging. This algorithm only requires the variable of interest; hence, it is a simple method to use and explain. OK performs well with this dataset (because of the sampling design) and produces readily available measures of local uncertainty. Hence, for a spatial location where we do not have direct observations, we can obtain an estimate for the variable of interest, along with a range of uncertainty where the true value for that location should lie.

Kriging is based on the variogram model, which provides a quantitative representation of the autocorrelation pattern in the study area (Webster and Oliver, 2008). The variogram is created by averaging the semi-variances, which are calculated as follows, of all pairs of points over distance bins:

$$\gamma(h) = \frac{1}{2} E[\{Z(x) - Z(x+h)\}^2] \qquad 1$$

In this equation, $\gamma(h)$ is the semi-variance between pairs of points that are separated by the vector h (referred to as the lag distance). By definition, the semi-variance is half the expected square difference between two values of the variable Z, namely, $Z(x)$ and $Z(x+h)$, that are separated by the vector h. Sampling design is extremely important for kriging. To obtain a robust representation of the autocorrelation structure, this algorithm requires samples to be separated by a range of lag distances, from short to very large, so that the variogram can well depict the spatial complexity of the area, which is important for obtaining a variogram model that is reliable and allows kriging to realize its optimal accuracy. The dataset that was sampled during the 2008 surveys presents exactly this characteristic, with areas where samples are clustered and areas where samples are farther apart. Hence, it is ideal for kriging and for this comparison.

Kriging is highly sensitive to skewed distributions because extreme positive values have a large impact on semi-variance calculations and, therefore, may render kriging estimates unstable (Webster and Oliver, 2007). For this reason, we examined the dependent variable SOC distribution and summary statistics (Section 3.1). Since we concluded that this distribution was highly skewed, we applied lognormal OK (Cressie, 1993), which interpolates the log-transformed values of the variable SOC. The back-transformed estimate for each interpolated location is obtained as follows (Laurent, 1963):

$$\hat{Z}_i = \exp\left(Z_i + \frac{\sigma_i^2}{2}\right), \qquad 2$$

where $\hat{Z}_i$ is the back-transformed estimated value for location i, Z_i is the estimated value that is obtained via lognormal kriging for location i, and σ_i^2 is the local lognormal kriging variance.

Similarly, the back-transformed local variance can be obtained via the following equation (Laurent, 1963):

$$\hat{\sigma}_i^2 = \exp(2 \cdot Z_i + \sigma_i^2) \times [\exp(\sigma_i^2) - 1], \qquad 3$$

where $\hat{\sigma}_i^2$ is the back-transformed kriging variance for location i. This value provides the range of uncertainty around the kriging prediction.

Multivariate Kriging

Other popular forms of kriging can include external factors, e.g., environmental predictors, into the model. These predictors can be used to model a trend; in this case, we consider kriging with an external drift. Another popular form of kriging is regression kriging, where kriging is used in combination with another algorithm. Regression kriging uses a regression (which may be ML) to compute estimates for each predicted location, followed by ordinary kriging to interpolate the residuals.

This method is well described in Hengl et al. (2003), where the authors remark that the additive nature of regression kriging (since residuals are added into the ML model) is transmitted to the local variance estimates, via the following equation:

$$\sigma_{RK}^2(i) = \sigma^2\{\hat{Z}_{ML}(i)\} + \sigma^2\{\hat{e}_K(i)\} \qquad 4$$

where $\sigma_{RK}^2(i)$ is the local variance for location i, $\sigma^2\{\hat{Z}_{ML}(i)\}$ is the local variance from the ML model and $\sigma^2\{\hat{e}_K(i)\}$ is the variance from the interpolation of the residuals. According to this equation, without a proper uncertainty estimate from the ML part of the equation, obtaining a reliable uncertainty map would be extremely difficult.

Kriging with an external drift was applied with a preliminary backward and forward stepwise feature selection approach that is based on the Akaike information criteria (Venables et al., 2013).

This kriging should increase the accuracy of the algorithm. Regression kriging was applied in the same way as by Vaysse & Lagacherie (2017). RF can compute a local uncertainty as a function of the spread of the values that are predicted by each tree in the forest (Veronesi & Hurni, 2014; Veronesi et al., 2016). Hence, it can be employed in regression kriging to estimate SOC values in test locations. The same procedure was also applied using boosted trees (Elith et al., 2008) as regressors. However, in this case, the standard algorithm was not employed, but the modified version, in which bootstrapping and random selection of predictors enable the algorithm to estimate the local uncertainty. As discussed previously, standard BRT does not provide local uncertainty by default; thus, it would be difficult to solve Eq. 4.

Random Forest

The random forest algorithm, which was developed by Breiman (2001), is very popular in the modeling community. In the past 10 years, more than 3000 environmental science papers in SCOPUS listed it as a topic. RF is based on ensembles of regression trees, which are classes of algorithms that partition the training set based on series of if-then rules that define classes of probabilities. RF is based on CART (Breiman, 1984), which fits a single tree using the entire training set. This method is well established and easy to interpret, but not highly accurate (James et al., 2013). RF overcomes the accuracy issue by fitting multiple trees and using their median value as the final predicted value. To realize this, RF employs a simulation that is based on bootstrapping (Hastie et al., 2001) as a resampling technique to generate multiple random realizations of the original training dataset. Moreover, not all the predictors are used to fit individual trees; RF randomly selects a specified number of predictors for each tree, which is a third of the total number of predictors by default and increases the accuracy of the algorithm because the trees are not correlated with one another (James et al., 2013). A schematic representation of the RF algorithm is presented in Fig. 3.

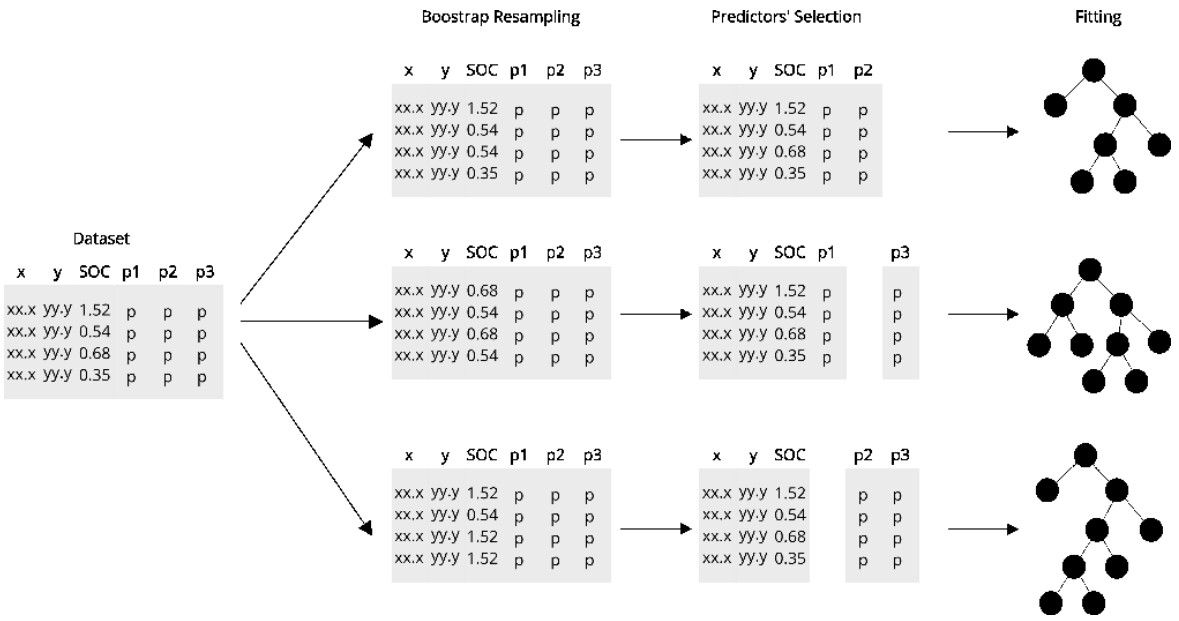


Figure 3: Graphical representation of the random forest algorithm. At each iteration, a slightly different training set is used, which is generated by bootstrapping, and only a specified proportion of the predictors (p) are included.

Since multiple trees are fitted to slightly different training set and the trees are not correlated with one another, RF is appealing to the spatial modeling community. Even though the final predicted value for a test location is the median value over all regression trees, the standard deviation of the estimated values of the forest can also be useful for assessing the local uncertainty.

RF was tested with a preliminary feature selection approach that is implemented using the recursive feature elimination algorithm (Guyon et al., 2002). This algorithm uses internal three-fold cross-validation and applies RF, which provides a measure of variable importance, to rank and extract only the most important predictors for the model, which reduces the noise in the predictors, thereby increasing the accuracy of the RF model. Then, the algorithm was trained on the calibration set using only the best predictors and fitting 1000 trees.

Linear Regression

This linear regression algorithm was included as a baseline method, particularly for the ML algorithms. However, since there are strong linear correlations between predictors and target variables in many environmental datasets, satisfactory performances from linear models are not surprising. Moreover, it is generally recognized in the statistics community that model simplicity is an important aspect to consider (Samulesson et al. 2017). Simple models are typically easier to explain and understand; therefore, if two algorithms perform with similar levels of accuracy, the simplest should be preferred. This algorithm was applied by including a preliminary backward and forward stepwise

feature selection method that is based on the Akaike information criteria (Venables et al., 2013). This algorithm is also one of the algorithms we modified, by employing the RF framework to compute its confidence intervals.

We created a simulation when bootstrapping and a random selection of predictors were employed to train the algorithm on various datasets at each run so that we would be able to assess the consistency of the estimates in relation to changes in the training set. This process was repeated 1000 times to obtain the same number of results as with RF and QRF. From this distribution of equiprobable values, we computed the median and standard deviation, which provide an estimate and confidence interval for each test location.

Boosted Regression Trees

Boosted regression trees is a relatively new method (developed by Elith et al., 2008); the first entry in the SCOPUS database is from 2005. However, it is rapidly gaining momentum and in the past 10 years, it was discussed in approximately 600 environmental science publications. This algorithm operates in a fundamentally different way compared to RF. Boosting is initialized by fitting a single regression tree to a subset of the entire dataset and evaluating its performance on the remaining data. The next iteration fits another tree; however, in this iteration, the objective is to decrease the error from the previous step. This process continues until adding more trees does not provide any improvement in accuracy (Elith et al. 2008). Hence, each successive tree is strongly correlated to the previous tree and, therefore, computing the variance of their estimates makes little sense. For a thorough explanation of this algorithm, please refer to Hastie et al. (2001); a schematic representation of BRT is presented in Fig. 4.

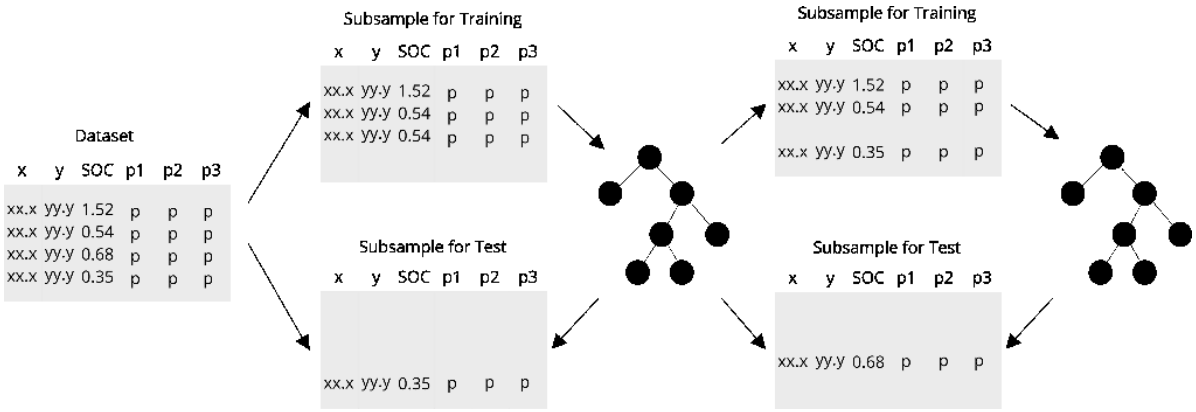


Figure 4: Graphical representation of boosted regression trees.

Since the trees are correlated, BRT cannot be used in RF; thus, it is not possible to extract any measure of uncertainty. Hence, BRT is another promising candidate for modification using the RF framework. The procedure that we followed is the same as for linear regression.

The R code for the modified versions of linear regression with feature selection and BRT is available on GitHub at: https://github.com/fveronesi/ModifiedGBM_LM

6.3 Results

6.3.1 Summary Statistics

Table 1 lists the summary statistics of the SOC data. The results demonstrate a skewed distribution with a wide range, which may create issues in computing the variogram and is the main reason why we apply the lognormal ordinary kriging.

Table 1: Summary statistics of soil organic carbon for the 2008 sampling campaign in Sicily.

Summary Statistic	**SOC Value in %**
Mean	1.51
Median	1.26
Standard Deviation	0.88
Interquartile Range	0.67
Range	0.22 – 8.79
Skewness	3.02
Kurtosis	18.46

6.3.2 Validation

The results of the cross-validation are presented in numerical format in Table 1 and in graphical format in Fig. 5.

Table 2: Results of the validation in which 25% of the samples were excluded and the process was repeated 100 times. The results are presented as average values, along with their standard deviations over the 100 replicates. The methods are ordinary kriging (OK), kriging with an external drift (KED), random forest (RF), quantile random forest (QRF), boosted regression trees (GBM), linear model (LM), regression kriging with random forest (RF_RK), and regression kriging with boosted regression trees (GBM_RK).

	OK	**KED**	**RF**	**QRF**
RMSE	0.68±0.27	0.67±0.28	0.73±0.35	0.74±0.35

MAE	0.48±0.13	0.48±0.16	0.48±0.16	0.48±0.15
CCC	0.4±0.17	0.43±0.26	0.33±0.22	0.31±0.21

	GBM	**LM**	**RF_RK**	**GBM_RK**
RMSE	0.72±0.28	1.01± 1.26	0.74±0.36	0.75±0.3
MAE	0.51±0.13	0.65±0.33	0.49±0.13	0.52±0.14
CCC	0.35±0.2	0.32±0.2	0.32±0.23	0.31±0.2

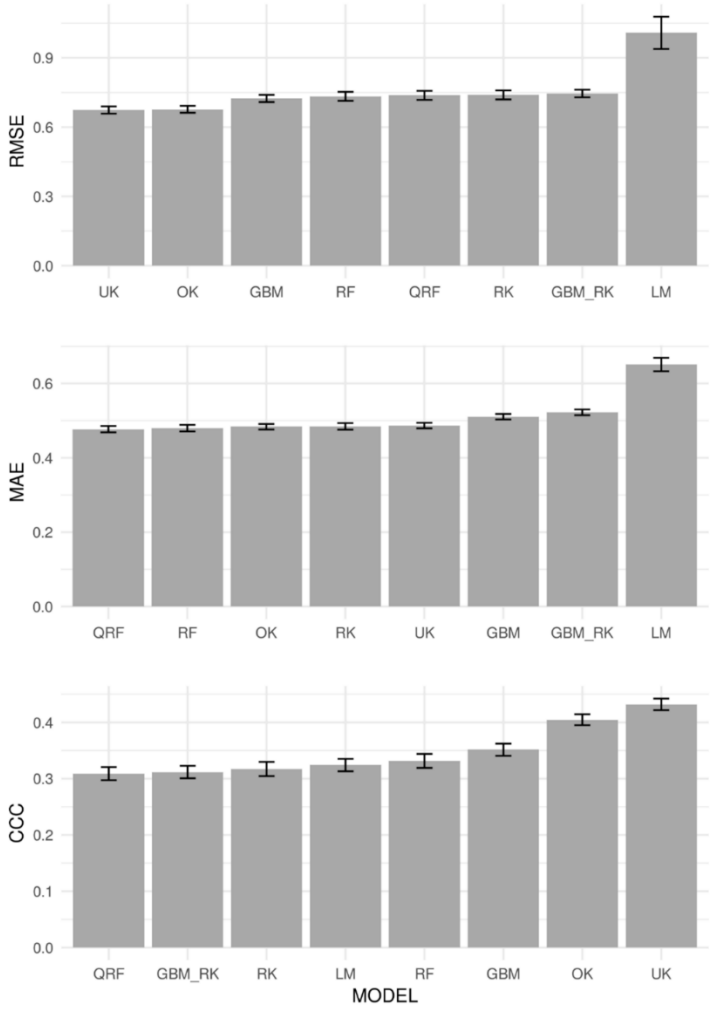


Figure 5: Bar chart that shows the mean values of RMSE, MAE and CCC with 95% confidence intervals. The bars are ordered according to increasing mean.

The results demonstrate that ordinary and universal kriging performed better than the other methods that were evaluated in this work. From the confidence intervals in Fig. 5, we conclude that OK and UK yield significantly lower values of RMSE and significantly higher values of CCC compared to tree-based ML methods, which rank second best. For MAE, the picture is more complex because several methods (QRF, RF, OK, UK and RK) present overlapping confidence intervals;

therefore, statistically, their accuracies cannot be distinguished. Methods QRF and RF are ranked first and second, respectively, in terms of MAE, but only fifth and forth in terms of RMSE. Therefore, the residual distributions for these two algorithms feature extreme values, namely, very large positive or negative residuals. Because the root-mean-square error squares the residuals, it is highly affected by extremes; that is the reason for these differences. Hence, RF and QRF might be accurate on average; however, in locations where they are not accurate, their errors can be much larger compared to other methods.

6.3.3 Transferability - Estimating the Test Set

We divide the original dataset into a calibration set, which will be used for training, and a test set, which consists of locations that are at least 9.8 km away from their nearest neighbor. Each algorithm will be discussed separately in this section. To facilitate the interpretation of the plots, in Fig. 6, we present the locations of the test samples with a numerical ID that can be traced back to the residuals plots that we show below, along with the SOC values to facilitate the interpretation of areas of higher/lower accuracy for each algorithm.

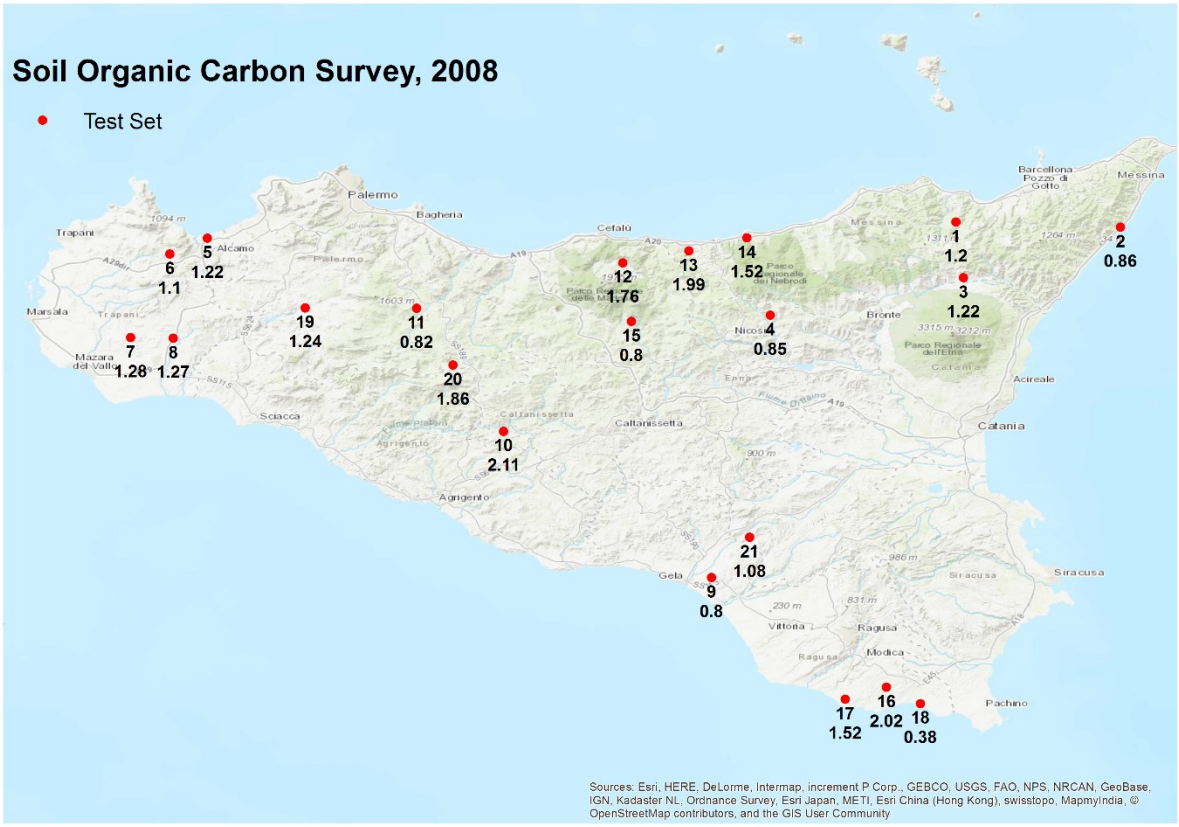


Figure 6: Locations of the test samples. Each is identified by a numerical ID and the local SOC measurement is specified.

Ordinary Kriging and Kriging with an External Drift

Traditionally, kriging has been considered the only DSM algorithm that can provide realistic and robust confidence intervals for each estimated location. This technique remains highly popular among environmental modelers and widely used in research. Moreover, both ordinary kriging and kriging with an external drift scored very well in the cross-validation. Therefore, these algorithms will provide a solid basis for understanding how ML deals with local uncertainty and reports confidence intervals.

For both algorithms, the SOC concentration was log-transformed to comply with the assumption of normality and Eq. 2 and 3 were used to compute the estimated (back-transformed) values and the standard deviations for the test locations. The results are presented in Fig. 7, where the residuals at each of the 21 test locations are plotted. This plot shows the 21 locations, which are ordered by ID and plotted one on top of the other along the Y-axis; on the X axis, the residuals are plotted. The red line represents the line of perfect fit, when the residuals are equal to 0 (units are SOC %). The empty dots represent the residuals for each point, which are normalized by the observed values. Hence, a dot on the left side of the red line corresponds to a negative residual and, thus, the estimated value exceeds the observed value; the opposite applies to dots on the right side of the red line. Also shown on the plot are the MAE, which is computed by comparing estimates and observed values for the test locations, and the average width of the confidence intervals. The number of times that confidence intervals cross the line of perfect fit is also specified on the plot, which is useful for evaluating the usefulness of these ranges of uncertainty in practice. Practitioners, including agronomists and advisors, need to know that the values they are getting from a map are the true values of SOC concentration; otherwise, the map would be useless for practical purposes.

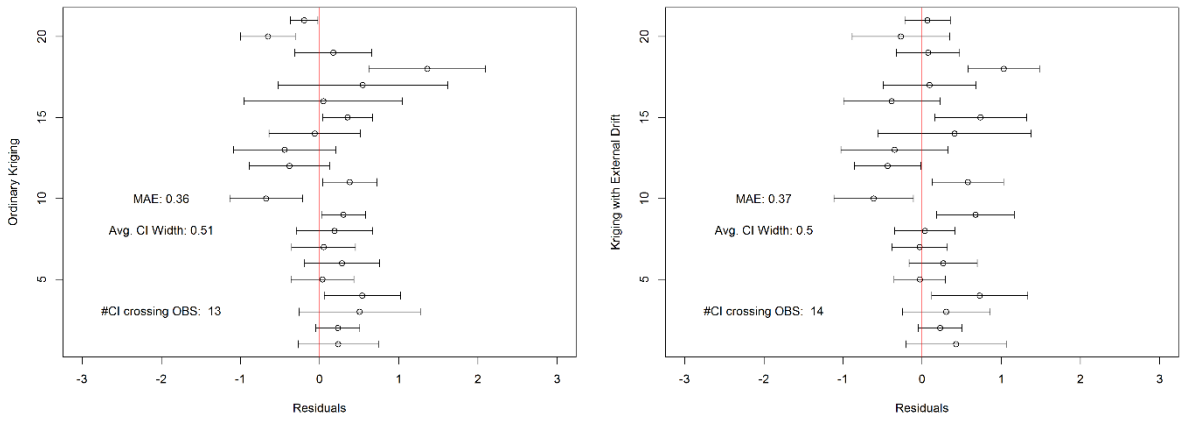


Figure 7: Residuals for the 21 test locations, which were computed from the estimates that were obtained via ordinary kriging and kriging with an external drift. The red line corresponds to the residuals being equal to 0.

Since these two methods were identified as the best predictors in the standard validation test, we consider their results as our baseline, with which we will compare to all results from the other algorithms. Ordinary Kriging yields an MAE of 0.36%, an RMSE of 0.47% and a CCC of 0.3, while UK yields an MAE of 0.37%, an RMSE of 0.46% and a CCC of 0.23. Therefore, in terms of the average error, both models yield satisfactory results; however, in terms of correlation, they perform worse than validation. In terms of confidence intervals, on average, the standard deviations around the estimate are ±0.5%; hence, the estimated values that are provided by kriging are typically very close to the real observations, or at least within an acceptable margin of error. There are cases where the confidence interval is large, for example, samples 16 and 17 for OK and sample 14 for KED. According to the map above, samples 16 and 17 have higher values compared to sample 18 (0.38% SOC), which is relatively close. This result may create difficulties for kriging, which estimates an average value for the whole area. KED performs better in dealing with this scenario, but is unable to estimate sample 14, which is on the coast but has similar values to nearby observations. There may be something in the predictors that reduces the accuracy in this area. In terms of reliability of the confidence intervals, 13 confidence intervals for OK and 14 for KED cross the line of perfect fit; hence, in approximately 66% of the cases, the range of uncertainty that is provided by the map is reliable.

Random Forest and Quantile Regression Forest

As discussed previously, RF and QRF operate similarly by fitting ensembles of regression trees; thus, they can assess local uncertainty. In this experiment, the two algorithms are compared by estimating the test set and examining the results, which are presented in Fig. 8. This figure was created in the same format as the previous figure so that the results can be compared easily. The confidence intervals are computed as the standard deviation of the distribution of the values that are estimated by each tree in the forest for RF and by extracting the 25th and 75th percentiles from QRF.

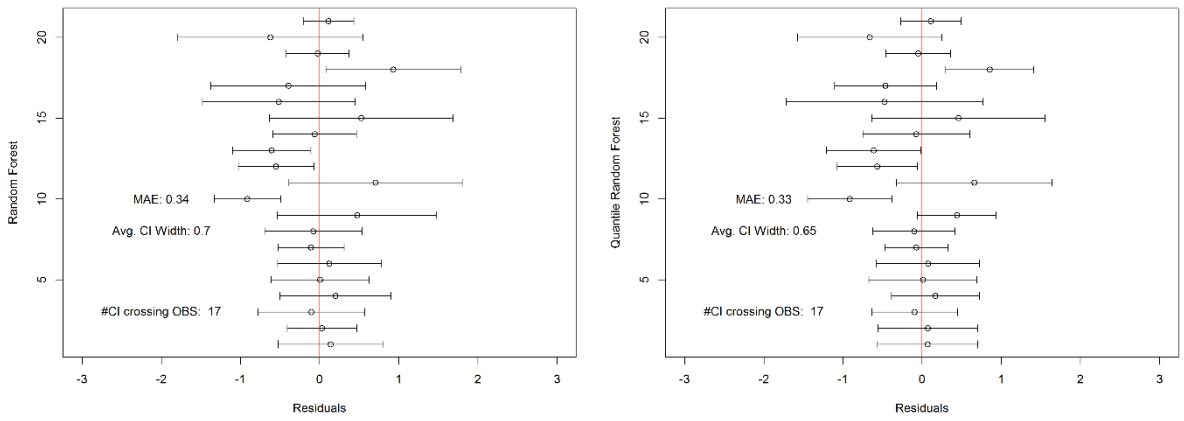


Figure 8: Residuals for the 21 test locations, which were computed from the estimates that were obtained via random forest and quantile regression forest. The red line corresponds to the residuals being equal to 0.

RF yields an MAE of 0.34%, an RMSE of 0.45% and a CCC of 0.11, while QRF yields an MAE of 0.33%, an RMSE of 0.44% and a CCC of 0.16. These results are similar to those that were reported in the general validation. However, from this work, we can also assess the way these algorithms report their local uncertainty, namely, the way in which they represent their confidence that what they are reporting is the true value of the SOC in a specified location.

In total, 17 confidence intervals cross the line of perfect fit; hence, in approximately 80% of the unsampled locations, both algorithms can provide reliable (meaning the true value of the sample is included in the confidence interval) uncertainty ranges. However, the downside is that, on average, the confidence intervals have ranges of ±0.7% SOC, which may not seem like a large difference compared with kriging, but according to Fig. 8, there are samples for which the potential range of confidence is large. For example, sample number 15 for both algorithms has a confidence interval of approximately ±1% SOC, which is a large uncertainty for an observation of 0.8% SOC. The range of confidence should be considered when using these methods because it can negatively affect the way in which SOC is represented and evaluated.

Linear Modelling

The linear modelling algorithm is the first algorithm for which we employed the RF framework to compute its confidence intervals. According to the cross-validation, linear regression is the least accurate of the algorithms that were evaluated in this experiment. However, it provides a satisfactory baseline for the other methods since it is widely considered the simplest mapping technique.

In terms of accuracy, LM yielded an MAE of 0.4%, RMSE of 0.51% and CCC of 0.09. The average standard deviation around the estimates is close to that computed via kriging. In terms of the reliability of the confidence intervals, linear regression outperforms kriging, with 16 confidence intervals crossing the line of perfect fit, which is only one less compared to RF and QRF.

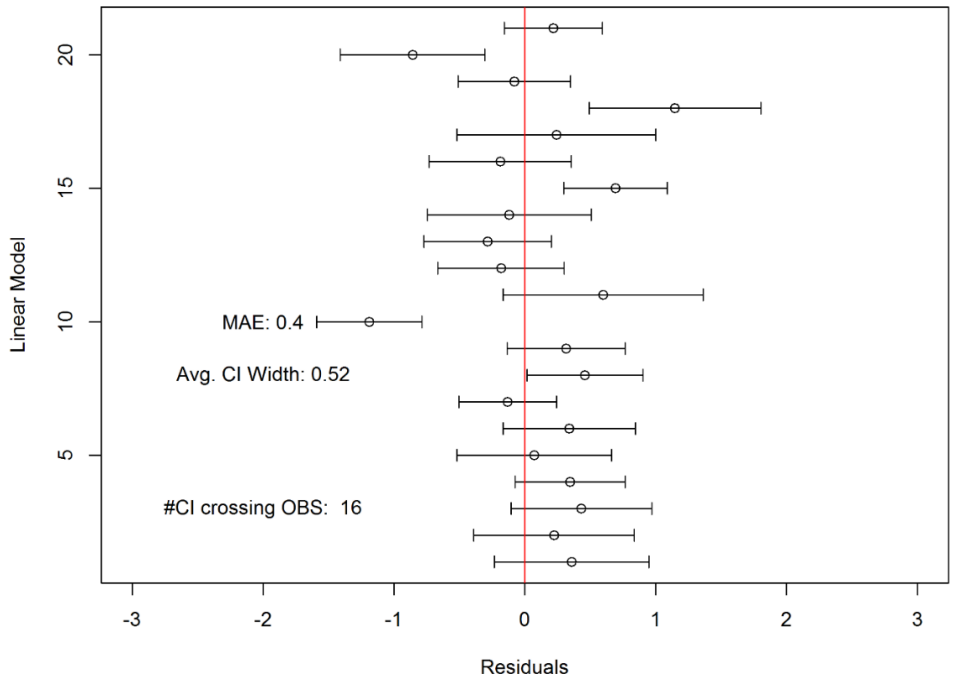


Figure 9: Residuals for the 21 test locations, which were computed from the estimates via the bagged linear regression algorithm. The red line corresponds to the residuals being equal to 0.

From Fig. 9, it is difficult to conclude that this method outperforms the method that was evaluated above. For example, numerous residuals are above or below 1% SOC and its correlation coefficient is low; according to the image, various values are predicted with very low accuracy, which must be carefully considered since a difference of 1% SOC may become significant when the soil map is employed for agricultural management purposes. Sample 10 has a particularly low residual value. This sample has the highest percentage of SOC among the samples that were selected for testing, which may demonstrate that LM is unable to estimate relatively high values of the variable of interest in cases where these values are at the extremes of the distribution of the values that are employed for training. Thus, LM is unsuitable for use as a reliable and robust estimator for heterogeneous datasets of this type.

Boosted Regression Trees

BRT is the other algorithm of which we proposed a modified form that incorporates a local uncertainty estimation into the RF simulation that is described above. The results of applying this modified implementation of the BRT algorithm to the test set are presented in Fig. 10.

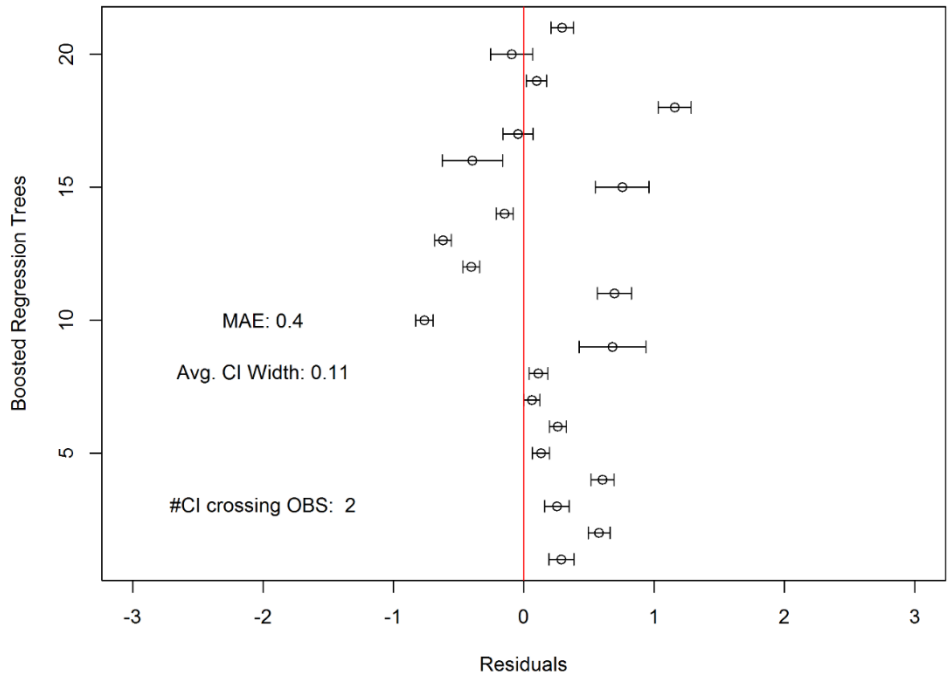


Figure 10: Residuals for the 21 test locations, which are computed via the bagged boosted regression tree algorithm. The red line corresponds to the residuals being equal to 0.

The results were an MAE of 0.4%, an RMSE of 0.5% and a CCC of -0.01. Therefore, BRT is not capable of estimating, with any degree of confidence, data that are far from the area where it was trained (at least for this dataset). This inability to estimate distant data is clear from the image, in which large discrepancies between observed and predicted values are observed.

However, despite the application of the same technique as was used by RF and QRF, the estimates from BRT tend to have a very small spread. In other words, the confidence intervals are highly precise but not highly accurate, which should be considered carefully when mapping SOC. It may be that for some datasets, BRT performs well, and in such cases, its estimates would be close to the real values at any unsampled location because of the narrow confidence intervals. However, in cases where BRT does not perform as well, we must be careful when using its estimates because we may be providing practitioners estimates that are not as reliable and robust as we might believe. Only 2 confidence intervals crossed the line of perfect fit and most of them have residuals that exceeded 0. Thus, in most cases, a soil map that is created with BRT will provide SOC values that are too optimistic.

Regression Kriging

Regression kriging with full estimation of the local uncertainty of the map is now possible, even with BRT. In this section, we present the results for regression kriging that is applied using both RF

and BRT as regressors to estimate the test set. The results are presented in Fig. 11. As discussed previously, the confidence intervals are computed by solving Eq. 4; thus, they are the sum of the error from RF and BRT, which is computed as the standard deviation of the estimates for each tree in the forest, and the kriging variance from the residual interpolation.

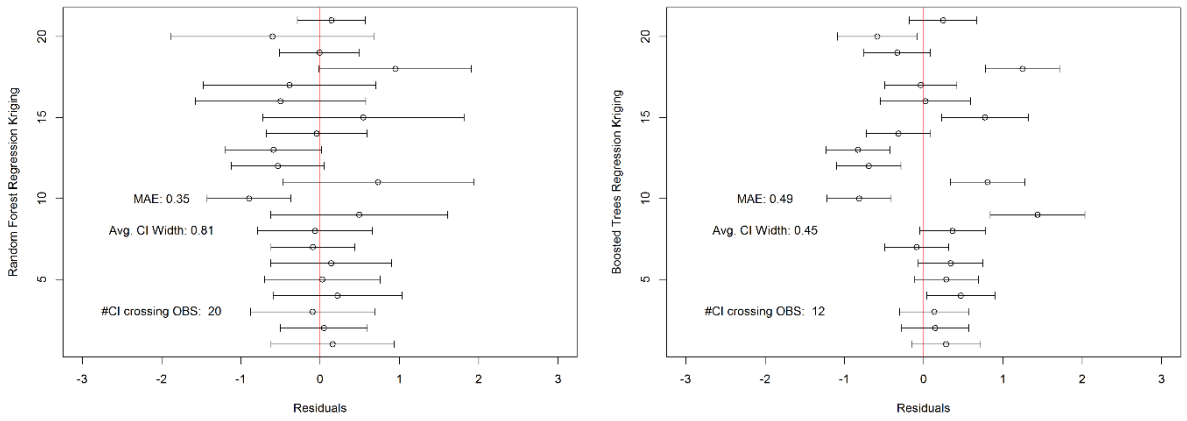


Figure 11: Residuals for the 21 test locations, which are computed via regression kriging based on RF (left) and BRT (right). The red line corresponds to the residuals being equal to 0.

RF produces relatively accurate estimates, but computes very wide confidence intervals, which does not change when kriging is employed in a regression kriging framework; instead, the widths of the confidence intervals increase. With regression kriging, all the confidence intervals have the lower value on the left side of the plot; thus, if we only consider the lower bound, we will always underestimate and, therefore, be conservative, for example, when planning soil management practices. There may be scenarios in which this is not optimal; however, there are datasets for which being conservative is appropriate.

Regarding BRT-based regression kriging, kriging of the residuals can partially overcome the issues that were identified previously. Regression kriging increased the widths of the error bars (by an average of 0.45% SOC), with 12 of them crossing the line of perfect fit. In terms of accuracy, for this dataset, BRT, even with kriging of the residuals, still performs worse than the other approaches. However, for sampling designs that are optimized for this method, the increased reliability of the error bars renders this algorithm more attractive for DSM than the standard version of BRT.

Averaging over All ML models

To complete our comparison, we evaluate the performance of a model that averages the estimates from all the algorithms. Several authors (Huang et al., 2012; Zhang, 2014; Zhang et al., 2017a; Zhang et al., 2017b) used this approach to incorporate uncertainty into ML modelling. We estimated the test locations using all the algorithm that are presented above. Then, we averaged their values; this

provides the new predicted value for each test location. The uncertainty was computed by calculating the standard deviation of the estimates from all the models. The results are presented in Fig. 12.

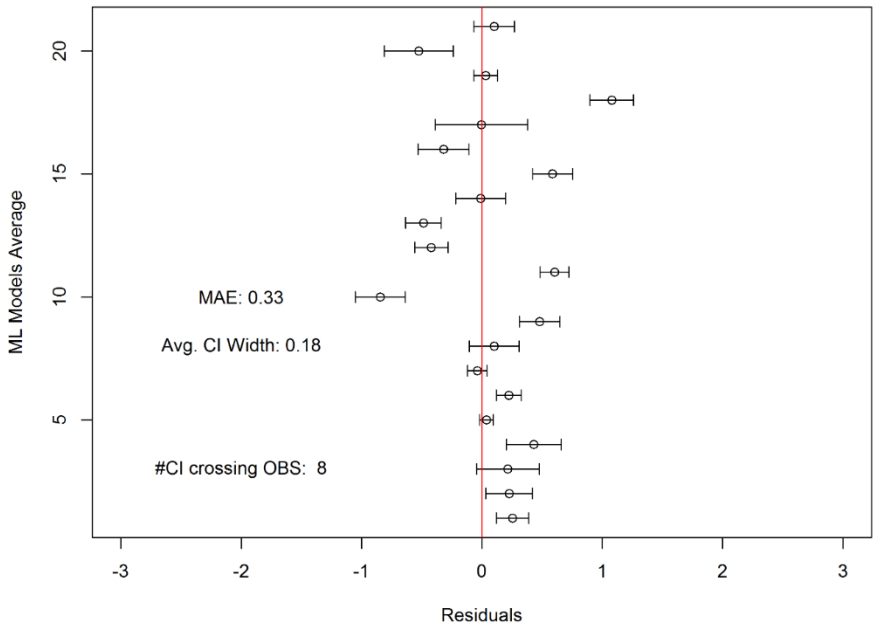


Figure 12: Residuals for the 21 test locations, which were computed by averaging the results over all ML methods that were evaluated in this experiment. The red line corresponds to the residuals being equal to 0.

The results are similar to those of RF and QRF, with MAE of 0.33%, RMSE of 0.44% and CCC equal to 0.17. However, the results are also erratic, with some locations badly misrepresented and confidence intervals that are probably too narrow to be used in practice.

The average width of the error bars is just 0.18% SOC; hence, in cases of relatively high or low residuals, the range of uncertainty crosses the line of perfect fit and, thus, is not reliable for practical purposes. Overall, these results demonstrate that averaging the results of multiple ML algorithms may be a satisfactory approach to realizing high accuracy and providing a measure of local uncertainty.

6.3.4 Discussions

Since the sampling design was based on geostatistics, kriging performing best is not surprising. The sample requirements for kriging and ML are fundamentally different: kriging requires a sampling design that is optimized for fitting the variogram (i.e., good mix of small, medium and large lag distances), whereas ML requires a sampling design that is optimized for predictor coverage, along the lines of the Latin hypercube (McKay et al., 1979), and extensively used for soil mapping (Ballabio et al., 2016; Ließ, 2015; Minasny and McBratney, 2006; Mulder et al., 2013). Therefore, this dataset may not be optimized for ML, which is the main driver of the lower performances of these methods.

The main issue with standard validation frameworks, such as the one that we employed or the standard cross-validation, where the dataset is randomly divided into non-overlapping folds (James et al., 2013), is the random sampling of the test set. Therefore, in areas with a high sample density, the test set may include samples that are very close to samples that are used to train the algorithm (Veronesi et al., 2017). This sample proximity may inflate the performance of kriging-based methods. Moreover, this is not a realistic approach to evaluating the accuracy of the algorithms; soil samples are not often used to estimate locations that are close to where we sampled since we may assume a degree of spatial similarity, but are used to estimate locations that may be farther afield and for which we do not have any information. To robustly evaluate each algorithm, we must construct a more realistic validation framework that accounts for DSM being employed to estimate areas for which no direct observations exist. We tried to realize this objective in the second part of the experiment (Section 3.3).

According to the main cross-validation, OK and UK obtained significantly more accurate results compared to all the other algorithms. The rankings of the algorithms differ among accuracy indices. In terms of only RMSE and CCC, OK and UK performed best; however, the MAE results appear to be less diverse. RMSE is more strongly affected by large residuals because of the squaring; therefore, it is a useful index for identifying algorithms that are accurate, on average, but produce large residuals. This situation is probably what occurs here, with OK and UK being capable of reducing the impact of extreme values. Since the 2008 sampling campaign was carried out using a sampling design that was developed via geostatistical techniques, this result is not surprising and demonstrates that the selection of the estimation method must be based on careful considerations. It is good practice to test algorithms that have already been used in the literature for similar purposes; however, it may be counterproductive to only consider an algorithm because it is new and has never been used before in a specified field. This situation sometimes occurs in science, where methods are selected only because they are fashionable and not based on real and tested experiments. In this case, for example, since the sampling design was based on geostatistics, very advanced methods such as deep-learning could be useful, but still less accurate than kriging, which can also provide robust and well-tested uncertainty estimations. In summary, even though we now have access to advanced algorithms, we should not use them blindly without fully considering what we are trying to achieve.

The ranking from the main cross-validation is reproduced almost exactly in the second part of the experiment, where algorithms were tested in locations that are far from the training area. This ranking demonstrates that even though we are only estimating 21 samples, the results that we obtained in terms of confidence intervals are reliable. However, this part of the experiment was not designed to evaluate the accuracy of the algorithms, only the way they represent uncertainty. This experiment

enabled us to obtain extremely useful information about the way error is reported by each algorithm. Soil maps are not produced only for the sake of creating beautiful geographical visualizations; they provide accurate and reliable information on soil properties that decision makers can use to plan interventions. When we perform a cross-validation, which is a method that is widely used to test mapping algorithms, we can only provide a measure of the overall accuracy of our map. The indices that we used in Table 2 only provide an average value of the residuals, which is suitable for comparing methods but provides little information on the accuracy of the map. We could better evaluate the residuals by either plotting them in a histogram or creating a map that shows the residuals in each sampled location using a color scale, which would present a more precise picture of the geographical distribution of the error of our model. However, it will provide almost no information about the error that would be generated when estimating the value at a location that was not encountered by the algorithm during training.

Several authors (Guio Blanco et al., 2018; Poggio and Gimona, 2015) interpolate the residuals that are computed after cross-validation to develop an uncertainty map for ML algorithms that are not capable of generating one. For example, in Song et al. (2016), the authors applied a deep-learning neural network to model the spatio-temporal pattern of soil moisture; the uncertainty was computed via sequential Gaussian simulation using the residuals. This approach is suggested in Kanevski et al. (2009); however, it may be too optimistic. The problem is related to the way in which cross-validation works, which is through a random selection of locations that does not account for spatial auto-correlation, but assumes samples to be independent. This validation approach enables algorithm testing in areas and locations that are very close to the area and locations that are used during training, as demonstrated in Veronesi et al. (2016). This validation approach may produce results that are too optimistic; therefore, it is not suitable for constructing a reliable uncertainty estimation map. The only way ensure that the uncertainty map we provide to the end user is reliable is to use an algorithm that has been thoroughly tested and is known to produce reliable error maps.

According to the literature, kriging provides a robust standard deviation around its estimates. In this example, it performs well in terms of accuracy; therefore, we used this algorithm as a baseline model and everything was compared with it. RF is the only ML algorithm that can report a measure of uncertainty for each value it estimates. According to our results, its standard deviation values are similar to those that are produced by kriging; even though the interval is slightly wider, it has eco-physiological meaning for the studied indicator and the climate (Lombardo et al., 2018, Schillaci et al., 2018a, 2017a). Hence, RF can be used to replace kriging in scenarios in which the latter is unsuitable. Since its confidence intervals cross the line of perfect fit most of the time, RF can be employed when we need to create maps of areas that are far away from where we trained the model.

QRF generated the most reliable confidence intervals. This result is consistent with findings of Szatmári and Pásztor (2018) which also found that QRF outperformed other kriging-based models. These confidence intervals enable practitioners to select an optimal safeguarding level. For example, users can decide to only plan management practices using the most conservative estimates, namely, the lower values. In the image, most confidence intervals have lower bounds that are above -1, which should provide an acceptable safeguarding level for planning, without being over-conservative. With RF, all the lower bounds have values that are below the true SOC concentration for a specified predicted location.

In this research, we also demonstrated how the RF framework can be used successfully to incorporate an uncertainty estimation into other algorithms. Furthermore, we developed modified versions of BRT and linear models; theoretically, the same code can be used to modify any currently used algorithm. Linear modelling was included to represent the lower end of the accuracy spectrum because we assumed that it would be the least accurate method. BRT is becoming popular in the environmental modelling community. Despite being included in a simulation where samples were bootstrapped and predictors were randomly excluded, BRT produced very consistent estimates. Thus, its performance is only slightly affected by changes in the predictors or the training dataset, which is very interesting because we sometimes create soil maps with samples that we did not collect ourselves and that are likely affected by errors that are difficult to assess. In such cases, BRT could be the best option since its estimates will not be strongly affected by these issues. However, we must be very careful when using BRT because the uncertainty estimates that it provides may not be reliable.

6.4 Conclusions

Agro-ecosystems are the pillar of our economy; however, their health closely depends on satisfactory soil management and factors such as parent materials and the weather regime. Knowledge of SOC and its stock requires detailed information. Moreover, the development of an accurate mapping methodology is very important for avoiding soil degradation and optimizing the crop yield. The main objective of this work was to compare SOC mapping algorithms in terms of not only their average accuracy but also their ability to provide reliable confidence intervals in unsampled locations that may be far away from the training area. Since many ML algorithms do not have a “built-in” function for computing local uncertainty, in this work, we propose to adapt the RF framework and modify existing algorithms to incorporate local uncertainty estimation.

In terms of validation, OK and UK were the most accurate mapping methods, which is not surprising since the sampling design was optimized for geostatistics. However, in terms of local

uncertainty, RF and QRF generated the most reliable confidence intervals, namely, intervals that included the observed value of SOC most of the time (even when estimating areas that are far away from their training locations), which is potentially important for practical uses; however, they must be handled carefully since the confidence intervals are also very wide. BRT, which did not perform well on this dataset, has a much lower variance compared to RF; hence, its estimates are much more robust against changes in the predictors, which is a desirable property in environmental and soil modelling. However, computing confidence intervals that, in many cases, do not include the true value of the unsampled location must be considered when employing BRT for mapping. In cases where BRT is highly accurate, its uncertainty map will be extremely useful. However, in other scenarios where BRT is not as accurate, we must be careful because its uncertainty map may be practically useless.

6.5 Acknowledgments

The authors would like to thank Maria Gabriella Matranga, Vito Ferraro and Fabio Guaitoli from the Regional Bureau for Agriculture, Rural Development and Mediterranean Fishery, the Department of Agriculture, Service 7UOS7.03 Geographical Information Systems, Cartography and Broadband Connection in Agriculture, Palermo. We are also very grateful to the editorial team of Ecological Indicators and the reviewers who took the time to provide very useful comments that helped us substantially improve the manuscript.

4. Dissertation conclusions

Soil data and quantitative models based on spatial covariate are growing in number and quality across the world. Share the knowledge among public management needs, stakeholders and others is still a challenge for the management of the environment. This especially occurs when there is a need to manage highly heterogeneous areas. The health of agroecosystems is due to many factors, one upon all anthropic control, while preserving the income of the area and reducing land degradation. Agricultural compartment has been often a weak chain of the economy; however, it has recently become easier to monitor, and to manage, thanks to the impressive earth observation programmes. Soil properties, especially SOC and SOC stock, are extremely important for crop production, and they can be used as a metric of ecological services. With this work, a particular aspect of regional assessment of soil resources was studied to enhance cropland soils monitoring. The results of this dissertation are: (i) an improvement of the modelling procedures to estimate SOC in Mediterranean areas; (ii) the production high detailed and accurate maps of SOC distribution; (iii) the quantification of the change in SOC over time. These maps can be used at a regional extent for management purposes. The link between soil managers and policy makers have improved the way to face land degradation and today subsidization, especially in U.E., is not only limited to sustain commodities production but also to support the implementation of best management practices with the aim to preserve the environment and yield potential in time and space. Interesting results about the SOC dynamics in the cropland were observed in the study. Knowing the space-time variation of this trait (both in term of SOC concentration and stock) is a prerequisite for sustainable land management, where sustainability includes

- the increase of soil C stock, which is directly related to the reduction of greenhouse gas concentration in the atmosphere and indirectly to land susceptibility to soil loss;the maintenance of yield potential, since SOC acts as an easily-accessible soil reservoir of nutrients for plants and soil biota and provides a wealth of ecosystem services to plant growth outliers is to be performed when the amount of data is high and from different sampling campaigns.

This can lead in individuating error in reporting the unit of measurements (e.g. g kg^{-1} confounded with %) or affected by a systematic error. Some of the cases were highlighted, and it has been demonstrated that their deletion can lead to a reduction of modelling error in the estimation and/or relative uncertainty.

Part of the database from the chapter 3, in particular those related to the 0-30 cm soil layer, were used for an estimation process of the SOC stock in Sicily (chapter 4). Since the database included data of samples collected from 1968 to 2008, the most of which in 1993 and 2008, the SOC stock computation was made referring to the year 2000 as an average of the historical records of the legacy data. In addition, soil bulk density was estimated with an in-situ pedo-transfer function. The model built showed a high reliability and clearly highlighted main predictors of the SOC stock. Part of these regressors, as expected, were the same of those found in the SOC mapping experiment, which was conducted using data collected in 1993 and 2008 and allowed for an analysis of SOC dynamic (chapter 5). The experiment shown in chapter 5 also yielded valuable information for assessing the effect of a climate change scenario on SOC stocks and their spatial distribution in semiarid areas, where low rainfall and high temperatures limit SOC accumulation. Notably, the models of the 1993 and 2008 showed very high reliability, with R^2 above 0.68. These results can help in understanding how internal (i.e. soil texture) and external variables (e.g. climate, land use and land use change, and those measured as remote sensing variable) influence models of SOC dynamic under various management or climate change scenarios. Furthermore, it will be crucial to make sure that the training and deployment datasets mirror have similar characteristics for the covariates.

The methodology implied in chapter 5 also highlighted that the estimated yearly variation of SOC (but see fig. Supplementary material Fig.6 in chapter 5) is on average 0.108 g kg^{-1} y^{-1}. A rough calculation of the SOC stock (using a hypothetic BD reference value for croplands of 1.4 kg dm^{-3}) multiplied by the average SOC sequestration rate found, would result in a gain of 0.48 t ha^{-1} yr^{-1}. This is as double as the theoretical value given by the 4*1000 project and other current published values. However, applying a bulk density of 1.4 kg dm^{-3} appear as excessively high in the area under study. Indeed, from the lass than one thousand data available of measured bulk density in the layer under study, we have found that bulk density in the few non-tilled conditions were 1.47±0.02 kg dm^{-3} in 2008 and 1.08±0.01 kg dm^{-3} in 1993. This partly explain the increase in the SOC. Also, most of sites were tilled and shown a bulk density measured around 0.5-0.7 g dm^{-3}, which low values were indeed due to the tillage itself. This strongly resizes the estimation of the SOC stock to values that are less than a half than those previously computed. Unfortunately, the strong lack of BD measurements and especially the difference in abundance of data per year impaired us to use a PTF as made in the previous work.

Thus, although the chapter in question demonstrated the usefulness of the DSM approach, it has opened some questions on the reliability of the legacy sample design and maybe we can further hypothesize that the different analysis protocols followed during the time make it difficult to

assimilate data into a single frame. In the light of these agronomic considerations, it is useful to remind that the data used in Chapter 5 for the DSM were taken in different places, as it is the nature of a legacy data, with no one single sample overlapping in the two periods. Such lack will be clarified in further work of sampling on coinciding locations.Further work will be needed to understand SOC dynamics under certain conditions, such as under extremely high carbonate content, or under hyper-arid areas. It is useful to scale up to the whole Mediterranean region in order to gain an increased understanding of how machine-learning models are able to map soil properties with their uncertainties and to verify whether similar outcomes from this study are confirmed. This is needed since many Mediterranean environments share similar micro-climatic traits but contrasting land use history.

Nonetheless, the procedure requires checking the reliability of original data and identifying a model able to handle data with given distribution among predictors and space (such as unevenness, scarce resolution and presence of extreme values in both feature space and field data). This is particularly frequent when legacy information are used. Thus, a test of models with contrasting properties was performed (chapter 6) for providing evaluation indices. Future planned works include a quantitative analysis of the international and grey literature and a test of the direct difference in resampled location after 25 years from the first sampling. This study will allow for the analysis of SOC sequestration potential in the Mediterranean Basin. This can be done, along with a sampling design optimization for detecting dynamics of soil properties, including SOC, in cropland (mainly cereals/legumes rotation) using data collected in sites where previous sampling campaigns were conducted. The number of these sites will be defined with a power analysis design and randomly chosen within the examined area. This procedure based on the direct results of SOC differences between two sampling times at the same site can test the efficacy of the indirect results obtained from modelling analysis.

Dissertation acknowledgments

I would like to express my gratitude to the three referees of the dissertation: Dr. Jorge Alvaro-Fuentes, Dr. Daniel Plaza-Bonilla and Prof. Igor Bogunovic for the valuable comments, reflections and ideas useful to improve the dissertation. I am indebted to them and I am extremely grateful to my colleagues with whom I have had stimulating discussions on the topic of thesis and on the methodologies used, Dr. Alessia Perego, Dr. Marcello E. Chiodini, Dr. Aldo Lipani, Dr. Fabio Veronesi, Dr. Maria Fantappiè, Prof. Michael Märker, Dr. Laura Poggio and Dr. Alessandro Gimona. I would like to thank the officers that allowed me to use the soil legacy data, and also for motivating me by providing valuable feedbacks not only linked to scientific validity but also to the practical value of the obtained results: Dr. Maria Gabriella Matranga, Dr. Vito Ferraro and Dr. Fabio Guaitoli (Regional Bureau for Agriculture, Rural Development and Mediterranean Fishery, the Department of Agriculture, Service

7UOS7.03 Geographical Information Systems, Cartography and Broadband Connection in Agriculture, Palermo).

References

Acutis, M., Alfieri, L., Giussani, A., Provolo, G., Di Guardo, A., Colombini, S., Bertoncini, G., Castelnuovo, M., Sali, G., Moschini, M., Sanna, M., Perego, A., Carozzi, M., Chiodini, M.E., Fumagalli, M., 2014. ValorE: An integrated and GIS-based decision support system for livestock manure management in the Lombardy region (northern Italy). Land use policy 41, 149–162. doi:10.1016/j.landusepol.2014.05.007

Adhikari, K., Hartemink, A.E., Minasny, B., Bou Kheir, R., Greve, M.B.M.H., Greve, M.B.M.H., 2014. Digital Mapping of Soil Organic Carbon Contents and Stocks in Denmark. PLoS One 9, e105519. doi:10.1371/journal.pone.0105519

Aguilera, E., Guzmán, G.I., Álvaro-Fuentes, J., Infante-Amate, J., García-Ruiz, R., Carranza-Gallego, G., Soto, D., González de Molina, M., 2018. A historical perspective on soil organic carbon in Mediterranean cropland (Spain, 1900–2008). Sci. Total Environ. 621, 634–648. doi:10.1016/j.scitotenv.2017.11.243

Aitkenhead, M.J., Coull, M.C., 2016. Mapping soil carbon stocks across Scotland using a neural network model. Geoderma 262, 187–198. doi:10.1016/j.geoderma.2015.08.034

Akpa, S.I.C., Odeh, I.O.A., Bishop, T.F.A., Hartemink, A.E., Amapu, I.Y., 2016. Total soil organic carbon and carbon sequestration potential in Nigeria. Geoderma 271, 202–215. doi:10.1016/j.geoderma.2016.02.021

Akpa, S.I.C., Ugbaje, S.U., Bishop, T.F.A., Odeh, I.O.A., 2016. Enhancing pedotransfer functions with environmental data for estimating bulk density and effective cation exchange capacity in a data-sparse situation. Soil Use Manag. 32, 644–658. doi:10.1111/sum.12310

Álvaro-Fuentes, J., López, M. V., Cantero-Martinez, C., Arrúe, J.L., 2008. Tillage Effects on Soil Organic Carbon Fractions in Mediterranean Dryland Agroecosystems. Soil Sci. Soc. Am. J. 72, 541. doi:10.2136/sssaj2007.0164

Álvaro-Fuentes, J., Morell, F.J., Plaza-Bonilla, D., Arrúe, J.L., Cantero-Martínez, C., 2012. Modelling tillage and nitrogen fertilization effects on soil organic carbon dynamics. Soil Tillage Res. 120, 32–39. doi:10.1016/j.still.2012.01.009

An, Y., Yang, L., Zhu, A.-X., Qin, C., Shi, J., 2018. Identification of representative samples from existing samples for digital soil mapping. Geoderma 311, 109–119. doi:10.1016/J.GEODERMA.2017.03.014

Anderson, D.W., Saggar, S., Bettany, J.R., Stewart, J.W.B., 1981. Particle Size Fractions and Their Use in Studies of Soil Organic Matter: I. The Nature and Distribution of Forms of Carbon, Nitrogen, and Sulfur1. Soil Sci. Soc. Am. J. 45, 767. doi:10.2136/sssaj1981.03615995004500040018x

Arrouays, D. et al., 2017. Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. GeoResJ 14, 1–19. doi:10.1016/j.grj.2017.06.001

Arrouays, D., Lagacherie, P., Hartemink, A.E., 2017. Digital soil mapping across the globe. Geoderma Reg. 9, 1–4. doi:https://doi.org/10.1016/j.geodrs.2017.03.002

Badagliacca, G., Ruisi, P., Rees, R.M., Saia, S., 2017. An assessment of factors controlling N2O and CO2 emissions from crop residues using different measurement approaches. Biol. Fertil. Soils. doi:10.1007/s00374-017-1195-z

Bardy, M., Arrouays, D., Jolivet, C., Laroche, B., Le Bas, C., Martin, M., Ratié, C., Richer-De-Forges, A.C., Saby, N., Antoni, V., Bispo, A., Brossard, M., Fort, J.-L., Sauter, J., Gascuel, C., 2018. Understanding Soils for Their More Efficient Management: A National Soil Information System, in: Berthelin, J., Valentin, C., Munch, J.C. (Eds.), Soils as a Key Component of the Critical Zone 1: Functions and Services, First Edition. ISTE Ltd and John Wiley & Sons, Inc., pp. 35–57.

Barré, P., Eglin, T., Christensen, B.T., Ciais, P., Houot, S., Kätterer, T., van Oort, F., Peylin, P., Poulton, P.R., Romanenkov, V., Chenu, C., 2010. Quantifying and isolating stable soil organic carbon using long-term bare fallow experiments. Biogeosciences 7, 3839–3850. doi:10.5194/bg-7-3839-2010

Batjes, N.H., 2016. Harmonized soil property values for broad-scale modelling (WISE30sec) with estimates of global soil carbon stocks. Geoderma 269, 61–68. doi:10.1016/J.GEODERMA.2016.01.034

Batjes, N.H., 2014. Batjes, N. H. 1996. Total carbon and nitrogen in the soils of the world: European Journal of Soil Science, 47, 151-163. Reflections by N.H. Batjes. Eur. J. Soil Sci. 65, 2–3. doi:10.1111/ejss.12115

Batjes, N.H., 2009. Harmonized soil profile data for applications at global and continental scales: updates to the WISE database. Soil Use Manag. 25, 124–127. doi:10.1111/j.1475-2743.2009.00202.x

Batjes, N.H., 1996. Total carbon and nitrogen in the soils of the world. Eur. J. Soil Sci. 47, 151–163. doi:10.1111/j.1365-2389.1996.tb01386.x

Bayliss, H.R., Haddaway, N.R., Eales, J., Frampton, G.K., James, K.L., 2016. Updating and amending systematic reviews and systematic maps in environmental management. Environ. Evid. 5, 20. doi:10.1186/s13750-016-0073-8

Beguin, J., Fuglstad, G.-A., Mansuy, N., Paré, D., 2017. Predicting soil properties in the Canadian boreal forest with limited data: Comparison of spatial and non-spatial statistical approaches. Geoderma 306, 195–205. doi:https://doi.org/10.1016/j.geoderma.2017.06.016

Behrens, T., Zhu, A.-X., Schmidt, K., Scholten, T., 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. Geoderma 155, 175–185. doi:10.1016/j.geoderma.2009.07.010

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. Environ. Model. Softw. 40, 1–20. doi:10.1016/j.envsoft.2012.09.011

Bleuler, M., Farina, R., Francaviglia, R., di Bene, C., Napoli, R., Marchetti, A., 2017. Modelling the impacts of different carbon sources on the soil organic carbon stock and CO 2 emissions in the Foggia province (Southern Italy). Agric. Syst. 157, 258–268. doi:10.1016/j.agsy.2017.07.017

Bocchi, S., 2015. Zolle : storie di tuberi, graminacee e terre coltivate. Cortina.

Bogunović, I., Pereira, P., Brevik, E.C., 2017a. Spatial distribution of soil chemical properties in an organic farm in

Croatia. Sci. Total Environ. 584–585, 535–545. doi:10.1016/j.scitotenv.2017.01.062

Bogunovic, I., Trevisani, S., Seput, M., Juzbasic, D., Durdevic, B., 2017. Short-range and regional spatial variability of soil chemical properties in an agro-ecosystem in eastern Croatia. Catena 154, 50–62. doi:10.1016/j.catena.2017.02.018

Bogunović, I., Trevisani, S., Šeput, M., Juzbašić, D., Đurđević, B., 2017b. Short-range and regional spatial variability of soil chemical properties in an agro-ecosystem in eastern Croatia. Catena (Cremlingen) 156, 1–11.

Boria, E., 2017. Mapping Power, in: Mapping Across Academia. Springer Netherlands, Dordrecht, pp. 223–257. doi:10.1007/978-94-024-1011-2_12

Borrelli, P., Paustian, K., Panagos, P., Jones, A., Schütt, B., Lugato, E., 2016. Effect of Good Agricultural and Environmental Conditions on erosion and soil organic carbon balance: A national case study. Land use policy 50, 408–421. doi:10.1016/j.landusepol.2015.09.033

Borrelli, P., Robinson, D.A., Fleischer, L.R., Lugato, E., Ballabio, C., Alewell, C., Meusburger, K., Modugno, S., Schütt, B., Ferro, V., Bagarello, V., Oost, K. Van, Montanarella, L., Panagos, P., 2017. An assessment of the global impact of 21st century land use change on soil erosion. Nat. Commun. 8, 2013. doi:10.1038/s41467-017-02142-7

Bossard, M., Feranec, J., Otahel, J., 2000. CORINE land cover technical guide - Addendum 2000, Technical Report.

Bou Kheir, R., Greve, M.H., Bøcher, P.K., Greve, M.B., Larsen, R., McCloy, K., 2010. Predictive mapping of soil organic carbon in wet cultivated lands using classification-tree based models: The case study of Denmark. J. Environ. Manage. 91, 1150–1160. doi:10.1016/j.jenvman.2010.01.001

Bradford, M.A., Wieder, W.R., Bonan, G.B., Fierer, N., Raymond, P.A., Crowther, T.W., 2016. Managing uncertainty in soil carbon feedbacks to climate change. Nat. Clim. Chang. 6, 751–758. doi:10.1038/nclimate3071

Bradley, R.I., Milne, R., Bell, J., Lilly, A., Jordan, C., Higgins, A., 2005. A soil carbon and land use database for the United Kingdom. Soil Use Manag. 21, 363–369. doi:10.1079/SUM2005351

Brahim, N., Blavet, D., Gallali, T., Bernoux, M., 2011. Application of structural equation modeling for assessing relationships between organic carbon and soil properties in semiarid Mediterranean region. Iran. J. Environ. Heal. Sci. Eng. 8, 305–320.

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. doi:10.1023/A:1010933404324

Breiman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and regression trees. Wadsworth Books.

Brilli, L., Bechini, L., Bindi, M., Carozzi, M., Cavalli, D., Conant, R., Dorich, C.D., Doro, L., Ehrhardt, F., Farina, R., Ferrise, R., Fitton, N., Francaviglia, R., Grace, P., Iocola, I., Klumpp, K., Léonard, J., Martin, R., Massad, R.S., Recous, S., Seddaiu, G., Sharp, J., Smith, P., Smith, W.N., Soussana, J.-F., Bellocchi, G., 2017. Review and analysis of strengths and weaknesses of agro-ecosystem models for simulating C and N fluxes. Sci. Total Environ. 598, 445–470. doi:10.1016/J.SCITOTENV.2017.03.208

Brus, D.J., 2015. Balanced sampling: A versatile sampling approach for statistical soil surveys. Geoderma 253–254,

111–121. doi:10.1016/j.geoderma.2015.04.009

Bruun, T.B., Elberling, B., de Neergaard, A., Magid, J., 2015. Organic Carbon Dynamics in Different Soil Types After Conversion of Forest to Agriculture. L. Degrad. Dev. 26, 272–283. doi:10.1002/ldr.2205

Bui, E., Henderson, B., Viergever, K., 2009. Using knowledge discovery with data mining from the Australian Soil Resource Information System database to inform soil carbon mapping in Australia. Global Biogeochem. Cycles 23. doi:Artn Gb4033\rDoi 10.1029/2009gb003506

Camilo, D.C., Lombardo, L., Mai, P.M., Dou, J., Huser, R., 2017. Handling high predictor dimensionality in slope-unit-based landslide susceptibility models through LASSO-penalized Generalized Linear Model. Environ. Model. Softw. 97, 145–156. doi:10.1016/J.ENVSOFT.2017.08.003

Campbell, G.., Lilly, A., Corstanje, R., Mayr, T.R., Black, H.I.., 2017. Are existing soils data meeting the needs of stakeholders in Europe? An analysis of practical use from policy to field. Land use policy 69, 211–223. doi:10.1016/J.LANDUSEPOL.2017.09.016

Cannarozzo, M., Noto, L.V., Viola, F., 2006. Spatial distribution of rainfall trends in Sicily (1921–2000). Phys. Chem. Earth, Parts A/B/C 31, 1201–1211. doi:10.1016/j.pce.2006.03.022

Castaldi, F., Castrignanò, A., Casa, R., 2016a. A data fusion and spatial data analysis approach for the estimation of wheat grain nitrogen uptake from satellite data. Int. J. Remote Sens. 37, 4317–4336. doi:10.1080/01431161.2016.1212423

Castaldi, F., Palombo, A., Santini, F., Pascucci, S., Pignatti, S., Casa, R., 2016b. Evaluation of the potential of the current and forthcoming multispectral and hyperspectral imagers to estimate soil texture and organic carbon. Remote Sens. Environ. 179, 54–65. doi:10.1016/j.rse.2016.03.025

Chan, K., Roberts, W., Heenan, D., 1992. Organic carbon and associated soil properties of a red earth after 10 years of rotation under different stubble and tillage practices. Aust. J. Soil Res. 30, 71. doi:10.1071/SR9920071

Chander, G., Haque, M.O., Micijevic, E., Barsi, J.A., 2010. A Procedure for Radiometric Recalibration of Landsat 5 TM Reflective-Band Data. IEEE Trans. Geosci. Remote Sens. 48, 556–574. doi:10.1109/TGRS.2009.2026166

Chartin, C., Stevens, A., Goidts, E., Krüger, I., Carnol, M., van Wesemael, B., 2017. Mapping Soil Organic Carbon stocks and estimating uncertainties at the regional scale following a legacy sampling strategy (Southern Belgium, Wallonia). Geoderma Reg. 9, 73–86. doi:10.1016/j.geodrs.2016.12.006

Chen, S., Martin, M.P., Saby, N.P.A., Walter, C., Angers, D.A., Arrouays, D., 2018a. Fine resolution map of top- and subsoil carbon sequestration potential in France. Sci. Total Environ. 630, 389–400. doi:10.1016/j.scitotenv.2018.02.209

Chen, S., Richer-de-Forges, A.C., Saby, N.P.A., Martin, M.P., Walter, C., Arrouays, D., 2018b. Building a pedotransfer function for soil bulk density on regional dataset and testing its validity over a larger area. Geoderma 312, 52–63. doi:10.1016/j.geoderma.2017.10.009

Chevallier, T., Cournac, L., Bernoux, M., Cardinael, R., Cozzi, T., Girardin, C., Chenu, C., 2017. Soil Inorganic Carbon and Climate Change in drylands? An emerging issue?, in: Global Symposium on Soil Organic Carbon. FAO,

Rome, Italy, pp. 482–485.

Chiti, T., Gardin, L., Perugini, L., Quaratino, R., Vaccari, F.P., Miglietta, F., Valentini, R., 2012. Soil organic carbon stock assessment for the different cropland land uses in Italy. Biol. Fertil. Soils 48, 9–17. doi:10.1007/s00374-011-0599-4

Conant, R.T., Ogle, S.M., Paul, E.A., Paustian, K., 2011. Measuring and monitoring soil organic carbon stocks in agricultural lands for climate mitigation. Front. Ecol. Environ. 9, 169–173. doi:10.1890/090153

Conant, R.T., Paustian, K., Elliott, E.T., 2001. GRASSLAND MANAGEMENT AND CONVERSION INTO GRASSLAND: EFFECTS ON SOIL CARBON. Ecol. Appl. 11, 343–355. doi:10.2307/3060893

Conforti, M., Matteucci, G., Buttafuoco, G., 2017. Monitoring soil organic carbon content using Vis-NIR spectroscopy: A case study in southern Italy. Rend. Online della Soc. Geol. Ital. 42, 38–41. doi:10.3301/ROL.2017.09

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, 1991–2007. doi:10.5194/gmd-8-1991-2015

Costantini, E.A.C., Barbetti, R., Fantappiè, M., L'Abate, G., Lorenzetti, R., Magini, S., 2013. Pedodiversity BT - The Soils of Italy, in: Costantini, E.A.C., Dazzi, C. (Eds.), . Springer Netherlands, Dordrecht, pp. 105–178. doi:10.1007/978-94-007-5642-7_6

Costantini, E.A.C., Fattappié, M., L'Abate, G., 2013. Climate and Pedoclimate of Italy, in: Costantini, E.A.C., Dazzi, C. (Eds.), The Soils of Italy, World Soils Book Series. Springer Netherlands, Dordrecht, pp. 19–37. doi:10.1007/978-94-007-5642-7

Costantini, E.A.C., L'Abate, G., 2016. Beyond the concept of dominant soil: Preserving pedodiversity in upscaling soil maps. Geoderma 271, 243–253. doi:10.1016/j.geoderma.2015.11.024

Cressie, N.A.C., 1993. Statistics for Spatial Data, Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA. doi:10.1002/9781119115151

Crowther, T.W., Todd-Brown, K.E.O., Rowe, C.W., Wieder, W.R., Carey, J.C., Machmuller, M.B., Snoek, B.L., Fang, S., Zhou, G., Allison, S.D., Blair, J.M., Bridgham, S.D., Burton, A.J., Carrillo, Y., Reich, P.B., Clark, J.S., Classen, A.T., Dijkstra, F.A., Elberling, B., Emmett, B.A., Estiarte, M., Frey, S.D., Guo, J., Harte, J., Jiang, L., Johnson, B.R., Kröel-Dulay, G., Larsen, K.S., Laudon, H., Lavallee, J.M., Luo, Y., Lupascu, M., Ma, L.N., Marhan, S., Michelsen, A., Mohan, J., Niu, S., Pendall, E., Peñuelas, J., Pfeifer-Meister, L., Poll, C., Reinsch, S., Reynolds, L.L., Schmidt, I.K., Sistla, S., Sokol, N.W., Templer, P.H., Treseder, K.K., Welker, J.M., Bradford, M.A., 2016. Quantifying global soil carbon losses in response to warming. Nature 540, 104–108. doi:10.1038/nature20150

Dai, F., Zhou, Q., Lv, Z., Wang, X., Liu, G., 2014. Spatial prediction of soil organic matter content integrating artificial neural network and ordinary kriging in Tibetan Plateau. Ecol. Indic. 45, 184–194. doi:10.1016/J.ECOLIND.2014.04.003

Darwish, T., Fadel, A., 2017. Mapping of soil organic carbon stock in the Arab countries to mitigate land degradation.

Arab. J. Geosci. 10, 486–490. doi:10.1007/s12517-017-3267-7

Davidson, E.A., Janssens, I.A., 2006. Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. Nature 440, 165–173. doi:10.1038/nature04514

de Brogniez, D., Ballabio, C., Stevens, A., Jones, R.J.A., Montanarella, L., van Wesemael, B., 2015. A map of the topsoil organic carbon content of Europe generated by a generalized additive model. Eur. J. Soil Sci. 66, 121–134. doi:10.1111/ejss.12193

de Gruijter, J.J., McBratney, A.B., Minasny, B., Wheeler, I., Malone, B.P., Stockmann, U., 2016. Farm-scale soil carbon auditing. Geoderma 265, 120–130. doi:10.1016/j.geoderma.2015.11.010

Di Bene, C., Marchetti, A., Francaviglia, R., Farina, R., 2016. Soil organic carbon dynamics in typical durum wheat-based crop rotations of southern Italy. Ital. J. Agron. 11, 209–216. doi:10.4081/ija.2016.763

Dobos, E., Bialkó, T., Micheli, E., Kobza, J., 2010. Legacy Soil Data Harmonization and Database Development BT - Digital Soil Mapping: Bridging Research, Environmental Application, and Operation, in: Boettinger, J.L., Howell, D.W., Moore, A.C., Hartemink, A.E., Kienast-Brown, S. (Eds.), . Springer Netherlands, Dordrecht, pp. 309–323. doi:10.1007/978-90-481-8863-5_25

Doetterl, S., Stevens, A., Six, J., Merckx, R., Van Oost, K., Casanova Pinto, M., Casanova-Katny, A., Muñoz, C., Boudin, M., Zagal Venegas, E., Boeckx, P., 2015a. Soil carbon storage controlled by interactions between geochemistry and climate. Nat. Geosci. 8, 780–783. doi:10.1038/ngeo2516

Doetterl, S., Stevens, A., Six, J., Merckx, R., Van Oost, K., Casanova Pinto, M., Casanova-Katny, A., Muñoz, C., Boudin, M., Zagal Venegas, E., Boeckx, P., 2015b. Soil carbon storage controlled by interactions between geochemistry and climate. Nat. Geosci. 8, 780–783. doi:10.1038/ngeo2516

Don, A., Osborne, B., Hastings, A., Skiba, U., Carter, M.S., Drewer, J., Flessa, H., Freibauer, A., Hyvönen, N., Jones, M.B., Lanigan, G.J., Mander, Ü., Monti, A., Djomo, S.N., Valentine, J., Walter, K., Zegada-Lizarazu, W., Zenone, T., 2012. Land-use change to bioenergy production in Europe: Implications for the greenhouse gas balance and soil carbon. GCB Bioenergy 4, 372–391. doi:10.1111/j.1757-1707.2011.01116.x

Dong, W., Song, A., Liu, X., Yu, B., Wang, B., Lu, Y., Li, Y., Yin, H., Li, J., Fan, F., 2018. Warming deferentially altered multidimensional soil legacy induced by past land use history. Sci. Rep. 8, 1546. doi:10.1038/s41598-018-19912-y

Dono, G., Cortignani, R., Dell'Unto, D., Deligios, P., Doro, L., Lacetera, N., Mula, L., Pasqui, M., Quaresima, S., Vitali, A., Roggero, P.P., 2016. Winners and losers from climate change in agriculture: Insights from a case study in the Mediterranean basin. Agric. Syst. 147, 65–75. doi:10.1016/j.agsy.2016.05.013

Egli, M., Alioth, L., Mirabella, A., Raimondi, S., Nater, M., Verel, R., 2007. Effect of climate and vegetation on soil organic carbon, humus fractions, allophanes, imogolite, kaolinite, and oxyhydroxides in volcanic soils of Etna (Sicily). Soil Sci. 172, 673–691. doi:10.1097/ss.0b013e31809eda23

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol. 77, 802–813. doi:10.1111/j.1365-2656.2008.01390.x

Evrendilek, F., Celik, I., Kilic, S., 2004. Changes in soil organic carbon and other physical soil properties along adjacent Mediterranean forest, grassland, and cropland ecosystems in Turkey. J. Arid Environ. 59, 743–752. doi:10.1016/j.jaridenv.2004.03.002

Fantappiè, M., Bocci, M., Paolanti, M., Perciabosco, M., Antinoro, C., Rivieccio, R., Costantini, E.A.C., 2011a. Realizzazione della carta digitale dei suoli della Sicilia utilizzando il rilevamento GIS-oriented e un modello CLORPT, in: Dazzi, C. (Ed.), La Percezione Del Suolo - Atti Del Workshop. Collana: Scienza Del Suolo. Le Penseur, pp. 139–142.

Fantappiè, M., L'Abate, G., Costantini, E.A.C., 2011b. The influence of climate change on the soil organic carbon content in Italy from 1961 to 2008. Geomorphology 135, 343–352. doi:10.1016/j.geomorph.2011.02.006

Fantappiè, M., L'Abate, G., Costantini, E.A.C., 2010. Factors influencing soil organic carbon stock variations in Italy during the last three decades, in: Zdruli, P. (Ed.), Land Degradation and Desertification: Assessment, Mitigation and Remediation. pp. 435–465. doi:10.1007/978-90-481-8657-0_34

Fantappiè, M., Priori, S., Costantini, E., 2016. Physiography of the Sicilian region (1:250,000 scale). J. Maps 12, 111–122. doi:10.1080/17445647.2014.984785

Fantappiè, M., Priori, S., Costantini, E.A.C., 2015. Soil erosion risk, Sicilian Region (1:250,000 scale). J. Maps 11, 323–341. doi:10.1080/17445647.2014.956349

Farina, R., Marchetti, A., Francaviglia, R., Napoli, R., Bene, C. Di, 2016. Modeling regional soil C stocks and CO2 emissions under Mediterranean cropping systems and soil types. Agric. Ecosyst. Environ. doi:10.1016/j.agee.2016.08.015

Farr, T.G., Rosen, P.A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., Alsdorf, D., 2007. The Shuttle Radar Topography Mission. Rev. Geophys. 45, n/a--n/a. doi:10.1029/2005RG000183

Fernández-Getino, A.P., Duarte, A.C., 2015. Soil management guidelines in Spain and Portugal related to EU Soil Protection Strategy based on analysis of soil databases. Catena 126, 146–154. doi:10.1016/j.catena.2014.11.003

Ferrara, R.M., Trevisiol, P., Acutis, M., Rana, G., Richter, G.M., Baggaley, N., 2009. Topographic impacts on wheat yields under climate change: two contrasted case studies in Europe. Theor. Appl. Climatol. 99, 53–65. doi:10.1007/s00704-009-0126-9

Ferro, V., Giordano, G., Iovino, M., 1991. Isoerosivity and erosion risk map for Sicily. Hydrol. Sci. J. 36, 549–564. doi:10.1080/02626669109492543

Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37, 4302–4315. doi:10.1002/joc.5086

Filippi, P., Minasny, B., Cattle, S.R., Bishop, T.F.A., 2016. Chapter Four – Monitoring and Modeling Soil Change: The Influence of Human Activity and Climatic Shifts on Aspects of Soil Spatiotemporally, in: Advances in Agronomy. pp. 153–214. doi:10.1016/bs.agron.2016.06.001

Florinsky, I. V., 2012. The Dokuchaev hypothesis as a basis for predictive digital soil mapping (on the 125th

anniversary of its publication). Eurasian Soil Sci. 45, 445–451. doi:10.1134/S1064229312040047

Forkuor, G., Hounkpatin, O.K.L., Welp, G., Thiel, M., Zhu, A.-X., Scholten, T., Koch, B., Shepherd, K., 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, e0170478. doi:10.1371/journal.pone.0170478

Francaviglia, R., Renzi, G., Doro, L., Parras-Alcántara, L., Lozano-García, B., Ledda, L., 2017a. Soil sampling approaches in Mediterranean agro-ecosystems. Influence on soil organic carbon stocks. Catena 158, 113–120. doi:10.1016/j.catena.2017.06.014

Francaviglia, R., Renzi, G., Ledda, L., Benedetti, A., 2017b. Organic carbon pools and soil biological fertility are affected by land use intensity in Mediterranean ecosystems of Sardinia, Italy. Sci. Total Environ. 599–600, 789–796. doi:10.1016/j.scitotenv.2017.05.021

Freibauer, A., Rounsevell, M.D.A., Smith, P., Verhagen, J., 2004. Carbon sequestration in the agricultural soils of Europe. Geoderma 122, 1–23. doi:10.1016/j.geoderma.2004.01.021

Freund, Y., Schapire, R.E., 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. J. Comput. Syst. Sci. 55, 119–139. doi:10.1006/jcss.1997.1504

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: A statistical view of boosting. Ann. Stat. 28, 337–407. doi:10.1214/aos/1016218223

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38, 367–378. doi:10.1016/S0167-9473(01)00065-2

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.

Galati, A., Crescimanno, M., Gristina, L., Keesstra, S., Novara, A., 2016. Actual provision as an alternative criterion to improve the efficiency of payments for ecosystem services for C sequestration in semiarid vineyards. Agric. Syst. 144, 58–64. doi:10.1016/j.agsy.2016.02.004

Gargouri, K., Rigane, H., Arous, I., Touil, F., 2013. Evolution of soil organic carbon in an olive orchard under arid climate. Sci. Hortic. (Amsterdam). 152, 102–108. doi:10.1016/j.scienta.2012.11.025

Gasch, C.K., Gräler, B., Meyer, H., Magney, T.S., Brown, D.J., 2015. Spatio-temporal interpolation of soil water, temperature, and electrical conductivity in 3D + T: The Cook Agronomy Farm data set. Spat. Stat. 14, 70–90. doi:10.1016/j.spasta.2015.04.001

Giovino, A., Scibetta, S., Saia, S., Guarino, C., 2014. Genetic and morphologic diversity of European fan palm (Chamaerops humilis L.) populations from different environments from Sicily. Bot. J. Linn. Soc. 176, 66–81. doi:10.1111/boj.12195

Gomez, C., Viscarra Rossel, R.A., McBratney, A.B., 2008. Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: An Australian case study. Geoderma 146, 403–411. doi:10.1016/j.geoderma.2008.06.011

Gosling, P., van der Gast, C., Bending, G.D., 2017. Converting highly productive arable cropland in Europe to grassland: –a poor candidate for carbon sequestration. Sci. Rep. 7, 10493. doi:10.1038/s41598-017-11083-6

Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island - Digital soil mapping using Random Forests analysis. Geoderma 146, 102–113. doi:10.1016/j.geoderma.2008.05.008

Grinand, C., Maire, G. Le, Vieilledent, G., Razakamanarivo, H., Razafimbelo, T., Bernoux, M., 2017. Estimating temporal changes in soil carbon stocks at ecoregional scale in Madagascar using remote-sensing. Int. J. Appl. Earth Obs. Geoinf. 54, 1–14. doi:10.1016/j.jag.2016.09.002

Gregorich E.G, Carter, M.R., Angers, D.A., Monreal, E.E., 1994. Toward a minimum data set to assess soil organic matter quality in agr soils. Can.J. Soil. Sci.

Grunwald, S., 2009. Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma 152, 195–207. doi:10.1016/J.GEODERMA.2009.06.003

Grunwald, S., Thompson, J.A., Boettinger, J.L., 2011. Digital Soil Mapping and Modeling at Continental Scales: Finding Solutions for Global Issues. Soil Sci. Soc. Am. J. 75, 1201. doi:10.2136/sssaj2011.0025

Guevara, M., Olmedo, G.F., Stell, E., Yigini, Y., Aguilar Duarte, Y., Arellano Hernández, C., Arévalo, G.E., Arroyo-Cruz, C.E., Bolivar, A., Bunning, S., Bustamante Cañas, N., Cruz-Gaistardo, C.O., Davila, F., Dell Acqua, M., Encina, A., Figueredo Tacona, H., Fontes, F., Hernández Herrera, J.A., Ibelles Navarro, A.R., Loayza, V., Manueles, A.M., Mendoza Jara, F., Olivera, C., Osorio Hermosilla, R., Pereira, G., Prieto, P., Ramos, I.A., Rey Brina, J.C., Rivera, R., Rodríguez-Rodríguez, J., Roopnarine, R., Rosales Ibarra, A., Rosales Riveiro, K.A., Schulz, G.A., Spence, A., Vasques, G.M., Vargas, R.R., Vargas, R., 2018. No silver bullet for digital soil mapping: country-specific soil organic carbon estimates across Latin America. SOIL 4, 173–193. doi:10.5194/soil-4-173-2018

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene Selection for Cancer Classification using Support Vector Machines. Mach. Learn. 46, 389–422. doi:10.1023/A:1012487302797

Guyot, G., Gu, X.-F., 1994. Effect of radiometric corrections on NDVI-determined from SPOT-HRV and Landsat-TM data. Remote Sens. Environ. 49, 169–180. doi:10.1016/0034-4257(94)90012-4

Haddaway, N.R., 2014. Maximizing legacy and impact of primary research: a call for better reporting of results. Ambio 43. doi:10.1007/s13280-014-0535-6

Haddaway, N.R., Hedlund, K., Jackson, L.E., Kätterer, T., Lugato, E., Thomsen, I.K., 2016. How does tillage intensity affect soil organic carbon? A systematic review protocol. Env. Evid 5. doi:10.1186/s13750-016-0052-0

Haddaway, N.R., Styles, D., Pullin, A.S., 2014. Environmental impacts of farm land abandonment in high altitude/mountain regions: a systematic map. Env. Evid. 3. doi:10.1186/2047-2382-3-17

Hallett, S.H., Sakrabani, R., Keay, C.A., Hannam, J.A., 2017. Developments in land information systems: examples demonstrating land resource management capabilities and options. Soil Use Manag. 33, 514–529. doi:10.1111/sum.12380

Hashimoto, S., Nanko, K., Ťupek, B., Lehtonen, A., 2016. Data-mining analysis of factors affecting the global distribution of soil carbon in observational databases and Earth system models. Geosci. Model Dev. Discuss. 1–22. doi:10.5194/gmd-2016-138

Hastie, T., Tibshirani, R., Friedman, J., 2009. Random Forests. pp. 587–604. doi:10.1007/978-0-387-84858-7_15

Heinze, S., Ludwig, B., Piepho, H.-P., Mikutta, R., Don, A., Wordell-Dietrich, P., Helfrich, M., Hertel, D., Leuschner, C., Kirfel, K., Kandeler, E., Preusser, S., Guggenberger, G., Leinemann, T., Marschner, B., 2018. Factors controlling the variability of organic matter in the top- and subsoil of a sandy Dystric Cambisol under beech forest. Geoderma 311, 37–44. doi:10.1016/j.geoderma.2017.09.028

Henderson, B.L., Bui, E.N., Moran, C.J., Simon, D.A.P., 2005. Australia-wide predictions of soil properties using decision trees. Geoderma 124, 383–398. doi:10.1016/j.geoderma.2004.06.007

Hendriks, C.M.J., Stoorvogel, J.J., Claessens, L., 2016. Exploring the challenges with soil data in regional land use analysis. Agric. Syst. 144, 9–21. doi:10.1016/j.agsy.2016.01.007

Hengl, T., de Jesus, J.M., MacMillan, R.A.R., Batjes, N.N.H., Heuvelink, G.B.M.G., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Gonzalez, M.R., 2014. SoilGrids1km — Global Soil Information Based on Automated Mapping. PLoS One 9, e105992. doi:10.1371/journal.pone.0105992

Hengl, T., Heuvelink, G.B.M., Stein, A., 2003. Comparison of kriging with external drift and regression-kriging *.

Hengl, T., Mendes de Jesus, J., Heuvelink, G.B.M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS One 12, e0169748. doi:10.1371/journal.pone.0169748

Henry, M., Valentini, R., Bernoux, M., 2009. Soil carbon stocks in ecoregions of Africa. Biogeosciences Discuss. 6, 797–823. doi:10.5194/bgd-6-797-2009

Hiederer, R., Köchy, M., 2012. Global soil organic carbon estimates and the harmonized world soil database, EUR Scientific and Technical Research series. doi:10.2788/13267

Hiemstra, P.H., Pebesma, E.J., Twenhöfel, C.J.W., Heuvelink, G.B.M., n.d. Real-time automatic interpolation of ambient gamma dose rates from the Dutch Radioactivity Monitoring Network.

Hijbeek, R., Cormont, A., Hazeu, G., Bechini, L., Zavattaro, L., Janssen, B., Werner, M., Schlatter, N., Guzmán, G., Bijttebier, J., Pronk, A.A., van Eupen, M., van Ittersum, M.K., 2017. Do farmers perceive a deficiency of soil organic matter? A European and farm level analysis. Ecol. Indic. 83, 390–403. doi:10.1016/J.ECOLIND.2017.08.023

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25, 1965–1978. doi:10.1002/joc.1276

Huang, J., Buchanan, S.M., Bishop, T.F.A., Triantafilis, J., 2017. terraGIS – a web GIS for delivery of digital soil maps in cotton-growing areas of Australia. Soil Use Manag. 33, 568–582. doi:10.1111/sum.12383

Huang, J., Minasny, B., McBratney, A.B., Padarian, J., Triantafilis, J., 2018. The location- and scale- specific correlation between temperature and soil carbon sequestration across the globe. Sci. Total Environ. 615, 540–548. doi:10.1016/J.SCITOTENV.2017.09.136

Hudson, B.D., 1994. Soil organic matter and available water capacity. J. Soil Water Conserv. 49, 189–194. doi:10.1081/E-ESS-120018496

Ingram, J., Mills, J., Dibari, C., Ferrise, R., Ghaley, B.B., Hansen, J.G., Iglesias, A., Karaczun, Z., McVittie, A., Merante, P., Molnar, A., Sánchez, B., 2016. Communicating soil carbon science to farmers: Incorporating credibility, salience and legitimacy. J. Rural Stud. 48, 115–128. doi:10.1016/J.JRURSTUD.2016.10.005

IUSS Working Group WRB, 2014. World reference base for soil resources 2014. International soil classification system for naming soils and creating legends for soil maps, World Soil Resources Reports No. 106. doi:10.1017/S0014479706394902

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. Statistical Learning. pp. 15–57. doi:10.1007/978-1-4614-7138-7_2

James, K.L., Randall, N.P., Haddaway, N.R., 2016. A methodology for systematic mapping in environmental sciences. Environ. Evid. 5. doi:10.1186/s13750-016-0059-6

Jandl, R., Rodeghiero, M., Martinez, C., Cotrufo, M.F., Bampa, F., van Wesemael, B., Harrison, R.B., Guerrini, I.A., Richter, D. deB, Rustad, L., Lorenz, K., Chabbi, A., Miglietta, F., 2014. Current status, uncertainty and future needs in soil organic carbon monitoring. Sci. Total Environ. 468–469, 376–383. doi:10.1016/J.SCITOTENV.2013.08.026

Kämpf, I., Hölzel, N., Störrle, M., Broll, G., Kiehl, K., 2016. Potential of temperate agricultural soils for carbon sequestration: A meta-analysis of land-use effects. Sci. Total Environ. 566, 428–435. doi:10.1016/j.scitotenv.2016.05.067

Kanevski, M., Pozdnoukhov, A., Timonin, V., 2009. Machine learning for spatial environmental data : theory, applications and software. EPFL Press.

Karunaratne, S.B., Bishop, T.F.A., Odeh, I.O.A., Baldock, J.A., Marchant, B.P., 2014. Estimating change in soil organic carbon using legacy data as the baseline: issues, approaches and lessons to learn. Soil Res. 52, 349. doi:10.1071/SR13081

Kempen, B., Brus, D.J., de Vries, F., 2015. Operationalizing digital soil mapping for nationwide updating of the 1:50,000 soil map of the Netherlands. Geoderma 241–242, 313–329. doi:10.1016/j.geoderma.2014.11.030

Kerry, R., Oliver, M.A., 2007. Comparing sampling needs for variograms of soil properties computed by the method of moments and residual maximum likelihood. Geoderma 140, 383–396. doi:10.1016/J.GEODERMA.2007.04.019

Kirschbaum, M.U.F., 1995. The temperature dependence of soil organic matter decomposition, and the effect of global warming on soil organic C storage. Soil Biol. Biochem. 27, 753–760. doi:10.1016/0038-0717(94)00242-S

Köchy, M., Don, A., van der Molen, M.K., Freibauer, A., 2015a. Global distribution of soil organic carbon – Part 2: Certainty of changes related to land use and climate. SOIL 1, 367–380. doi:10.5194/soil-1-367-2015

Köchy, M., Hiederer, R., Freibauer, A., 2015b. Global distribution of soil organic carbon – Part 1: Masses and frequency distributions of SOC stocks for the tropics, permafrost regions, wetlands, and the world. Soil 1, 351–365. doi:10.5194/soil-1-351-2015

Krol, B.G.C.M., 2008. Towards a Data Quality Management Framework for Digital Soil Mapping with Limited Data BT - Digital Soil Mapping with Limited Data, in: Hartemink, A.E., McBratney, A., Mendonça-Santos, M. de L. (Eds.), . Springer Netherlands, Dordrecht, pp. 137–149. doi:10.1007/978-1-4020-8592-5_11

Kurganova, I., Lopes de Gerenyu, V., Six, J., Kuzyakov, Y., 2014. Carbon cost of collective farming collapse in Russia. Glob. Chang. Biol. 20, 938–947. doi:10.1111/gcb.12379

Lacoste, M., Minasny, B., McBratney, A., Michot, D., Viaud, V., Walter, C., 2014. High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. Geoderma 213, 296–311. doi:10.1016/J.GEODERMA.2013.07.002

Lagacherie, P., Álvaro-Fuentes, J., Annabi, M., Bernoux, M., Bouarfa, S., Douaoui, A., Grünberger, O., Hammani, A., Montanarella, L., Mrabet, R., Sabir, M., Raclot, D., 2018. Managing Mediterranean soil resources under global change: expected trends and mitigation strategies. Reg. Environ. Chang. 18, 663–675. doi:10.1007/s10113-017-1239-9

Lal, R., 2007. Soil carbon stocks under present and future climate with specific reference to European ecoregions. Nutr. Cycl. Agroecosystems 81, 113–127. doi:10.1007/s10705-007-9147-x

Lal, R., 2004. Soil carbon sequestration impacts on global climate change and food security. Science 304, 1623–1627. doi:10.1126/science.1097396

Laurent, A.G., 1963. The Lognormal Distribution and the Translation Method: Description and Estimation Problems. J. Am. Stat. Assoc. 58, 231–235. doi:10.1080/01621459.1963.10500844

Lee, J., Hopmans, J.W., Rolston, D.E., Baer, S.G., Six, J., 2009. Determining soil carbon stock changes: Simple bulk density corrections fail. Agric. Ecosyst. Environ. 134, 251–256. doi:10.1016/j.agee.2009.07.006

Legendre, P., Legendre, L., 1998. Numerical Ecology - Second English Edition, Developments in Environmental Modelling. doi:10.1017/CBO9781107415324.004

Leip, a., Marchi, G., Koeble, R., Kempen, M., Britz, W., Li, C., 2008. Linking an economic model for European agriculture with a mechanistic model to estimate nitrogen and carbon losses from arable soils in Europe. Biogeosciences 5, 73–94. doi:10.5194/bg-5-73-2008

Li, Y., Shi, S., Waqas, M.A., Zhou, X., Li, J., Wan, Y., Qin, X., Gao, Q., Liu, S., Wilkes, A., 2018. Long-term (≥20 years) application of fertilizers and straw return enhances soil carbon storage: a meta-analysis. Mitig. Adapt. Strateg. Glob. Chang. 23, 603–619. doi:10.1007/s11027-017-9751-2

Lin, L.I.-K., 1989. A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics 45, 255. doi:10.2307/2532051

Liu, S., Wei, Y., Post, W.M., Cook, R.B., Schaefer, K., Thornton, M.M., 2013. The Unified North American Soil Map and its implication on the soil organic carbon stock in North America. Biogeosciences 10, 2915–2930.

doi:10.5194/bg-10-2915-2013

Lobry de Bruyn, L., Jenkins, A., Samson-Liebig, S., 2017. Lessons Learnt: Sharing Soil Knowledge to Improve Land Management and Sustainable Soil Use. Soil Sci. Soc. Am. J. 81, 427. doi:10.2136/sssaj2016.12.0403

Lombardo, L., Cama, M., Conoscenti, C., Märker, M., Rotigliano, E., 2015. Binary logistic regression versus stochastic gradient boosted decision trees in assessing landslide susceptibility for multiple-occurring landslide events: application to the 2009 storm event in Messina (Sicily, southern Italy). Nat. Hazards 79, 1621–1648. doi:10.1007/s11069-015-1915-3

Lombardo, L., Saia, S., Schillaci, C., Mai, P., Huser, R., 2018. Modeling soil organic carbon with Quantile Regression: Dissecting predictors' effects on carbon stocks. Geoderma. Geoderma 318, in press. doi:10.1016/j.geoderma.2017.12.011

Lugato, E., Bampa, F., Panagos, P., Montanarella, L., Jones, A., 2014a. Potential carbon sequestration of European arable soils estimated by modelling a comprehensive set of management practices. Glob. Chang. Biol. 20, 3557–3567. doi:10.1111/gcb.12551

Lugato, E., Panagos, P., Bampa, F., Jones, A., Montanarella, L., 2014b. A new baseline of organic carbon stock in European agricultural soils using a modelling approach. Glob. Chang. Biol. 20, 313–326. doi:10.1111/gcb.12292

Lugato, E., Paustian, K., Panagos, P., Jones, A., Borrelli, P., 2016. Quantifying the erosion effect on current carbon budget of European agricultural soils at high spatial resolution. Glob. Chang. Biol. 22, 1976–1984. doi:10.1111/gcb.13198

Luo, Z., Wang, E., Zheng, H., Baldock, J.A., Sun, O.J., Shao, Q., 2015. Convergent modelling of past soil organic carbon stocks but divergent projections. Biogeosciences 12, 4373–4383. doi:10.5194/bg-12-4373-2015

Ma, L., Yuan, S., Guo, C., Wang, R., 2014. Carbon and nitrogen dynamics of native Leymus chinensis grasslands along a 1000 km longitudinal precipitation gradient in northeastern China. Biogeosciences 11, 7097–7106. doi:10.5194/bg-11-7097-2014

Macrae, R.J., Mehuys, G.R., 1985. The effect of green manuring on the physical properties of temperate-area soils. Adv. soil Sci. Vol. 3.

Maia, S.M.F., Ogle, S.M., Cerri, C.E.P., Cerri, C.C., 2010. Soil organic carbon stock change due to land use activity along the agricultural frontier of the southwestern Amazon, Brazil, between 1970 and 2002. Glob. Chang. Biol. 16, 2775–2788. doi:10.1111/j.1365-2486.2009.02105.x

Malone, B.P., McBratney, A.B., Minasny, B., Laslett, G.M., 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma 154, 138–152. doi:10.1016/j.geoderma.2009.10.007

Malone, B.P., Styc, Q., Minasny, B., McBratney, A.B., 2017. Digital soil mapping of soil carbon at the farm scale: A spatial downscaling approach in consideration of measured and uncertain data. Geoderma 290, 91–99. doi:10.1016/j.geoderma.2016.12.008

Martin, M.P., Orton, T.G., Lacarce, E., Meersmans, J., Saby, N.P.A., Paroissien, J.B., Jolivet, C., Boulonne, L., Arrouays, D., 2014. Evaluation of modelling approaches for predicting the spatial distribution of soil organic

carbon stocks at the national scale. Geoderma 223–225, 97–107. doi:10.1016/j.geoderma.2014.01.005

Martin, M.P., Wattenbach, M., Smith, P., Meersmans, J., Jolivet, C., Boulonne, L., Arrouays, D., 2011. Spatial distribution of soil organic carbon stocks in France. Biogeosciences 8, 1053–1065. doi:10.5194/bg-8-1053-2011

Martin, M.P., Wattenbach, M., Smith, P., Meersmans, J., Jolivet, C., Boulonne, L., Arrouays, D., 2010. Spatial distribution of soil organic carbon stocks in France. Biogeosciences 8, 1053–1065. doi:10.5194/bg-8-1053-2011

Mateu-Andrés, I., Aguilella, A., Boisset, F., Currás, R., Guara, M., Laguna, E., Marzo, A., Puche, M.F., Pedrola, J., 2013. Geographical patterns of genetic variation in rosemary (Rosmarinus officinalis) in the Mediterranean basin. Bot. J. Linn. Soc. 171, 700–712. doi:10.1111/boj.12017

McBratney, A., Mendonça Santos, M.L., Minasny B., 2003. On digital soil mapping. Geoderma 117, 3–52. doi:10.1016/S0016-7061(03)00223-4

McBratney, A.B., Odeh, I.O.A., Bishop, T.F.A., Dunbar, M.S., Shatar, T.M., 2000. An overview of pedometric techniques for use in soil survey. Geoderma 97, 293–327. doi:10.1016/S0016-7061(00)00043-4

McKay, M.D., Beckman, R.J., Conover, W.J., 1979. Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. Technometrics 21, 239–245. doi:10.1080/00401706.1979.10489755

Meersmans, J., De Ridder, F., Canters, F., De Baets, S., Van Molle, M., DERIDDER, F., Canters, F., DEBAETS, S., VANMOLLE, M., 2008. A multiple regression approach to assess the spatial distribution of Soil Organic Carbon (SOC) at the regional scale (Flanders, Belgium). Geoderma 143, 1–13. doi:10.1016/j.geoderma.2007.08.025

Meersmans, J., Van Wesemael, B., Van Molle, M., 2009. Determining soil organic carbon for agricultural soils: a comparison between the Walkley & Black and the dry combustion methods (north Belgium). Soil Use Manag. 25, 346–353. doi:10.1111/j.1475-2743.2009.00242.x

Miller, B.A., Koszinski, S., Hierold, W., Rogasik, H., Schröder, B., Van Oost, K., Wehrhan, M., Sommer, M., 2016. Towards mapping soil carbon landscapes: Issues of sampling scale and transferability. Soil Tillage Res. 156, 194–208. doi:10.1016/j.still.2015.07.004

Miller, B.A., Koszinski, S., Wehrhan, M., Sommer, M., 2015. Comparison of spatial association approaches for landscape mapping of soil organic carbon stocks. SOIL 1, 217–233. doi:10.5194/soil-1-217-2015

Miller, B.A., Koszinski, S., Wehrhan, M., Sommer, M., 2015. Impact of multi-scale predictor selection for modeling soil properties. Geoderma 239, 97–106. doi:10.1016/j.geoderma.2014.09.018

Milne, E., Banwart, S.A., Noellemeyer, E., Abson, D.J., Ballabio, C., Bampa, F., Bationo, A., Batjes, N.H., Bernoux, M., Bhattacharyya, T., Black, H., Buschiazzo, D.E., Cai, Z., Cerri, C.E., Cheng, K., Compagnone, C., Conant, R., Coutinho, H.L.C., de Brogniez, D., Balieiro, F. de C., Duffy, C., Feller, C., Fidalgo, E.C.C., da Silva, C.F., Funk, R., Gaudig, G., Gicheru, P.T., Goldhaber, M., Gottschalk, P., Goulet, F., Goverse, T., Grathwohl, P., Joosten, H., Kamoni, P.T., Kihara, J., Krawczynski, R., La Scala, N., Lemanceau, P., Li, L., Li, Z., Lugato, E., Maron, P.-A., Martius, C., Melillo, J., Montanarella, L., Nikolaidis, N., Nziguheba, G., Pan, G., Pascual, U., Paustian, K., Piñeiro, G., Powlson, D., Quiroga, A., Richter, D., Sigwalt, A., Six, J., Smith, J., Smith, P., Stocking, M.,

Tanneberger, F., Termansen, M., van Noordwijk, M., van Wesemael, B., Vargas, R., Victoria, R.L., Waswa, B., Werner, D., Wichmann, S., Wichtmann, W., Zhang, X., Zhao, Y., Zheng, J., Zheng, J., 2015. Soil carbon, multiple benefits. Environ. Dev. 13, 33–38. doi:10.1016/J.ENVDEV.2014.11.005

Minasny, B., Malone, B.P., McBratney, A.B., Angers, D.A., Arrouays, D., Chambers, A., Chaplot, V., Chen, Z.-S., Cheng, K., Das, B.S., Field, D.J., Gimona, A., Hedley, C.B., Hong, S.Y., Mandal, B., Marchant, B.P., Martin, M., McConkey, B.G., Mulder, V.L., O'Rourke, S., Richer-de-Forges, A.C., Odeh, I., Padarian, J., Paustian, K., Pan, G., Poggio, L., Savin, I., Stolbovoy, V., Stockmann, U., Sulaeman, Y., Tsui, C.-C., Vågen, T.-G., van Wesemael, B., Winowiecki, L., 2017. Soil carbon 4 per mille. Geoderma 292, 59–86. doi:10.1016/j.geoderma.2017.01.002

Minasny, B., McBratney, A.B., Malone, B.P., Wheeler, I., 2013. Digital Mapping of Soil Carbon. Adv. Agron. 118, 1–47. doi:10.1016/B978-0-12-405942-9.00001-3

Mishra, U., Lal, R., Slater, B., Calhoun, F., Liu, D.S., Van Meirvenne, M., 2009., Mishra, U., Lal, R., Slater, B., Calhoun, F., Liu, D., Van Meirvenne, M., 2009. Predicting soil organic carbon stock using profile depth distribution functions and ordinary kriging. Soil Sci. Soc. Am. J. 73 (2), 614–621 73, 614. doi:10.2136/sssaj2007.0410

Mondal, A., Khare, D., Kundu, S., 2016. Impact assessment of climate change on future soil erosion and SOC loss. Nat. Hazards 82, 1515–1539. doi:10.1007/s11069-016-2255-7

Mongeon, P., Paul-Hus, A., 2016. The journal coverage of Web of Science and Scopus: a comparative analysis. Scientometrics 106, 213–228. doi:10.1007/s11192-015-1765-5

Moore, I., Gessler, P., Nielsen, G.A., Peterson, G.A., 1993. Soil attribute prediction using terrain analysis. Soil Sci. Soc. Am. J. 57, 443–452. doi:10.2136/sssaj1993.572NPb

Moore, I.D., Grayson, R.B., Ladson, A.R., 1991. Digital Terrain Modeling : A Review of Hydrological Geomorphological and Biological Applications. Hydrol. Process. 5, 3–30. doi:DOI: 10.1002/hyp.3360050103

Morvan, X., Saby, N.P.A., Arrouays, D., Le Bas, C., Jones, R.J.A., Verheijen, F.G.A., Bellamy, P.H., Stephens, M., Kibblewhite, M.G., 2008. Soil monitoring in Europe: A review of existing systems and requirements for harmonisation. Sci. Total Environ. 391, 1–12. doi:10.1016/j.scitotenv.2007.10.046

Nachtergaele, F., van Velthuizen, H., Verelst, L., Batjes, N., Dijkshoorn, K., van Engelen, V., Fischer, G., Jones, A., Montanarella, L., Petri, M., 2008. Harmonized world soil database. Food Agric. Organ. United Nations.

Nadeu, E., Gobin, A., Fiener, P., van Wesemael, B., van Oost, K., 2015. Modelling the impact of agricultural management on soil carbon stocks at the regional scale: the role of lateral fluxes. Glob. Chang. Biol. 21, 3181–3192. doi:10.1111/gcb.12889

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M.E., Papritz, A., 2018. Evaluation of digital soil mapping approaches with large sets of environmental covariates. SOIL 4, 1–22. doi:10.5194/soil-4-1-2018

Napoli, R., Piccini, C., Di Bene, C., Farina, R., Marchetti, A., Gazzetti, C., Sarandrea, P., 2017. Agricultural activities effects on groundwater contamination in a Nitrate Vulnerable Zone of Latina Province. Rend. Online della Soc.

Geol. Ital. 42, 46–49. doi:10.3301/ROL.2017.11

Novara, A., Gristina, L., Kuzyakov, Y., Schillaci, C., Laudicina, V.A., La Mantia, T., 2013. Turnover and availability of soil organic carbon under different Mediterranean land-uses as estimated by ^{13}C natural abundance. Eur. J. Soil Sci. doi:10.1111/ejss.12038

Novara, A., Gristina, L., Sala, G., Galati, A., Crescimanno, M., Cerdà, A., Badalamenti, E., La Mantia, T., 2017. Agricultural land abandonment in Mediterranean environment provides ecosystem services via soil carbon sequestration. Sci. Total Environ. 576, 420–429. doi:10.1016/j.scitotenv.2016.10.123

Novara, A., La Mantia, T., Rühl, J., Badalucco, L., Kuzyakov, Y., Gristina, L., Laudicina, V.A., 2014. Dynamics of soil organic carbon pools after agricultural abandonment. Geoderma 235–236, 191–198. doi:10.1016/j.geoderma.2014.07.015

Odeh, I.O.A., Leenaars, J., Hartemink, A., 2012. The challenges of collating legacy data for digital mapping of Nigerian soils. Digit. Soil Assessments Beyond 453–458. doi:10.1201/b12728-88

Odeh, I.O.A., McBratney, A.B., Chittleborough, D.J., 1994. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. Geoderma 63, 197–214. doi:10.1016/0016-7061(94)90063-9

Ogle, S.M., Breidt, F.J., Easter, M., Williams, S., Killian, K., Paustian, K., 2010. Scale and uncertainty in modeled soil organic carbon stock changes for US croplands using a process-based model. Glob. Chang. Biol. 16, 810–822. doi:10.1111/j.1365-2486.2009.01951.x

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., Fernández-Ugalde, O., 2017. LUCAS Soil, the largest expandable soil dataset for Europe: a review. Eur. J. Soil Sci. doi:10.1111/ejss.12499

Orton, T.G., Pringle, M.J., Page, K.L., Dalal, R.C., Bishop, T.F.A., 2014. Spatial prediction of soil organic carbon stock using a linear model of coregionalisation. Geoderma 230, 119–130. doi:10.1016/j.geoderma.2014.04.016

Panagos, P., Ballabio, C., Yigini, Y., Dunbar, M.B., 2013a. Estimating the soil organic carbon content for European NUTS2 regions based on LUCAS data collection. Sci. Total Environ. 442, 235–246. doi:10.1016/J.SCITOTENV.2012.10.017

Panagos, P., Hiederer, R., Van Liedekerke, M., Bampa, F., 2013b. Estimating soil organic carbon in Europe based on data collected through an European network. Ecol. Indic. 24, 439–450. doi:10.1016/j.ecolind.2012.07.020

Panagos, P., Van Liedekerke, M., Jones, A., Montanarella, L., 2012. European Soil Data Centre: Response to European policy support and public data requirements. Land use policy 29, 329–338. doi:10.1016/j.landusepol.2011.07.003

Paolanti, M., Costantini, E.A.C., Fantappiè, M., Barbetti, R., 2010. La descrizione del suolo, in: Costantini, E.A.C. (Ed.), Linee Guida Dei Metodi per Il Rilevamento e l'informatizzazione Dei Dati Pedologici. SELCA, Firenze, Italia.

Parras-Alcántara, L., Lozano-García, B., Keesstra, S., Cerdà, A., Brevik, E.C., 2016. Long-term effects of soil management on ecosystem services and soil loss estimation in olive grove top soils. Sci. Total Environ. 571, 498–506. doi:10.1016/j.scitotenv.2016.07.016

Parras-Alcantara, L., Lozano-Garci-a, B., Keesstra, S., Cerda, A., Brevik, E.C., Parras-Alcántara, L., Lozano-García, B., Keesstra, S., Cerdà, A., Brevik, E.C., 2016. Long-term effects of soil management on ecosystem services and soil loss estimation in olive grove top soils. Sci. Total Environ. 571, 498–506. doi:10.1016/j.scitotenv.2016.07.016

Pellegrini, S., Vignozzi, N., Costantini, E.A.C., L'Abate, G., 2007. A new pedotransfer function for estimating soil bulk density, in: Dazzi, C. (Ed.), Changing Soils in a Changing World: The Soils of Tomorrow. 5th International Congress of European Society for Soil Conservation. 978–88–9572–09–2.

Pereira, P., Brevik, E., Trevisani, S., 2018. Mapping the environment. Sci. Total Environ. 610–611, 17–23. doi:10.1016/j.scitotenv.2017.08.001

Perri, E., Bernasconi, M.P., Cefalà, M., 2016. Quaternary carbonate deposition and climate variation (Tyrrhenian coast, Calabria, Southern Italy). Rend. Online della Soc. Geol. Ital. 38, 73–76. doi:10.3301/ROL.2016.21

Persichillo, M.G., Bordoni, M., Cavalli, M., Crema, S., Meisina, C., 2017. Evaluation of anthropogenic effects on the sediment delivery dynamics in response to slope instability. Rend. Online della Soc. Geol. Ital. 42, 5–9. doi:10.3301/ROL.2017.01

Phachomphon, K., Dlamini, P., Chaplot, V., 2010. Estimating carbon stocks at a regional level using soil information and easily accessible auxiliary variables. Geoderma 155, 372–380. doi:10.1016/j.geoderma.2009.12.020

Pielke, R.A., Marland, G., Betts, R.A., Chase, T.N., Eastman, J.L., Niles, J.O., Niyogi, D. d. S., Running, S.W., 2002. The influence of land-use change and landscape dynamics on the climate system: relevance to climate-change policy beyond the radiative effect of greenhouse gases. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 360, 1705–1719. doi:10.1098/rsta.2002.1027

Pisante, M., Stagnari, F., Acutis, M., Bindi, M., Brilli, L., Di Stefano, V., Carozzi, M., 2015. Conservation Agriculture and Climate Change, in: Conservation Agriculture. Springer, pp. 579–620.

Poeplau, C., Vos, C., Don, A., 2017. Soil organic carbon stocks are systematically overestimated by misuse of the parameters bulk density and rock fragment content. SOIL 3, 61–66. doi:10.5194/soil-3-61-2017

Poggio, L., Gimona, A., Brewer, M.J., 2013. Regional scale mapping of soil properties and their uncertainty with a large number of satellite-derived covariates. Geoderma 209, 1–14. doi:10.1016/j.geoderma.2013.05.029

Post, W.M., Emanuel, W.R., Zinke, P.J., Stangenberger, A.G., 1982. Soil carbon pools and world life zones. Nature 298, 156–159. doi:10.1038/298156a0

Post W.M and Kwon K.C.., 2000. Soil carbon sequestration and land-use change: processes and potential. Glob. Chang. Biol. 6, 317–327. doi:10.1046/j.1365-2486.2000.00308.x

Priori, S., Fantappiè, M., Bianconi, N., Ferrigno, G., Pellegrini, S., Costantini, E.A.C., 2016. Field-Scale Mapping of Soil Carbon Stock with Limited Sampling by Coupling Gamma-Ray and Vis-NIR Spectroscopy. Soil Sci. Soc. Am. J. 80, 954. doi:10.2136/sssaj2016.01.0018

Purton, K., Pennock, D., Leinweber, P., Walley, F., 2015. Will changes in climate and land use affect soil organic matter composition? Evidence from an ecotonal climosequence. Geoderma 253–254, 48–60. doi:10.1016/j.geoderma.2015.04.007

R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. Vienna Austria R Found. Stat. Comput. doi:10.1007/978-3-540-74686-7

Rabbi, S.M.F., Tighe, M., Delgado-Baquerizo, M., Cowie, A., Robertson, F., Dalal, R., Page, K., Crawford, D., Wilson, B.R., Schwenke, G., Mcleod, M., Badgery, W., Dang, Y.P., Bell, M., O'Leary, G., Liu, D.L., Baldock, J., 2016. Climate and soil properties limit the positive effects of land use reversion on carbon storage in Eastern Australia. Sci. Rep. 5, 17866. doi:10.1038/srep17866

Ramos, T.B., Horta, A., Gonçalves, M.C., Pires, F.P., Duffy, D., Martins, J.C., 2017. The INFOSOLO database as a first step towards the development of a soil information system in Portugal. CATENA 158, 390–412. doi:10.1016/J.CATENA.2017.07.020

Randall, N.P., James, K.L., 2012. The effectiveness of integrated farm management, organic farming and agri-environment schemes for conserving biodiversity in temperate Europe - A systematic map. Environ. Evid. 1, 4. doi:10.1186/2047-2382-1-4

Rhee, J., Im, J., 2017. Meteorological drought forecasting for ungauged areas based on machine learning: Using long-range climate forecast and remote sensing data. Agric. For. Meteorol. 237–238, 105–122. doi:https://doi.org/10.1016/j.agrformet.2017.02.011

Rial, M., Martínez Cortizas, A., Rodríguez-Lado, L., 2017a. Understanding the spatial distribution of factors controlling topsoil organic carbon content in European soils. Sci. Total Environ. 609, 1411–1422. doi:10.1016/j.scitotenv.2017.08.012

Rial, M., Martínez Cortizas, A., Taboada, T., Rodríguez-Lado, L., 2017b. Soil organic carbon stocks in Santa Cruz Island, Galapagos, under different climate change scenarios. CATENA 156, 74–81. doi:10.1016/j.catena.2017.03.020

Ribeiro, E., Batjes, N.H., Leenaars, J.G.B., Oostrum, A. Van, Jesus, J.M. De, 2015. Towards the standardization and harmonization of world soil data.

Riffaldi, R., Saviozzi, A., Levi-Minzi, R., Menchetti, F., 1994. Chemical characteristics of soil after 40 years of continuous maize cultivation. Agric. Ecosyst. Environ. 49, 239–245. doi:10.1016/0167-8809(94)90053-1

Rivera, D., Sandoval, M., Godoy, A., 2015. Exploring soil databases: a self-organizing map approach. Soil Use Manag. 31, 121–131. doi:10.1111/sum.12169

Rodeghiero, M., Rubio, A., Díaz-Pinés, E., Romanyà, J., Marañón-Jiménez, S., Levy, G.J., Fernandez-Getino, A.P., Sebastià, M.T., Karyotis, T., Chiti, T., Sirca, C., Martins, A., Madeira, M., Zhiyanski, M., Gristina, L., La Mantia, T., 2011. Soil Carbon in Mediterranean Ecosystems and Related Management Problems, in: Soil Carbon in Sensitive European Ecosystems. John Wiley & Sons, Ltd, Chichester, UK, pp. 175–218. doi:10.1002/9781119970255.ch8

Rodríguez Martín, J.A., Álvaro-Fuentes, J., Gonzalo, J., Gil, C., Ramos-Miras, J.J., Grau Corbí, J.M., Boluda, R., 2016. Assessment of the soil organic carbon stock in Spain. Geoderma 264, 117–125. doi:10.1016/j.geoderma.2015.10.010

Ross, C.W., Grunwald, S., Myers, D.B., 2013. Spatiotemporal modeling of soil organic carbon stocks across a subtropical region. Sci. Total Environ. 461–462, 149–157. doi:10.1016/j.scitotenv.2013.04.070

Rouse Jr, J., Haas, R., Schell, J., Deering, D., 1974. Monitoring Vegetation Systems in the Great Plains with ERTS, in: Fraden, S., Marcanti, E., Becker, M. (Eds.), Third ERTS-1 Symposium. NASA-SP-351, Washington DC, pp. 309–317.

Ruisi, P., Giambalvo, D., Saia, S., Di Miceli, G., Frenda, A.S.A.S., Plaia, A., Amato, G., 2014. Conservation tillage in a semiarid Mediterranean environment: results of 20 years of research. Ital. J. Agron. 9, 1. doi:10.4081/ija.2014.560

Saia, S., Benítez, E., García-Garrido, J.M., Settanni, L., Amato, G., Giambalvo, D., 2014. The effect of arbuscular mycorrhizal fungi on total plant nitrogen uptake and nitrogen recovery from soil organic material. J. Agric. Sci. 152, 370–378. doi:10.1017/S002185961300004X

Saia, S., Cappelletti, G.M., Russo, C., Nicoletti, G.M., Lo Storto, M.C., De Vita, P., 2017a. The Life Cycle Assessment of durum wheat yield in contrasting management systems in Mediterranean environment, in: Ventura, F., Seddaiu, G., Cola, G. (Eds.), Stategie Integrate per Affrontare Le Sfide Climatiche e Agronomiche Nella Gestione Dei Sistemi Agroalimentari - Joint Strategies to Cope with Climate Change and Agronomical Management in Food and Feed Systems. Atti Del XX Convegno AIAM e XLVI Convegno SIA. Milan, 12-14 September 2017, pp. 189–191. doi:10.6092/unibo/amsacta/5692

Saia, S., Schillaci, C., Lipani, A., Fantappiè, M., Märker, M., Lombardo, L., Matranga, M.G., Ferraro, V., Guaitoli, F., Acutis, M., 2017b. Protection of soil from the loss of organic carbon by taking into account erosion and managing land use at varying soil type: indication from a model semiarid area, in: Global Symposium on Soil Organic Carbon. Rome, Italy, pp. 510–515.

Said, M.E.-A., Militello, M., Saia, S., Settanni, L., Aleo, A., Mammina, C., Bombarda, I., Vanloot, P., Roussel, C., Dupuy, N., 2016. Artemisia arborescens Essential Oil Composition, Enantiomeric Distribution, and Antimicrobial Activity from Different Wild Populations from the Mediterranean Area. Chem. Biodivers. 13, 1095–1102. doi:10.1002/cbdv.201500510

Samuelsson, O., Björk, A., Zambrano, J., Carlsson, B., 2017. Gaussian process regression for monitoring and fault detection of wastewater treatment processes. Water Sci. Technol. 75, 2952–2963. doi:10.2166/wst.2017.162

Sánchez, B., Iglesias, A., McVittie, A., Álvaro-Fuentes, J., Ingram, J., Mills, J., Lesschen, J.P., Kuikman, P.J., 2016. Management of agricultural soils for greenhouse gas mitigation: Learning from a case study in NE Spain. J. Environ. Manage. 170, 37–49. doi:10.1016/j.jenvman.2016.01.003

Schapire, R.E., Freund, Y., 2012. Boosting: Foundations and algorithms. MIT press, Cambridge, Massachussets.

Schillaci, C.; Braun, A.; Kropácek, J., 2015. TerrainAnalysis and Landform Recognition, in: In: Clarke, L.E & Nield, J.M. (Eds. . (Ed.), Geomorphological Techniques (Online Edition). Chapter 2.4.2. British Society for Geomorphology; London, UK. ISSN: 2047-0371.

Schillaci, C., Acutis, M., Lombardo, L., Lipani, A., Fantappiè, M., Märker, M., Saia, S., 2017a. 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. 601–602, 821–832.

doi:10.1016/j.scitotenv.2017.05.239

Schillaci, C., Lombardo, L., Saia, S., Fantappiè, M., Märker, M., Acutis, M., 2017b. Modelling the topsoil carbon stock of agricultural lands with the Stochastic Gradient Treeboost in a semi-arid Mediterranean region. Geoderma 286, 35–45. doi:10.1016/j.geoderma.2016.10.019

Schillaci, C., Acutis, M., Vesely, F., Saia, S., 2018a. A simple pipeline for the assessment of legacy soil datasets: an example and test with soil organic carbon from a highly variable area. Catena. doi:10.1016/j.catena.2018.12.015

Schillaci, C., Saia, S., Acutis, M., 2018b. Modelling of Soil Organic Carbon in the Mediterranean area: a systematic map. Rend. Online della Soc. Geol. Ital. 4, 161–166.

Schmolke, A., Thorbek, P., DeAngelis, D.L., Grimm, V., 2010. Ecological models supporting environmental decision making: a strategy for the future. Trends Ecol. Evol. 25, 479–486. doi:10.1016/j.tree.2010.05.001

Schoeneberger, P.J., Wysocki, D.A., Benham, E.C., Soil_Survey_Staff, 2012. Field book for describing and sampling soils, Version 3.0. Lincoln, NE.

Scull, P., Franklin, J., Chadwick, O.A., McArthur, D., 2003. Predictive soil mapping: a review. Prog. Phys. Geogr. 27, 171–197. doi:10.1191/0309133303pp366ra

Seen, D. Lo, Ramesh, B.R., Nair, K.M., Martin, M., Arrouays, D., Bourgeon, G., 2010. Soil carbon stocks, deforestation and land-cover changes in the Western Ghats biodiversity hotspot (India). Glob. Chang. Biol 16, 1777–1792. doi:10.1111/j.1365-2486.2009.02127.x

Sequeira, C.H., Wills, S.A., Seybold, C.A., West, L.T., 2014. Predicting soil bulk density for incomplete databases. Geoderma 213, 64–73. doi:10.1016/j.geoderma.2013.07.013

Shary, P.A., Sharaya, L.S., Mitusov, A. V., 2002. Fundamental quantitative methods of land surface analysis. Geoderma 107, 1–32. doi:10.1016/S0016-7061(01)00136-7

Six, J., Paustian, K., 2014. Aggregate-associated soil organic matter as an ecosystem property and a measurement tool. Soil Biol. Biochem. 68, A4–A9. doi:10.1016/J.SOILBIO.2013.06.014

Smith, P., 2004. Carbon sequestration in croplands: the potential in Europe and the global context. Eur. J. Agron. 20, 229–236. doi:10.1016/J.EJA.2003.08.002

Smith, P., Cotrufo, M.F., Rumpel, C., Paustian, K., Kuikman, P.J., Elliott, J.A., McDowell, R., Griffiths, R.I., Asakawa, S., Bustamante, M., House, J.I., Sobocká, J., Harper, R., Pan, G., West, P.C., Gerber, J.S., Clark, J.M., Adhya, T., Scholes, R.J., Scholes, M.C., 2015. Biogeochemical cycles and biodiversity as key drivers of ecosystem services provided by soils. SOIL 1, 665–685. doi:10.5194/soil-1-665-2015

Smith, P., Lutfalla, S., Riley, W.J., Torn, M.S., Schmidt, M.W.I., Soussana, J.-F., 2018. The changing faces of soil organic matter research. Eur. J. Soil Sci. 69, 23–30. doi:10.1111/ejss.12500

Soane, B.D., 1990. The role of organic matter in soil compactibility: A review of some practical aspects. Soil Tillage Res. 16, 179–201. doi:10.1016/0167-1987(90)90029-D

Söderström, B., Hedlund, K., Jackson, L.E., Kätterer, T., Lugato, E., Thomsen, I.K., Bracht Jørgensen, H., 2014. What

are the effects of agricultural management on soil organic carbon (SOC) stocks? Environ. Evid. 3, 2. doi:10.1186/2047-2382-3-2

Sokos, C.K., Mamolos, A.P., Kalburtji, K.L., Birtsas, P.K., 2013. Farming and wildlife in Mediterranean agroecosystems. J. Nat. Conserv. 21, 81–92. doi:10.1016/J.JNC.2012.11.001

Song, Y.-Q., Yang, L.-A., Li, B., Hu, Y.-M., Wang, A.-L., Zhou, W., Cui, X.-S., Liu, Y.-L., Song, Y.-Q., Yang, L.-A., Li, B., Hu, Y.-M., Wang, A.-L., Zhou, W., Cui, X.-S., Liu, Y.-L., 2017. Spatial Prediction of Soil Organic Matter Using a Hybrid Geostatistical Model of an Extreme Learning Machine and Ordinary Kriging. Sustainability 9, 754. doi:10.3390/su9050754

Sperow, M., 2016. Estimating carbon sequestration potential on U.S. agricultural topsoils. Soil Tillage Res. 155, 390–400. doi:10.1016/J.STILL.2015.09.006

Stockmann, U., Adams, M.A., Crawford, J.W., Field, D.J., Henakaarchchi, N., Jenkins, M., Minasny, B., McBratney, A.B., Courcelles, V. de R. de, Singh, K., Wheeler, I., Abbott, L., Angers, D.A., Baldock, J., Bird, M., Brookes, P.C., Chenu, C., Jastrow, J.D., Lal, R., Lehmann, J., O'Donnell, A.G., Parton, W.J., Whitehead, D., Zimmermann, M., 2013. The knowns, known unknowns and unknowns of sequestration of soil organic carbon. Agric. Ecosyst. Environ. 164, 80–99. doi:10.1016/j.agee.2012.10.001

Strobl, C., Malley, J., Tutz, G., 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol. Methods 14, 323–48. doi:10.1037/a0016973

Stumpf, F., Schmidt, K., Behrens, T., Schönbrodt-Stitt, S., Buzzo, G., Dumperth, C., Wadoux, A., Xiang, W., Scholten, T., 2016. Incorporating limited field operability and legacy soil samples in a hypercube sampling design for digital soil mapping. J. Plant Nutr. Soil Sci. 179, 499–509. doi:10.1002/jpln.201500313

Sulaeman, Y., Minasny, B., McBratney, A.B., Sarwani, M., Sutandi, A., 2013. Harmonizing legacy soil data for digital soil mapping in Indonesia. Geoderma 192, 77–85. doi:10.1016/j.geoderma.2012.08.005

Sun, X.-L., Wu, Y.-J., Lou, Y.-L., Wang, H.-L., Zhang, C., Zhao, Y.-G., Zhang, G.-L., 2015. Updating digital soil maps with new data: a case study of soil organic matter in Jiangsu, China. Eur. J. Soil Sci. 66, 1012–1022. doi:10.1111/ejss.12295

Suuster, E., Ritz, C., Roostalu, H., Kõlli, R., Astover, a., 2012. Modelling soil organic carbon concentration of mineral soils in arable land using legacy soil data. Eur. J. Soil Sci. 63, 351–359. doi:10.1111/j.1365-2389.2012.01440.x

Szatmári, G., Pásztor, L., 2018. Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms. Geoderma. doi:10.1016/J.GEODERMA.2018.09.008

Teng, H., Viscarra Rossel, R.A., Shi, Z., Behrens, T., 2018. Updating a national soil classification with spectroscopic predictions and digital soil mapping. CATENA 164, 125–134. doi:10.1016/j.catena.2018.01.015

Thorn, J.P.R., Friedman, R., Benz, D., Willis, K.J., Petrokofsky, G., 2016. What evidence exists for the effectiveness of on-farm conservation land management strategies for preserving ecosystem services in developing countries? A systematic map. Environ. Evid. 5, 13. doi:10.1186/s13750-016-0064-9

Underwood, E.C., Viers, J.H., Klausmeyer, K.R., Cox, R.L., Shaw, M.R., 2009. Threats and biodiversity in the mediterranean biome. Divers. Distrib. 15, 188–197. doi:10.1111/j.1472-4642.2008.00518.x

Vaysse, K., Heuvelink, G.B.M., Lagacherie, P., 2017. Spatial aggregation of soil property predictions in support of local land management. Soil Use Manag. 33, 299–310. doi:10.1111/sum.12350

Vaysse, K., Lagacherie, P., 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma 291, 55–64. doi:10.1016/j.geoderma.2016.12.017

Vaysse, K., Lagacherie, P., 2015. Evaluating Digital Soil Mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France). Geoderma Reg. 4, 20–30. doi:10.1016/j.geodrs.2014.11.003

Vega, G., Pertierra, L.R., Olalla-Tárraga, M.Á., 2017. MERRAclim, a high-resolution global dataset of remotely sensed bioclimatic variables for ecological modelling. Sci. Data 4, 170078. doi:10.1038/sdata.2017.78

Velthof, G.L., Kuikman, P.J., Oenema, O., 2002. Nitrous oxide emission from soils amended with crop residues. Nutr. Cycl. Agroecosystems 62, 249–261. doi:10.1023/A:1021259107244

Venables, W.N., Smith, D.M., 1990. An Introduction to R Notes on R: A Programming Environment for Data Analysis and Graphics Version 3.4.3 (2017-11-30). R. Gentlem. R. Ihaka Copyr. c.

Ventrella, D., Fiore, A., Vonella, A.V., Fornaro, F., 2011. Effectiveness of the GAEC cross-compliance standard management of stubble and crop residues in the maintenance of adequate contents of soil organic carbon. Ital. J. Agron. 6, 7. doi:10.4081/ija.2011.6.s1.e7

Vereecken, H., Schnepf, A., Hopmans, J.W., Javaux, M., Or, D., Roose, T., Vanderborght, J., Young, M.H., Amelung, W., Aitkenhead, M., Allison, S.D., Assouline, S., Baveye, P., Berli, M., Brüggemann, N., Finke, P., Flury, M., Gaiser, T., Govers, G., Ghezzehei, T., Hallett, P., Hendricks Franssen, H.J., Heppell, J., Horn, R., Huisman, J.A., Jacques, D., Jonard, F., Kollet, S., Lafolie, F., Lamorski, K., Leitner, D., McBratney, A., Minasny, B., Montzka, C., Nowak, W., Pachepsky, Y., Padarian, J., Romano, N., Roth, K., Rothfuss, Y., Rowe, E.C., Schwen, A., Šimůnek, J., Tiktak, A., Van Dam, J., van der Zee, S.E.A.T.M., Vogel, H.J., Vrugt, J.A., Wöhling, T., Young, I.M., 2016. Modeling Soil Processes: Review, Key Challenges, and New Perspectives. Vadose Zo. J. 15.

Vermeulen, D., Van Niekerk, A., 2017. Machine learning performance for predicting soil salinity using different combinations of geomorphometric covariates. Geoderma 299, 1–12. doi:10.1016/J.GEODERMA.2017.03.013

Veronesi, F., Corstanje, R., Mayr, T., 2012. Mapping soil compaction in 3D with depth functions. Soil Tillage Res. 124, 111–118. doi:10.1016/J.STILL.2012.05.009

Veronesi, F., Grassi, S., Raubal, M., 2016. Statistical learning approach for wind resource assessment. Renew. Sustain. Energy Rev. 56, 836–850. doi:10.1016/j.rser.2015.11.099

Veronesi, F., Hurni, L., 2014. Random Forest with semantic tie points for classifying landforms and creating rigorous shaded relief representations. Geomorphology 224, 152–160. doi:10.1016/J.GEOMORPH.2014.07.020

Veronesi, F., Korfiati, A., Buffat, R., Raubal, M., 2017. Assessing Accuracy and Geographical Transferability of Machine Learning Algorithms for Wind Speed Modelling. pp. 297–310. doi:10.1007/978-3-319-56759-4_17

Verrelst, J., Rivera, J.P., Moreno, J., Camps-Valls, G., 2013. Gaussian processes uncertainty estimates in experimental Sentinel-2 LAI and leaf chlorophyll content retrieval. ISPRS J. Photogramm. Remote Sens. 86, 157–167. doi:https://doi.org/10.1016/j.isprsjprs.2013.09.012

Viola, F., Liuzzo, L., Noto, L. V., Lo Conti, F., La Loggia, G., 2014. Spatial distribution of temperature trends in Sicily. Int. J. Climatol. 34, 1–17. doi:10.1002/joc.3657

Viscarra Rossel, R.A., Brus, D.J., Lobsey, C., Shi, Z., McLachlan, G., 2016. Baseline estimates of soil organic carbon by proximal sensing: Comparing design-based, model-assisted and model-based inference. Geoderma 265, 152–163. doi:10.1016/j.geoderma.2015.11.016

Walkley, A., Black, I.A., 1934. An examination of the degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method. Soil Sci. 37, 29–38. doi:10.1097/00010694-193401000-00003

Wang, S., Wang, Q., Adhikari, K., Jia, S., Jin, X., Liu, H., 2016. Spatial-Temporal Changes of Soil Organic Carbon Content in Wafangdian, China. Sustainability 8, 1154. doi:10.3390/su8111154

Webster, R. (Richard), Oliver, M.A., 2007. Geostatistics for environmental scientists. Wiley.

Wei, H., Guenet, B., Vicca, S., Nunan, N., Asard, H., AbdElgawad, H., Shen, W., Janssens, I.A., 2014. High clay content accelerates the decomposition of fresh organic matter in artificial soils. Soil Biol. Biochem. 77, 100–108. doi:10.1016/j.soilbio.2014.06.006

Weiss, A.D., 2001. Topographic Position and Landforms analysis. doi:http://www.jennessent.com/downloads/TPI-poster-TNC_18x22.pdf

Wiesmeier, M., Urbanski, L., Hobley, E., Lang, B., von Lützow, M., Marin-Spiotta, E., van Wesemael, B., Rabot, E., Ließ, M., Garcia-Franco, N., Wollschläger, U., Vogel, H.-J., Kögel-Knabner, I., 2019. Soil organic carbon storage as a key function of soils - A review of drivers and indicators at various scales. Geoderma 333, 149–162. doi:10.1016/j.geoderma.2018.07.026

Wilson, J.P. , Gallant, J.C., 2000. Terrain Analysis: Principles and Applications. Wiley.

Xiong, X., Grunwald, S., Myers, D.B., Kim, J., Harris, W.G., Bliznyuk, N., 2015. Assessing uncertainty in soil organic carbon modeling across a highly heterogeneous landscape. Geoderma 251–252, 105–116. doi:10.1016/j.geoderma.2015.03.028

Yaalon, D.H., 1997. Soils in the Mediterranean region: what makes them different? CATENA 28, 157–169. doi:10.1016/S0341-8162(96)00035-5

Yang, R., Rossiter, D.G., Liu, F., Lu, Y., Yang, F., Yang, F., Zhao, Y., Li, D., Zhang, G., 2015. Predictive Mapping of Topsoil Organic Carbon in an Alpine Environment Aided by Landsat TM. PLoS One 10, e0139042. doi:10.1371/journal.pone.0139042

Yao, X., Fu, B., Lü, Y., Sun, F., Wang, S., Liu, M., 2013. Comparison of four spatial interpolation methods for estimating soil moisture in a complex terrain catchment. PLoS One 8, e54660. doi:10.1371/journal.pone.0054660

Yigini, Y., Panagos, P., 2016. Assessment of soil organic carbon stocks under future climate and land cover changes in Europe. Sci. Total Environ. 557, 838–850. doi:10.1016/j.scitotenv.2016.03.085

Zamora, J., Verdú, J.R., Galante, E., 2007. Species richness in Mediterranean agroecosystems: Spatial and temporal analysis for biodiversity conservation. Biol. Conserv. 134, 113–121. doi:10.1016/J.BIOCON.2006.08.011

Zevenbergen, L.W., Thorne, C.R., 1987. Quantitative analysis of land surface topography. Earth Surf. Process. Landforms 12, 47–56. doi:10.1002/esp.3290120107

Zinn, Y.L., Lal, R., Resck, D.V.S., 2005a. Texture and organic carbon relations described by a profile pedotransfer function for Brazilian Cerrado soils, Geoderma. doi:10.1016/j.geoderma.2005.02.010

Zinn, Y.L., Lal, R., Resck, D.V.S.S., 2005b. Changes in soil organic carbon stocks under agriculture in Brazil. Soil Tillage Res. 84, 28–40. doi:10.1016/j.still.2004.08.007

Useful links

Agriforecast MARS

http://agri4cast.jrc.ec.europa.eu/DataPortal/Index.aspx

https://g4aw.spaceoffice.nl/en/projects/international/international-initiatives/monitoring-agricultural-resources-mars/

Centro nazionale cartografia pedologica

http://www.soilmaps.it/en/downloads.html

TIMESAT

http://web.nateko.lu.se/timesat/timesat.asp?cat=0

MODIS

https://newsroom.gsfc.nasa.gov/sdptoolkit/HEG/HEGHome.html

CORINE

https://www.eea.europa.eu/data-and-maps/data#c0=5&c11=&c5=all&b_start=0&c12=land+cover

Examples of country Agriculture department sites and level of information offered

http://www.abs.gov.au/Agriculture

Environmental information data centre

http://eidc.ceh.ac.uk/citing-data

Environmental monitoring

http://www.openforis.org/home.html

Github - 4geocomPP Lovelace

https://github.com/geocomPP/sdvwR/blob/master/Section4_Final_Example_Conclusions.md

Basemaps

http://www.naturalearthdata.com/downloads/

Metadata

https://geonetwork-opensource.org/

NOAA

https://www.nesdis.noaa.gov/JPSS-1

Open data Europe

https://www.europeandataportal.eu/elearning/it/#/id/co-01

Esdac ecopedological maps

https://esdac.jrc.ec.europa.eu/Library/Data/250000/Italy/MapRegions.htm

European soil database

https://esdac.jrc.ec.europa.eu/ESDB_Archive/ESDBv2/fr_thema.htm

European regional data

https://esdac.jrc.ec.europa.eu/content/regional-data

European land observation

https://land.copernicus.eu