New pedotransfer approaches to predict soil bulk density using WoSIS soil data and environmental covariates in Mediterranean agro-ecosystems

Calogero Schillaci, Alessia Perego, Elena Valkama, Michael Märker, Sergio Saia, Fabio Veronesi, Aldo Lipani, Luigi Lombardo, Tommaso Tadiello, Hannes A. Gamper, Luigi Tedone, Cami Moss, Elena Pareja-Serrano, Gabriele Amato, Kersten Kühl, Claudia Damatirca, Alessia Cogato, Nada Mzid, Rasu Eeswaran, Marya Rebelo, Giorgio Sperandio, Alberto Bosino, Margherita Bufalini, Tülay Tunçağ, Jianqi Ding, Marco Fiorentini, Guadalupe Tiscornia, Sarah Conradt, Marco Botta, Marco Acutis

PII: S0048-9697(21)01677-6
DOI: https://doi.org/10.1016/j.scitotenv.2021.146609
Reference: STOTEN 146609

To appear in: Science of the Total Environment

Received date: 26 December 2020
Revised date: 24 February 2021
Accepted date: 16 March 2021


This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
New pedotransfer approaches to predict soil bulk density using WoSIS soil data and environmental covariates in Mediterranean agro-ecosystems

Calogero Schillaci 1, Alessia Perego 1*, Elena Valkama 2, Michael Märker 3, Sergio Saia 4, Fabio Veronesi 5, Aldo Lipani 6, Luigi Lombardo 7, Tommaso Tadiello 1, Hannes A. Gamper 8, Luigi Tedone 9, Cami Moss 10, Elena Pareja-Serrano 11, Gabriele Amato 12, Kersten Kühl 13, Claudia Damatirca 14, Alessia Cogato 15, Nada Mzid 16, Rasu Eeswaran 17, Marya Rebelo 18, Giorgio Sperandio 19, Alberto Bosino 3, Margherita Bufalini 20, Tülay Tunçay 21, Jianqi Ding 22, Marco Fiorentini 23, Guadalupe Tiscornia 24, Sarah Conradt 25, Marco Botta 1, Marco Acutis 1
1 Department of Agricultural and Environmental Science, University of Milan, Via Via Celoria 2
2 Milan, Italy
2 Natural Resources Institute Finland (Luke), Bioeconomy and Environment, FI-31600, Tietotie 4, Jokioinen, Finland
3 Department of Earth and Environmental Sciences, University of Pavia, Via Ferrata, 1, 27100 Pavia, Italy
4 Department of Veterinary Sciences, University of Pisa, Via delle Piagge 2, Pisa 56129, Italy
5 Water Research Centre Limited, Frankland Road, Blagrove, Swindon, Wiltshire, SN56 8YF, England
6 Department of Web Intelligence Group, University College London (UCL) 90 High Holborn
7 London, London WC1E 7HT, UK
8 Department of Agricultural and Environmental Science, University of Bari Aldo Moro, Via Amendola 165/A, 70126 Bari, Italy
9 10 Department of Population Health, London School of Hygiene & Tropical Medicine, London WC1E 7HT, UK
11 Remote Sensing and GIS Group, Instituto de Desarrollo Regional, University of Castilla-La Mancha, 02071 Campus of Albacete, Spain
12 IFAC CNR, Via Madonna del Piano, 10, 50019 Sesto Fiorentino, Florence, Italy
13 Department of Geography Ludwig-Maximilians-University of Munich, Germany
14 Department of Agriculture, Forest and Food Sciences, University of Torino, largo Braccini 2, 10095 Grugliasco, Italy
15 Department of Land, Environmental, Agriculture and Forestry, University of Padova, 35020 Legnaro, Italy
16 Department of Sciences and Technology for Agriculture, Forest, Environment and Energy, DAFNE, Università della Tuscia, Viterbo, Italy
17 Department of Plant, Soil and Microbial Sciences, Michigan State University, East Lansing, MI48824, USA
18 Department of Agriculture, Food and Environment, University of Pisa, via del Borghetto 80, 56124 Pisa, Italy
19 Dipartimento di Medicina Molecolare e Traslazionale, Università di Brescia, Viale Europa, 11, 25123 Brescia, Italy
20 University of Camerino, School of Science and Technology-Geology Division, Via Gentile III da Varano, Camerino 62032, Italy
21 Soil Fertilizer and Water Resources Central Research Institute, Ankara-Turkey
Abstract

For the estimation of the soil organic carbon stocks, bulk density (BD) is a fundamental parameter but measured data are usually not available especially when dealing with legacy soil data. It is possible to estimate BD by applying pedotransfer function (PTF). We applied different estimation methods with the aim to define a suitable PTF for BD of arable land for the Mediterranean Basin, which has peculiar climate features that may influence the soil carbon sequestration. To improve the existing BD estimation methods, we used a set of public climatic and topographic data along with the soil texture and organic carbon data. The present work consisted of the following steps: i) development of three PTFs models separately for top (0-0.4 m) and subsoil (0.4-1.2 m), ii) a 10-fold cross-validation, iii) model transferability using an external dataset derived from published data.

The development of the new PTFs was based on the training dataset consisting of World Soil Information Service (WoSIS) soil profile data, climatic data from WorldClim at 1 km spatial resolution and Shuttle Radar Topography Mission (SRTM) digital elevation model at 30 m spatial resolution.

The three PTFs models were developed using: Multiple Linear Regression stepwise (MLR-S), Multiple Linear Regression backward stepwise (MLR-BS), and Artificial Neural Network (ANN).

The predictions of the newly developed PTFs were compared with the BD calculated using the
PTF proposed by Manrique and Jones (MJ) and the modelled BD derived from the global SoilGrids dataset.

For the topsoil training dataset (N=129), MLR-S, MLR-BS and ANN had a R² 0.35, 0.58 and 0.86, respectively. For the model transferability, the three PTFs applied to the external topsoil dataset (N=59), achieved R² values of 0.06, 0.03 and 0.41. For the subsoil training dataset (N=180), MLR-S, MLR-BS and ANN the R² values were 0.36, 0.46 and 0.83, respectively. When applied to the external subsoil dataset (N=29), the R² values were 0.05, 0.06 and 0.41. The cross-validation for both top and subsoil dataset, resulted in an intermediate performance compared to calibration and validation with the external dataset. The new ANN PTF outperformed MLR-S, MLR-BS, MJ and SoilGrids approaches for estimating BD. Further improvements may be achieved by additionally considering the time of sampling, agricultural soil management and cultivation practices in predictive models.

**Highlights**
- Three PTFs were developed to calculate bulk density of arable top- and subsoil
- WoSIS, WorldClim, and topographic data of the Mediterranean Basin were used
- Model transferability of the three new PTFs was validated with external dataset
- Topsoil ANN-PTF had R² of 0.89 in training and 0.45 in model transferability
- ANN-PTF outperformed the commonly employed PTF by Manrique and Jones

**Keywords:** Agriculture, bulk density, pedotransfer functions, PTFs, soil carbon, soil texture

**1. Introduction**

Soil bulk density (BD) is directly linked to soil functionality including mechanical support of crop plants, circulation of soil solution, and soil aeration (Håkansson and Lipiec, 2000). Relatively high values of BD indicate soil compaction which may lead to reduced water infiltration especially in agricultural land, where it can hamper the growth of crop root systems (Colombi et al., 2018). Soil BD is calculated as the dry weight of soil divided by its volume. Volumes include soil particle volume and pore space between soil particles. Soil BD is typically
expressed in g cm$^{-3}$ or Mg m$^{-3}$ (SI). Along with soil organic carbon (SOC) concentrations, soil BD is necessary to calculate SOC stocks (Minasny et al., 2013) and to assess carbon sequestration (Tao et al., 2019). Many soil physical and chemical properties are expressed on a volumetric basis and in particular the estimation of soil biological properties depend on BD estimates (Tejada et al., 2009). In arable lands, tillage and other management practices cause high variation of BD during the year. Scientists have tried to infer BD from soil properties that are routinely measured such as textural information and organic carbon content (Acutis and Donatelli, 2003; Alvarez-Acosta et al., 2012; Pachepsky et al., 1996; Van Looy et al., 2017). The functions enabling the estimation of a given soil property (e.g. BD) from other variables, routinely obtained through laboratory measurement, are called pedotransfer functions (PTF) (Bouma, 1989; Patil and Singh, 2016). PTFs have been used at global scale to estimate the soil water retention, soil particle size, soil BD and SOC stock (Batjes and Dijkshoorn, 1999; Rawls, 1983; Rawls and Pachepsky, 2002; Reynolds et al., 2000; Saxton et al., 1986). At this scale, soil BD models had limited predictive ability (Rawls, 1983; Tietje and Tapkenhinrichs, 1993). Unfortunately, PTFs are not able to fully replace direct measurements, as highlighted in a recent publication which compared >50 PTFs using high resolution geodata in at district scale (Nasta et al., 2020; Xiangsheng et al., 2016). PTF are also frequently chosen at district scales after a sensitivity analysis (Basile et al., 2019).

Accurate models are of high interest for land management and policy-making especially where sparse data are available.

Today, BD estimates are used to quantify and model the SOC stocks in top- and subsoil at regional and global scales (Valkama et al., 2020). For example, Sun et al. (2020) recently used PTF in a meta-analysis to assess the effect of conservation agriculture on carbon stocks but did
not provide an assessment of the PTF function performance.

One of the first attempts to estimate BD was made by Manrique and Jones (1991) who proposed a PTF based on SOC alone (BD=1.660-0.318·SOC\(^{0.5}\)) for all soil types. Since then, other PTFs for BD estimation have been developed based on the fine earth fractions and SOC, which is important to BD due to its effect on the ratio between soil macro- and micropores (Martín et al., 2017; Throop et al., 2012). Furthermore, many other functions have been proposed to describe regional (Akpa et al. 2016; Chagas et al., 2016; Chen et al., 2016; Makovníková et al., 2017; Montzka et al., 2017; Ramcharan et al., 2017; Román Dobarco et al., 2019; Wösten et al., 2013, 1999) and local conditions (Benites et al., 2007; De Vos et al., 2005; Picciafuoco et al., 2019; Sevastas et al., 2018) and to predict BD in different soil horizons (Hollis et al., 2012; Reidy et al., 2016; Sequeira et al., 2014).

In the absence of measured soil data, the availability of new topographic data such as digital elevation models and morphometric indices has also improved soil assessment (Lombardo et al. 2018, Schillaci et al., 2017a, b, 2019; Veronesi and Schillaci, 2019) and in particular to develop PTF (Leij et al., 2004; Romano and Chirico, 2004). Other geodata (e.g., climate, satellite-derived, land cover) correlated with BD have also been used to improve estimates (Aitkenhead and Coull, 2020). Various researchers have recently developed new methods to estimate BD. Bondi et al., (2018) estimated BD for peat soils using soil visual assessment, and decision trees achieving similar performances, with around 0.6 explained variance. Premrov et al., (2018) achieved similar performances (R\(^2\) from 0.4 to 0.6) using optimal power-transformation of measured physical and chemical soil parameters.

Chen et al. (2018) formalized an analytical protocol to test the PTF prediction at regional scales in France by building a Boosted Regression Tree (BRT) model to obtain reliable predictions (R\(^2\))
and also applied the advanced deep learning modelling framework for the evaluation of in situ spectral measurement of SOC with in situ vis-NIR spectroscopy in southeastern Tibet (Chen et al., 2020) achieving \(R^2 = 0.92\). Rodríguez-Lado et al. (2015) used a dataset consisting of 115 topsoil observations in a catchment of approximately 100 km\(^2\) to map soil BD and compared three methods: Stepwise Multiple Linear Regression (MLR-S), Random Forest (RF) and Artificial Neural Networks (ANN). In this procedure, RF and ANN appeared the most suitable approaches to predict the measured data, producing \(R^2\) of 0.90 and 0.86, respectively. These results suggest that soil samples remain essential to obtain good estimates, and that PTFs derived from data collected in given locations can fail to give accurate estimates when applied elsewhere (Akpa et al., 2016). PTFs modelling is a relatively new subject and many important steps have been carried out recently (Chen et al., 2018; Sevastas et al., 2018). To extract all the contributions on soil BD, simple query can be used to gather publications from SCOPUS and Web of Knowledge (Schillaci et al., 2018). Out of this search the most used approach for the BD estimation with PTFs is multiple linear regression (60%) followed by ANN (20%), therefore these two approaches are investigated here.

At present the available PTF models offer wide predictive ranges and none are specifically developed for the Mediterranean area. The aim of this study was to develop new regionally-specific BD prediction models using data gathered from the literature on soil texture, SOC, topography and climate in Mediterranean agro-ecosystems. As well as providing a modelling framework that can be applied in each environmental setting. In the Mediterranean basin area, soil organic matter mineralization is boosted by high-temperature conditions (Álvaro-Fuentes and Paustian, 2011), in which rainfall has a peculiar pattern (availability during a short season vs long dry period). Moreover, the agricultural systems are conventionally plough-based
(Mazzoncini et al., 2011) causing soil compaction and reduced SOC stocks.

2. Material and Methods

The study was conceptualized during the first annual summer school module “Statistical Analysis of Spatial Data in Agro-Environmental Research”, organized in cooperation with Lake Como Advanced School (https://sdae.lakecomoschool.org/), and held from August 26-30, 2019. As a practical teaching activity, soil legacy data and topographic datasets were compiled to develop a PTF. The school participants were mainly PhD students and early career researchers. The present work was carried out after the school as a collaboration between students and teachers.

2.1 Operational procedures

Study work streams included PTF development using training datasets from public databases, and PTF validation using an independent validation dataset compiled from systematic review of the literature (Table 1 and Fig. 1). In the training step, we defined three PTFs – two based on statistical approaches and one based on ANN. In the validation step, we applied the three newly-defined PTFs to an external dataset. We then compared the performances of the three PTFs and benchmarked them against those of the MJ PTF and SoilGrids estimates (see below). The training and validation datasets were each split into topsoil and subsoil to infer separate PTFs.

**Table 1. Study overview and workflow to develop pedotransfer functions (PTF) to infer soil bulk density (BD) in top- (0-0.4m) and subsoil (0.4-1.2) of arable fields in the Mediterranean**

<table>
<thead>
<tr>
<th>Study stage</th>
<th>New PTFs to estimate BD</th>
<th>Reference PTF and BD data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLR-BS: backward and stepwise</td>
<td>SoilGrids: estimated BD values derived from WoSIS data</td>
</tr>
<tr>
<td></td>
<td>Regression</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>------------</td>
<td>-------</td>
</tr>
<tr>
<td><strong>Training</strong></td>
<td>Developed using WoSIS database* + topographic + climatic data</td>
<td></td>
</tr>
<tr>
<td><strong>Validation and Benchmarking</strong></td>
<td>Applied on external database** + topographic + climatic data</td>
<td>Applied on external database**</td>
</tr>
</tbody>
</table>

* WoSIS database: measured data of bulk density (BD, Mg m$^{-3}$), soil organic carbon (%), sand
** newly compiled database of soil bulk density, organic carbon, sand, silt and clay measurements of studies from the Mediterranean
Figure 1. Features of the datasets used to train and validate (training and model transferability) three new pedotransfer functions (PTF) for arable soils in the Mediterranean. For a description of the WorldClim Bioclimatic data, please see (Fick and Hijmans, 2017).

2.2 Data used in the training and validation stages
2.2.1. Soil datasets

*Training dataset used for PTFs model development:* The World Soil Information Service WoSIS (https://www.isric.org/explore/wosis) was used to retrieve soil textural values, SOC content and bulk density. WoSIS is a world scale database containing 196,000 geo-referenced, standardized soil profile entries for soil data from multiple origins. Approximately 40 different organizations around the world provide free access to the data via WoSIS and the Soil Profile (https://www.isric.org/explore/wosis/wosis-contributing-institutions-and-experts). More information on WoSIS inclusion criteria, quality assurance, and standardization procedures are available in Batjes et al. (2017). We note that for Europe, one of the main providers of WoSIS data is the Joint Research Center of the European Community, which has made available the entire collection of soil profiles included within the Soil Profile Analytical Database (SPADE-2) (de Souza et al., 2016; Hiederer et al., 2006; Panagos et al., 2013). Using ARCGIS, we selected all the profiles of the WoSIS database belonging to the Mediterranean basin (and defined surrounding areas) with geographic coordinates in metric resolution as well as attributes including sand, silt, clay, organic carbon and bulk density data in at least one soil horizon.

*External dataset used to test model transferability:* To assess the model transferability, validation of the three developed PTFs was required. Accordingly, we conducted a systematic literature analysis to collate information on soil textures, SOC, and BD from studies of field crops cultivated on mineral soils in Mediterranean basin and close surrounding areas. The search was carried out in SCOPUS and Web of Science (WoS). The selection criterion was the same as that applied during the extraction of the WoSIS data: required data were BD, SOC, texture and geolocalization. It is needed to remark that systematic queries did not result in a adequate number of suitable articles, so that we used different approaches such as searching for soil dataset within
To compare the performances of the PTF models developed in this study with well-known approaches, in the validation phase we applied the MJ PTF (1991; \(BD = 1.660 - 0.318 \cdot SOC^{0.5}\)) and we fitted the available SoilGrids BD values (Hengl et al., 2017) with the data of the external validation database constructed as above. SoilGrids is a system for digital soil mapping that uses machine learning methods to map the spatial distribution of soil properties across the globe using WoSIS data and environmental predictors.

For both training and validation datasets, the analysis focused on samples that alternatively fall within the 0-0.4 m layer (i.e. topsoil) or the 0.4-1.2 m layer (i.e. subsoil). Due to the presence of multiple genetic horizons inside the topsoil and subsoil, single observations which are part of the training dataset were not averaged. The soil sampling depth were considered as predictor. Furthermore, the inclusion of predictors such as soil properties (soil particle size fractions and SOC stock) allows to describe the soil sample at the given profile depth (e.g., SOC and clay content tend to decrease along the soil profile).

The data points used in the training phase which were derived from WoSIS were 129 and 180 for topsoil and subsoil, respectively.

As SoilGrids data are provided for six soil layers at the fixed depths (0-5, 5-15, 15-30, 30-60, 60-100, 100-200 cm), we computed a weighted average of SoilGrids BD for the comparison with the BD from the external dataset. For example, if BD is measured for the 10-25 layer then 5 cm belongs to the 5-15 cm SoilGrids layer and 10 cm to the 15-30 cm layer. Consequently, to obtain the sample value, we computed a weighted mean between the SoilGrids BD values given for the 5-15 and the 15-30 layers, using a weighting factor of 5 and 10 for the two layers, respectively.

We excluded the BD values lower than 0.9 Mg m\(^{-3}\) because they were not representative of agronomic journals.
mineral soils in semiarid regions and, when present, they were likely due to tillage operations occurred close to the sampling moment. We also excluded BD values greater than 2 Mg m\(^{-3}\) because they are not representative for agricultural land. Textural plots were prepared using the gtern software (Hamilton and Ferry, 2018).

2.2.2. Geodata

For the terrain analysis, the Shuttle radar topography mission SRTM 30 m DEM (Farr et al., 2007) was used to obtain topographic data with a resampling at 90 m. The digital elevation model was downloaded in ten tiles from the open topography website (https://opentopography.org/). The topographic indices were obtained for the whole study area using the geo-processing terrain analysis tool in SAGA (Conrad et al., 2015). Data pre-processing and maps were prepared using ArcGIS. The WorldClim climatic data (Fick and Hijmans, 2017) was used to obtain climate data (e.g., mean annual rainfall, average annual temperature). For EU countries, CORINE land cover (Bossard et al., 2000) was used to select agricultural land use. To assign the target land cover (Agriculture) CLC was checked for all the available periods, 2000, 2006, 2012, 2018. For non-EU countries – except for Turkey, which was included in CORINE land cover data – we selected soil profiles belonging to agricultural areas by observing satellite and aerial imagery available in ArcMap and Google Earth-Pro.

2.3. Study area

The study focused on the Mediterranean Basin, which covers the territory between 30\(^{\circ}\) and 45\(^{\circ}\) latitudes and, according to the Köppen climate classification system, belongs to the three main climate groups: \(B\) (dry), \(C\) (temperate), and \(D\) (continental) (Francaviglia et al., 2020) (see Fig. 2). The influence of the sea plays a key role in shaping the environment including relief characteristics, which determine the characteristic Mediterranean climate at basin scale (Lionello et al., 2006). Mediterranean soils are the result of a complex genesis (Lagacherie et al., 2018).
Carbonatic and limestone parent materials are the most widespread minerals in the Mediterranean (Verheye and De La Rosa, 2005; Zdruli et al., 2011). Long-term agricultural use has altered soil structure and degraded carbon content. Soil characteristics indicate different ages of soil development and depths and there is evidence of clay particle translocation within the soil profile (Zdruli et al., 2011).

According to the World Reference Base for Soil Resources (WRB 2006), approximately a dozen soil orders can be found in the Mediterranean basin: histosols, aluvialsols, leptosols, vertisols, fluvisols, gleysols, andosols, kastanozems and phaeozems, durisols, gypsisols, luvisols, arenosols, cambisols, and regosols. A brief description of these soil orders can be found in Zdruli et al., (2011). Figure 2 shows the locations of the sites included in the training (WoSIS data) and validation (extracted from the literature) datasets.

![Figure 2](image)

**Figure 2.** Study area. The location of the sampling sites (WoSIS data for training and external data for model transferability).

### 2.4 Development of PTFs to estimate BD

In this study, we evaluated three methods to estimate BD, namely Multiple Linear Regression
models each using two-variable selection criteria, and Artificial Neural Networks (ANN). These methods were chosen in our analyses because they are suitable when data are sparse and no spatial structure can be defined. A 10-fold cross-validation frame was used to assess the prediction accuracy (Veronesi and Schillaci, 2019). The three models were defined using a wide set of predictors, i.e. independent variables (soil properties, bioclimatic and topographic indicators). These predictors were derived from the soil and additional database (Fig. 1).

2.4.1 Multiple linear regression (MLR)
The first method (MLR-S) used was a stepwise multiple linear regression starting from no dependent variables (a constant-only model); the first dependent variable that will be included in the model is the variable that produces the maximum increase in $R^2$; if the increase in the explicated variance is significant (partial F test) at a given $P(F)$, called inclusion threshold, the variable is retained in the model (forward step). The same procedure is done to evaluate the possibility to include a second independent variable and so on. At each inclusion step, there is an exclusion step too, where, among the variables included in the model, the variable that is excluded causes the lower reduction in explicated variance. If the decrease of explained variability is not significant at a given $P(F)$, called exclusion threshold and higher than the inclusion threshold, the variable is excluded from the model. The process stops when no more dependent variables are included or excluded (Noryani et al., 2019). In the MLR-S, a predictor is included in the model if its regression coefficient is significant at $P \leq 0.05$ and excluded if the partial F test has a $P > 0.1$ (Draper and Smith, 1998).

The second approach was a stepwise variable selection, which started by including all independent variables, then excluded non-significant variables one by one using a backward stepwise approach (MLR-BS). Variables were excluded when their contribution did not affect the model explication capability (i.e., when the partial F test have a $P>0.1$) (Ghani and Ahmad,
For both methods (MLR-S and MLR-BS), the normality test of Kolmogorov-Smirnov and the Breush-Pagan test for the homogeneity of variances (Breusch et al., 1979) were applied to the residuals of the regression models.

2.4.2 Artificial Neural Network

An ANN is part of a computing system, which is developed to mimic the way the human brain processes information. ANN allows finding non-linear behavior of the system that cannot be discovered with traditional regression-based methods. To develop a PTF, the ANN is generally made by three layers of neurons, i.e. an input layer, a hidden layer and an output layer (Ebrahimi et al., 2019; Minasny and McBratney, 2002; Schaap et al., 1998). This kind of ANN architecture is known as Multi-Layer Perceptron (MLP). ANN imposes minimal requirements for model structure or assumptions because the shape of the relationship is determined during the learning process (Haykin, 2008). We used an MLP implementation in the IBM-SPSS 26.0.0.1. One hidden layer was used with three neurons according to the default settings, using the hyperbolic tangent activation function and the identity function for the output layer. This is an identity function because this task is a regression problem. The weighted connections feed forward from the input layer to the output layer. The training algorithm works by back-propagating the prediction error, through the parameters of the neural network. In this study, the MLP had 18 input predictors and one output variable, i.e. BD. The independent variables used as predictors in the three statistical models for the BD estimation are shown in Fig. 1. The optimal fit was reached in cross-validation by using 1 hidden layer, combined with three neurons. The Hyperparameters tuning was iteratively tested by applying an ANN with one hidden layer with 2 to 10 neurons and, alternatively, an ANN with two hidden layers with 2 to 5 neurons in the first hidden layer combined with 2 or 3 neurons in the second hidden layer. The use of one single
hidden layer resulted to be more effective. This result agreed with the automatic parameterization proposed by the software: (ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/27.0/en/client/Manuals/IBM_SPSS_Statistics_Algorithms.pdf). Regarding computation time, the model training phase takes few seconds.

2.5. Analysis of models’ performance
The following evaluation indices were calculated to test the model performance in estimating BD: i) R² coefficient of determination of the scatter plot of the predicted against the observed values;
ii) Bias and %Bias (Addiscott and Whitmore, 1987), optimal value is 0, range is from +∞ to -∞; when the Bias% is < 10% it may be considered very favorable (Moriasi et al., 2007);
iii) Root Mean Square Error (RMSE) and %RMSE (RMSE/(Observed Mean) *100) (Fox, 1981), optimal value is 0, range is from 0 to +∞; %RMSE value lower than 10% is considered to be favorable (Bellocchi et al., 2002);
iv) The Pearson correlation coefficient, optimal value is 1, range is from +1 to −1;
v) The slope of the regression of observed data to the estimated ones, optimal value is 1, range is from +∞ to -∞ (Piñeiro et al., 2008).

Note that Bias is always equal to 0 when the ordinary least square (OLS) method is applied, which was the case in the two regression training sets. Moreover, in OLS analysis the slope of observed values against the estimated values is equal to 1. All indices were computed using Irene-DLL (Fila et al., 2003).

3. Results

3.1 Descriptive statistics

3.1.1 Soil properties
The highest average BD value was observed in the subsoil training dataset (1.51±0.17 Mg m\(^{-3}\)). The lowest average BD value was observed in the topsoil validation dataset (1.38±0.12 Mg m\(^{-3}\)). The SOC was higher in the topsoil testing (1.28±1 %), and lower in the subsoil testing dataset (0.61±0.35 %) (Table 3). The most variable soil property was the sand content with a coefficient of variation ranging from 48 to 85%, while BD was less variable with a coefficient of variation ranging from 9 to 14%. The references of the independent dataset for validation are listed in Table 2. The independent external dataset for validation comprised 59 observations for the topsoil and 29 for the subsoil, Table 3. Textural plots of the training and validation datasets are shown in Table 3.

**Table 2. Independent dataset for validation with country, climate and reference.**

<table>
<thead>
<tr>
<th>Country</th>
<th>Köppen climate classification</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algeria</td>
<td>BSk</td>
<td>Chennafi et al., 2006</td>
</tr>
<tr>
<td>Croatia</td>
<td>Cfa</td>
<td>Bogunovic et al., 2018</td>
</tr>
<tr>
<td>Egypt</td>
<td>BWh</td>
<td>Mahmoud et al., 2019</td>
</tr>
<tr>
<td></td>
<td>BWh</td>
<td>Salem et al., 2015</td>
</tr>
<tr>
<td>France</td>
<td>Csa</td>
<td>Cardinael et al., 2017</td>
</tr>
<tr>
<td>Greece</td>
<td>Csa</td>
<td>Antonopoulous et al., 2013</td>
</tr>
<tr>
<td>Israel</td>
<td>BSh</td>
<td>Stavi et al., 2008</td>
</tr>
<tr>
<td>Italy</td>
<td>Cfa</td>
<td>Pezzuolo et al., 2017</td>
</tr>
<tr>
<td></td>
<td>Csa</td>
<td>Carozzi et al., 2013</td>
</tr>
<tr>
<td></td>
<td>Cfa</td>
<td>Francaviglia et al., 2015</td>
</tr>
<tr>
<td></td>
<td>Cfa</td>
<td>Valboa et al., 2015</td>
</tr>
<tr>
<td></td>
<td>Cfa</td>
<td>Perego et al., 2019</td>
</tr>
<tr>
<td></td>
<td>Cfa</td>
<td>Ceotto et al., 2018</td>
</tr>
<tr>
<td></td>
<td>Cfa</td>
<td>Diacono et al., 2018</td>
</tr>
<tr>
<td></td>
<td>Csa</td>
<td>Vitale et al., 2017</td>
</tr>
<tr>
<td>Lebanon</td>
<td>Csa</td>
<td>Karam et al., 2007</td>
</tr>
<tr>
<td>Morocco</td>
<td>Csa</td>
<td>Ichir et al., 2003</td>
</tr>
<tr>
<td>Spain</td>
<td>BSk</td>
<td>Pareja-Sánchez et al., 2017</td>
</tr>
<tr>
<td></td>
<td>Cfa</td>
<td>Bescansa et al., 2006</td>
</tr>
<tr>
<td></td>
<td>BSk</td>
<td>Pardo et al., 2020</td>
</tr>
<tr>
<td></td>
<td>BSk</td>
<td>Tolon-Becerra et al., 2011</td>
</tr>
</tbody>
</table>
Table 3. Soil properties of the training and testing data for topsoil (0-0.4m) and subsoil (0.4-1.2 m): Bulk Density (BD), Soil Organic Carbon (SOC), Fine earth fractions.

<table>
<thead>
<tr>
<th>Region</th>
<th>Soil Type</th>
<th>BD (Mg m$^{-3}$)</th>
<th>SOC (%)</th>
<th>Sand (%)</th>
<th>Silt (%)</th>
<th>Clay (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topsoil Training</td>
<td>(N=129)</td>
<td>mean 1.44</td>
<td>1.26</td>
<td>24.4</td>
<td>36.5</td>
<td>39.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stdv 0.20</td>
<td>0.64</td>
<td>17.1</td>
<td>14.1</td>
<td>18.1</td>
</tr>
<tr>
<td>Topsoil Testing</td>
<td>(N=59)</td>
<td>mean 1.41</td>
<td>1.28</td>
<td>31.39</td>
<td>40.68</td>
<td>28.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stdv 0.11</td>
<td>1.0</td>
<td>13.82</td>
<td>9.09</td>
<td>14.25</td>
</tr>
<tr>
<td>Subsoil Training</td>
<td>(N=180)</td>
<td>mean 1.51</td>
<td>1.15</td>
<td>20.1</td>
<td>38.2</td>
<td>41.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stdv 0.17</td>
<td>0.67</td>
<td>17.0</td>
<td>15.7</td>
<td>17.4</td>
</tr>
<tr>
<td>Subsoil Testing</td>
<td>(N=29)</td>
<td>mean 1.48</td>
<td>0.61</td>
<td>29.04</td>
<td>39.49</td>
<td>31.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stdv 0.16</td>
<td>0.35</td>
<td>19.38</td>
<td>12.58</td>
<td>18.4</td>
</tr>
</tbody>
</table>

Figure 3. Textural plots, a) topsoil validation dataset, b) subsoil validation dataset, c) topsoil test dataset d) subsoil test dataset.
3.1.2 Environmental variables
Average precipitation reported in the training dataset was highly variable in the study area with a minimum value of 426 and a maximum of 1693 mm yr$^{-1}$. The validation dataset showed a minimum annual rainfall of 189 and a maximum of 1155 mm yr$^{-1}$. Mean annual temperature, Elevation (m), Slope (%) are reported in Table 4.

Table 4. Descriptive statistics of the selected environmental variables.

<table>
<thead>
<tr>
<th></th>
<th>Annual Precipitation (mm yr$^{-1}$)</th>
<th>Mean annual temperature (° C)</th>
<th>Elevation (m)</th>
<th>Slope (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N=77 sites)</td>
<td>mean</td>
<td>774.4</td>
<td>10.5</td>
<td>321</td>
</tr>
<tr>
<td></td>
<td>std</td>
<td>294.4</td>
<td>1.5</td>
<td>332</td>
</tr>
<tr>
<td></td>
<td><strong>Testing</strong></td>
<td>495.7</td>
<td>16</td>
<td>318</td>
</tr>
<tr>
<td>(N=36 sites)</td>
<td>mean</td>
<td>300</td>
<td>3.1</td>
<td>379</td>
</tr>
<tr>
<td></td>
<td>std</td>
<td>300</td>
<td>3.1</td>
<td>379</td>
</tr>
</tbody>
</table>

3.2 Model performance and transferability
Homogeneity of variance and normality tests for the MLR models were conducted using the Breush-Pagan test and Kolmogorov-Smirnov test (Table 5).

Table 5 Homogeneity of variance and normality tests for Multiple Linear Regression (MLR) models.

<table>
<thead>
<tr>
<th></th>
<th>MLR-S</th>
<th>MLR-BS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Topsoil</td>
<td>Subsoil</td>
</tr>
</tbody>
</table>

Homogeneity of variance of residuals* | 0.056 | 0.051 | 0.065 | 0.051
Normality of residuals** | >0.2 | >0.2 | >0.2 | >0.2

*Breush-Pagan test; ** Kolmogorov-Smirnov test

Topsoil model metrics are shown in Table 6. The RMSE of the topsoil training dataset (Table 6a) ranged from 0.07 (ANN) to 0.17 (MLR-S), and similar performances were obtained with the MLR-BS models. The Bias of the ANN was close to zero. The ANN model showed the highest $R^2$ (0.89), whereas the MLR-S model showed the lowest $R^2$ (0.24).

Table 6. Performance of the newly developed pedotransfer function (PTF) as developed with the topsoil training and cross validation (a) and tested with the independent external datasets for model transferability (b). Indices values reported in brackets refer to the cross-validation results.

<table>
<thead>
<tr>
<th>a)</th>
<th>MLR-S - MLR-S</th>
<th>MLR-BS - MLR-BS</th>
<th>ANN - ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>0.17 (0.16)</td>
<td>0.14 (0.15)</td>
<td>0.07 (0.16)</td>
</tr>
<tr>
<td>rRMSE %</td>
<td>11.91 (11.53)</td>
<td>8.88 (10.81)</td>
<td>4.56 (11.4)</td>
</tr>
<tr>
<td>Bias</td>
<td>(0.0007)</td>
<td>(0.0047)</td>
<td>0.00 (0.01)</td>
</tr>
<tr>
<td>Bias %</td>
<td>(0.144)</td>
<td>(0.37)</td>
<td>0.10 (1.11)</td>
</tr>
<tr>
<td>$r$</td>
<td>0.51 (0.49)</td>
<td>0.72 (0.57)</td>
<td>0.94 (0.67)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.26 (0.31)</td>
<td>0.51 (0.37)</td>
<td>0.89 (0.48)</td>
</tr>
<tr>
<td>Slope b</td>
<td>(0.84)</td>
<td>(0.71)</td>
<td>1.00 (0.78)</td>
</tr>
<tr>
<td>Estimated Max</td>
<td>1.61 (1.6)</td>
<td>1.80 (1.67)</td>
<td>1.90 (1.75)</td>
</tr>
<tr>
<td>Estimated Min</td>
<td>1.16 (1.22)</td>
<td>1.02 (1.14)</td>
<td>0.88 (1.13)</td>
</tr>
<tr>
<td>N</td>
<td>129</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>b)</th>
<th>MLR-S</th>
<th>MLR-BS</th>
<th>ANN</th>
<th>MJ</th>
<th>SoilGrids</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>0.14</td>
<td>0.32</td>
<td>0.16</td>
<td>0.17</td>
<td>0.13</td>
</tr>
</tbody>
</table>
The RMSE in the topsoil validation dataset (Table 6b) range from 0.13 (SoilGrids) to 0.32 (MLR-BS). All the Bias values were ≤0.5. The $R^2$ ranged between 0.09 to 0.41, in SoilGrids and ANN, respectively.

**Table 7.** Performance of the newly developed pedotransfer function (PTF) as developed with the subsoil training and cross validation (a) and tested with the independent external datasets for model transferability (b). Indices values reported in brackets refer to the cross-validation results.

<table>
<thead>
<tr>
<th></th>
<th>MLR-S</th>
<th>MLR-BS</th>
<th>ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>0.14 (0.13)</td>
<td>0.12 (0.13)</td>
<td>0.07 (0.11)</td>
</tr>
<tr>
<td>rRMSE %</td>
<td>9.04 (9.24)</td>
<td>8.04 (8.67)</td>
<td>4.53 (7.79)</td>
</tr>
<tr>
<td>Bias</td>
<td>-0.0003</td>
<td>-0.0004</td>
<td>0.00 (0.003)</td>
</tr>
<tr>
<td>Bias %</td>
<td>0.0056</td>
<td>0.0008</td>
<td>-0.16 (0.21)</td>
</tr>
<tr>
<td>r</td>
<td>0.49 (0.47)</td>
<td>0.70 (0.58)</td>
<td>0.92 (0.67)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.24 (0.21)</td>
<td>0.48 (0.38)</td>
<td>0.84 (0.48)</td>
</tr>
<tr>
<td>Slope b</td>
<td>0.90</td>
<td>0.90</td>
<td>0.98 (0.84)</td>
</tr>
<tr>
<td>Estimated Max</td>
<td>1.77 (1.68)</td>
<td>1.79 (1.71)</td>
<td>1.93 (1.76)</td>
</tr>
<tr>
<td>Estimated Min</td>
<td>1.12 (1.32)</td>
<td>1.35 (1.28)</td>
<td>1.10 (1.26)</td>
</tr>
</tbody>
</table>
Subsoil model metrics are shown in Table 7. The RMSE in the subsoil training dataset (Table 7a) ranged from 0.07 (ANN) to 0.14 (MLR-S). The Bias of the ANN was close to zero. The ANN model showed the highest $R^2$ of 0.84, whereas the MLR-S model showed the lowest $R^2$ of 0.24. The RMSE in the subsoil external dataset (Table 7b) are very similar and ranged from 0.14 to 0.39. The Bias % values ranging from -4.6 (MJ) to 2.8% (MLR-S). The $R^2$ ranged between 0.07 (MLR-S) to 0.45 (ANN), respectively.

Since the best performance was achieved with the ANN, we provide a .xlm spreadsheet file that can be used to execute the PTF developed with the ANN using the soil data, topography and WorldClim. Furthermore, to allow users to apply the PTF based on the ANN in different statistical packages a Predictive Model Markup Language file (PMML), which is an XML-based predictive model interchange format, is available in the supplemental materials.
Figure 4. Predicted vs observed data (training topsoil and subsoil A and B, validation topsoil and subsoil C and D), MLR-S model, MLR-BS model, ANN neural network model, MJ PTF SoilGrids.

3.3 Variable importance

Table 8 shows the absolute standardized regression coefficient for each MLR model, considering 100% the highest beta value, to obtain a comparable result to the ANN model. Clay was the most important predictor in the topsoil MLR-S model. In the topsoil, SOC contributed approximately 25% of BD in the MLR-BS PTF, but it was not present in the MLR-S models. Similarly, Clay$^2$ was not present in the MLR-S models, slope and SOC$^2$ were the most important predictors in the
subsoil using MLR-BS. Bioclimatic predictors such as BIO1 (Annual Mean Temperature), BIO2 (Mean Diurnal) and BIO7 (Annual T°C Range) were the most influential predictors in both topsoil and subsoil using MLR models. In topsoils, predictors included BIO7 (Annual T°C Range) and BIO14 (Precipitation of the Driest Month), heavily contributed to BD estimates within the MLR-BS and MLR-S of BD. BIO7 (Annual T°C Range) was more important than BIO14 (Precipitation of the Driest Month) in any model. Among the topographic predictors, the elevation was important in subsoil MLR-S models (contributing 24%), whereas it was not important in the topsoil or subsoil MLR-BS models. In subsoils, BIO3 (Isothermality) contributed 8% and 7% of subsoil BD in MLR-S and MLR-BS.

Table 8 Normalized variable importance in the MLR-S and MLR-BS (standardized regression coefficient in %). Conditional formatting is applied, Red color marks the minimum, green color the maximum and the yellow marks the middle values.

<table>
<thead>
<tr>
<th></th>
<th>MLR-S TOP</th>
<th>MLR-BS TOP</th>
<th>MLR-S SUB</th>
<th>MLR-BS SUB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clay</td>
<td></td>
<td></td>
<td>5.83</td>
<td>4.62</td>
</tr>
<tr>
<td>Sand</td>
<td>5.83</td>
<td>14.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Silt</td>
<td>2.97</td>
<td></td>
<td>4.69</td>
<td></td>
</tr>
<tr>
<td>SOC</td>
<td>19.15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MeanDepth</td>
<td></td>
<td></td>
<td>11.73</td>
<td></td>
</tr>
<tr>
<td>Elevation</td>
<td>2.34</td>
<td></td>
<td>4.53</td>
<td></td>
</tr>
<tr>
<td>Slope</td>
<td>24.97</td>
<td></td>
<td>4.35</td>
<td></td>
</tr>
<tr>
<td>Profile Curvature</td>
<td>13.91</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIO1</td>
<td>5.70</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIO2</td>
<td>5.91</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIO3</td>
<td>6.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIO4</td>
<td>36.46</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIO5</td>
<td>4.59</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIO7</td>
<td>14.75</td>
<td></td>
<td>14.75</td>
<td></td>
</tr>
<tr>
<td>BIO12</td>
<td></td>
<td>5.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIO13</td>
<td></td>
<td>6.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIO14</td>
<td>5.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIO15</td>
<td>5.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIO17</td>
<td>4.85</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIO18</td>
<td>4.74</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIO19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 9 shows the importance of the predictors included in the ANN models, based on sensitivity analyses using the default option in the MLP tool in IBM SPSS (independent variable importance analysis). The models included all the predictors except interactions between soil properties which were calculated within the ANN procedure but hidden to the user. Among the main physical and chemical properties, Sand was the most important in topsoil (5.6) and SOC in subsoil (5.5).

Within the ANN, the most important predictors were BIO7 (Annual T°C Range) and Profile Curvature for topsoil and subsoil, respectively. Other important predictors were MeanDepth and the BIO12 (Annual Rainfall) in topsoil, SOC, BIO1 (Annual Mean Temperature) and BIO16 (Precipitation of the Wettest Quarter) were important in topsoil and subsoil. Soil properties predicted the BD of subsoil. Clay predicted BD in subsoil (3.2) and topsoil (2.8). Among the topographic predictors, Profile and plan curvature played a stronger role predicting the BD of the topsoil (5.3) compared to the subsoil (5.6).

Table 9. Normalized variable* importance for predicting the bulk density of the top- and subsoil by means of the pedotransfer function developed by the Artificial Neural Network optimization approach. *variable description is available in figure 1, Conditional formatting is applied, Red color marks the minimum, green color the maximum and the yellow marks the middle values.

<table>
<thead>
<tr>
<th></th>
<th>Topsoil</th>
<th>Subsoil</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clay</td>
<td>3.2</td>
<td>2.8</td>
</tr>
<tr>
<td>Sand</td>
<td>5.6</td>
<td>1.2</td>
</tr>
<tr>
<td>Silt</td>
<td>3.4</td>
<td>3.7</td>
</tr>
<tr>
<td>SOC</td>
<td>3.8</td>
<td>5.5</td>
</tr>
<tr>
<td>MeanDepth</td>
<td>1.4</td>
<td>1.5</td>
</tr>
</tbody>
</table>
Elevation | 3.8 | 2.7
Slope | 2.6 | 5.3
SIN Aspect | 3.3 | 1.7
Profile Curvature | 5.3 | 4.5
Plan Curvature | 3.8 | 5.6
BIO1 | 2.5 | 2.8
BIO2 | 5.3 | 5.0
BIO3 | 2.0 | 2.7
BIO4 | 3.7 | 4.4
BIO5 | 4.7 | 1.6
BIO6 | 3.4 | 1.8
BIO7 | 3.9 | 5.4
BIO8 | 3.5 | 4.0
BIO9 | 3.8 | 2.9
BIO10 | 1.6 | 2.6
BIO11 | 2.9 | 1.8
BIO12 | 1.4 | 2.6
BIO13 | 3.0 | 2.3
BIO14 | 3.2 | 3.2
BIO15 | 5.6 | 2.3
BIO16 | 3.3 | 1.8
BIO17 | 3.9 | 3.5
BIO18 | 4.4 | 3.3
BIO19 | 1.7 | 2.7

4 Discussion

Collaborative work among researchers from different branches of geosciences facilitated a systematic literature review and compilation of an up-to-date, regionally-relevant geo-dataset; and this permitted the development and validation of new pedotransfer functions (PTF) to predict the BD of top- and subsoil in the Mediterranean.

4.1 Performance of pedotransfer functions

In this study, the performances of three new PTFs to estimate BD (MLR-S, MLR-BS and ANN) were defined used WoSIS soil data in combination with environmental data. The model transferability of the new PTFs was carried out using external dataset from Mediterranean
locations derived from literature. Results were also benchmarked against the widely used MJ-PTF, which uses only soil organic carbon to predict BD, and the global SoilGrids data which are based on topographic and remote-sensed estimates of BD. Among the MLR approaches, MLR-BS performed slightly better than MLR-S. The ANN model outperformed the MLR models.

Our PTF development strategy, which implies the use of topographic and climatic variables along with soil properties, agreed with the approach of Wang et al., (2014) and Akpa et al., (2016). However, we also validated our new PTFs by estimating the BD in top- and subsoil using an external and independent dataset. This resulted in less accurate predictions than those made by the training datasets as already remarked by (Khaledian and Miller, 2020; Morin and Davis, 2017; Thompson, 2006). Nevertheless, these authors suggested that the use of external dataset rather than internal validation methods provides direct evidence about whether study results will replicate (Thompson, 2006).

Here we initially used MLR because it is a hands-on tool that provides direct quantitative and easily interpretable results. By contrast, the ANN has provided an alternative machine learning approach used in relatively recent analyses (Alvarez-Acosta et al., 2012; Ballabio et al., 2016; Chen et al., 2018; Ghehi et al., 2012; Nussbaum et al., 2018). In this study, the MJ PTF was used as a simple comparator because it is independent of other physical soil parameters except SOC and is the most widely used PTF. In our MLR-BS topsoil, MLR-S subsoil, and MLR-BS subsoil models, we considered inclusion of the key determinant used (i.e., SOC square root), but it was not included in the final model because it did not significantly improve the predictions. Our MLR-S and MLR-BS performed better than the MJ because our dataset included additional factors that directly determine BD over the long-term (such as those related to the climate or topography), thus raising the prediction capability. the SoilGrids database yielded a lower
prediction ability in comparison to the other models. SoilGrids is considered as an interesting solution because it is a gridded multiple depth dataset at a 250m spatial resolution and it is available worldwide. However, the present results suggest that SoilGrids BD estimations may not adequately match the observed BD values (i.e., external dataset) which were measured in specific sites located in the Mediterranean area.

4.2 Data groupings and reliability of PTFs

The fit of the MLR subsoil models was more satisfactory than for subsoil linear model MLR-S with an $R^2$ 0.12. Subsoil models were also more satisfactory for the ANN ($R^2 = 0.45$). This is in contrast to previous publications in which grouping input data by soil depths did not improve the prediction of BD in tropical soils (De Vos et al., 2005), which might have been attributable to different level of disturbance of the soil in the study areas (Hollis et al., 2012), or differences in the additional factors analyzed.

Arable soils undergo significant changes over time due to tillage and cultivation. Therefore, physical soil properties such as BD are more stable in the subsoil than in the topsoil. Statistics for soil texture and SOC agreed with data reported for other Mediterranean countries (Çelik et al., 2019; Evrendilek et al., 2004). The MJ and SoilGrids models yielded a similar result (Table 6 and 7). Notably, MLR-BS showed an $R^2$ close to zero for the validation datasets. This hampers discussion of model variability comparing the training and validation datasets. Generally, negative Bias is observed in the subsoil external dataset. As for the external dataset, slightly positive Bias indicates that MLR-S, ANN, SoilGrids overestimate the average BD of 2.8%, 1.7 % and 0.7%, respectively; MJ and MLR-BS underestimate the BD of -4.6% and -2.3% which is not preferable especially for the subsoil, an exception has been encountered with the topsoil MLR-BS that has predicted in few cases values very far from the true BD. Mediterranean soils
are diverse, their hydraulic properties reflect pedogenetic factors as well as recent changes in management and climate (Yaalon, 1997).

4.3 Importance of predictor variables

Previous attempts to estimate soil BD by PTF (Çelik et al., 2019; Gozubuyuk et al., 2014; Tranter et al., 2007) did not include climate parameters because they were often not readily available or not immediately obvious as a determinant of BD. Many factors related to climate, such as bioclimatic indices, affect BD (e.g., rainfall intensity or pattern, high soil temperature in summer) (Basile et al., 2019; Chen et al., 2018). Bioclimatic indices and topographic predictors contributed greatly to the performance of the MLR and comprised 100% of the variables in the MLR-S for the topsoil, and about the 33% in the MLR-BS. The regression models (MLR-S and MLR-BS) included soil textural data (MLR-S topsoil) or SOC related (MLR-S subsoil and MLR-BS topsoil and subsoil). Our results showed that important predictors of BD in the MLR models were slope, clay, SOC$^2$, and bioclimatic variables such as BIO1 (Annual Mean Temperature), BIO2 (Mean Diurnal Range) and BIO7 (Annual T°C Range). This is consistent with previous reports (Akpa et al., 2016). In fact, part of the BD variability is due to the diverse bioclimatic zones within the Mediterranean Basin (Beck et al., 2018).

In our study, the inclusion of the climatic and topographic data increased the model reliability. Indeed, models without topographic and climatic predictors had very low performance at the training stage (data not shown). However, the lack of field management information, which strongly affects arable soils (e.g., crop type, tillage methods, irrigation, input of organic matter), hampers the ability to infer a relationship with factors of soil formation and processes (Wadoux et al., 2019) which would potentially improve the model prediction. In the Mediterranean Basin, significant effects of cropping systems and field managements on BD have been demonstrated in
field studies (Álvaro-Fuentes et al., 2008; Bogunovic et al., 2020; Çelik et al., 2019; Perego et al., 2019; Pezzuolo et al., 2017).

5. Conclusions

Arable soils are widely distributed and the estimation of their fertility and carbon sequestration ability is a prerequisite for their management at wide scale. Reliable PTF to estimate BD are thus a needed instrument for arable soils management at the regional or higher levels. In the present study, we developed a robust PTF for BD estimation by exploiting the WoSIS resource, and it was the first time that such a broad set of data are valorized for PTF development. Moreover, we considered relevant predictors such as climatic and topographic parameters, which are fully and freely available and responsible for remarkably improving the predictive capability of the PTF models.

One of the three developed PTF (i.e., ANN) showed a better capability of estimating BD data than the well-known function Manrique Jones and the SoilGrids estimation approach; this outcome proved that the work hypothesis was correct and then developing the PTF with climate-specific set of data and adding topographic and climate predictors leads to a better predictive capability.

A relevant result of the present work is a ready to be used PTF model (i.e., ANN) for to separate soil layers (i.e., topsoil and subsoil) for the arable soils in the Mediterranean basin. The potential users of this result are public authorities interested in estimating soil carbon stock by exploiting legacy soil data in which bulk density is an often-missing parameter in the large monitoring campaigns. Researchers can be also interested in a more robust method of BD estimation when elaborating sets of soil data, especially when the aim is to estimate spatial and temporal variation.
The robustness of the ANN PTF is ensured by the use of an independent external dataset compiled from the literature for the validation of the PTF models transferability.

Results from the present work provide a reproducible and externally tested tool that can be applied to obtain a BD estimation at a regional level more reliable than the presently used PTF or gridded benchmarks. Thus, the present results are an option for policy making and management at a regional level.

6. Acknowledgments

All authors acknowledge the funding for the Lake Como Advanced School provided by the grant for the SDAE Summer School 2019 from the Italian Society of Agronomy, University of Milan and University of Pavia. WLC. Landsupport H2020, grant number ID: 774234. Project “Carbon Market - Innovative cropping systems for carbon market” funded by Natural Resources Institute Finland (Luke). We express our gratitude to Cami Moss for editing the manuscript.

Views, thoughts, and opinions expressed in the text belong solely to the author and do not reflect the opinion of the author's employer, organization, committee members, or that of other course participants.

This paper is dedicated to the dear memory of Prof: Dario Sacco, who was a teacher of the SDAE School.

References


https://doi.org/10.1007/978-1-4612-3532-3_4


https://doi.org/10.1016/j.geoderma.2017.10.009


https://doi.org/10.1016/j.scitotenv.2018.01.129

environmental covariates: Rio Doce basin. Sci. Agric. 73, 525–534.


https://doi.org/10.2134/agronj2003.1330


https://doi.org/10.1016/j.scitotenv.2017.05.239


https://doi.org/10.1016/j.geoderma.2016.10.019


https://doi.org/10.1016/j.geoderma.2013.07.013


Mediterranean Soils.


Pezzuolo, A., Dumont, B., Sartori, L., Marinello, F., De Antoni Migliorati, M., Basso, B., 2017. Evaluating the impact of soil conservation measures on soil organic carbon at the farm...


Declaration of competing interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:
Credit Author Statement
Calogero Schillaci, Alessia Perego, Marco Acutis  Conceptualization;
Calogero Schillaci, Alessia Perego, Marco Acutis, Tommaso Tadiello, Luigi Lombardo, Sergio Saia, Elena Pareja-Serrano  Data curation;
Formal analysis; Calogero Schillaci, Alessia Perego, Marco Acutis
Funding acquisition;
Calogero Schillaci, Alessia Perego, Elena Valkama, Michael Märker, Sergio Saia, Fabio Veronesi, Aldo Lipani, Luigi Lombardo, Tommaso Tadiello, Hannes A. Gamper, Luigi Tedone, Cami Moss, Elena Pareja-Serrano, Gabriele Amato, Kersten Kühl, Claudia Damatirca, Alessia Cogato, Nada Mzid, Rasu Eeswaran, Marya Rebolo, Giorgio Sperandio, Alberto Bosino, Margherita Bufalini, Tülay Tunçay, Jianqi Ding, Marco Fiorentini, Guadalupe Tiscornia, Sarah Conradt, Marco Acutis, Investigation;
Calogero Schillaci, Alessia Perego, Marco Acutis Methodology;
Maro Acutis Project administration;
Calogero Schillaci, Alessia Perego, Elena Valkama, Michael Märker, Sergio Saia, Fabio Veronesi, Aldo Lipani, Resources;
Calogero Schillaci, Alessia Perego, Marco Acutis, Luigi Lombardo, Tommaso Tadiello, Sergio Saia, Marco Botta Software;
Elena Valkama, Michael Märker, Marco Acutis Supervision;
Calogero Schillaci, Alessia Perego, Marco Acutis Validation;
Calogero Schillaci, Alessia Perego, Marco Acutis, Luigi Lombardo, Tommaso Tadiello, Sergio Saia, Marco Botta Visualization;
Calogero Schillaci, Alessia Perego, Marco Acutis, Roles/Writing - original draft;
Calogero Schillaci, Alessia Perego, Elena Valkama, Michael Märker, Sergio Saia, Fabio Veronesi, Aldo Lipani, Luigi Lombardo, Tommaso Tadiello, Hannes A. Gamper, Luigi Tedone, Cami Moss, Elena Pareja-Serrano, Gabriele Amato, Kersten Kühl, Claudia Damatirca, Alessia Cogato, Nada Mzid, Rasu Eeswaran, Marya Rebolo, Giorgio Sperandio, Alberto Bosino, Margherita Bufalini, Tülay Tunçay, Jianqi Ding, Marco Fiorentini, Guadalupe Tiscornia, Sarah Writing - review & editing.
Graphical abstract

New pedotransfer approaches to predict soil BD using WoSIS soil data and environmental covariates in Mediterranean agro-ecosystems

Training dataset for the development of three new PTFs: MLR-S, MLR-BS, ANN

Source: WoSIS open database of georeferenced soil profiles sampling with soil properties: Bulk density, Texture, SOC, Rock fragment content, soil depth

Additional Topographic and Climatic data

Model transferability and comparison with benchmarks Mansirique and Jones and SoilGrids estimations

Source: Soil properties of studies carried out in the Mediterranean: Bulk density, Texture, SOC, Rock fragment content, soil depth

Additional Topographic and Climatic data

- ANN-PTF had a good performance in the training and model transferability steps.
- ANN-PTF outperformed the Mansirique and Jones PTF and SoilGrids estimations in predicting BD.
Highlights
- Three PTFs were developed to calculate bulk density of arable top- and subsoil
- WoSIS, WorldClim, and topographic data of the Mediterranean Basin were used
- Model transferability of the three new PTFs was validated with external dataset
- Topsoil ANN-PTF had $R^2$ of 0.89 in training and 0.45 in model transferability
- ANN-PTF outperformed the commonly employed PTF by Manrique and Jones