Catena 158 (2017) 390–412

Contents lists available at ScienceDirect

Catena

journal homepage: www.elsevier.com/locate/catena

The INFOSOLO database as a first step towards the development of a soil information system in Portugal

Tiago B. Ramos[a,*], Ana Horta[b], Maria C. Gonçalves[c], Fernando P. Pires[c], Deanna Duffy[d], José C. Martins[c]

[a] MARETEC, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal
[b] Institute of Land, Water and Society, Charles Sturt University, Albury-Wodonga, Australia
[c] INIAV, Instituto Nacional de Investigação Agrária e Veterinária, Oeiras, Portugal
[d] Spatial Data Analysis Network, Charles Sturt University, Albury-Wodonga, Australia

ARTICLE INFO

Keywords:
Digital soil mapping
Soil database
Organic carbon content
pH
Cation exchange capacity

ABSTRACT

Nowadays there is an increasing effort for raising global awareness for the importance of soils to ensure food security, to improve agricultural and environmental planning and monitoring, and to establish effective and sustainable land management policies to counteract soil degradation. This study presents the INFOSOLO legacy database as the first effort to develop a soil information system in Portugal, suitable to compile soil data produced in the country, and to support stakeholders and land managers in decision-making. The database currently includes soil data from a set of 9934 horizons/layers studied in 3461 soil profiles across the country between 1966 and 2014. Data was extracted from scattered soil surveys, research projects, and academic studies carried out by public Portuguese and other European institutions, with a series of validation tests and harmonization procedures being implemented in order to access and improve the quality of the data. The importance of the INFOSOLO legacy dataset was discussed and exemplified with the analysis of the spatial and temporal distribution of selected soil properties, namely the organic carbon content, pH, and cation exchange capacity. For these properties, 1 km grid maps were also developed for the topsoil horizons/layers in Portugal using different spatial modelling approaches. To highlight the importance of using INFOSOLO for soil characterization, the EU-wide soil database LUCAS was used to compare both datasets in terms of data distributions, spatial continuity and accuracy of modelling outputs. The comparison also included the digital soil maps provided by the SoilGrids product for Portugal. The comparison results highlighted specific areas in the country for which INFOSOLO is capable to deliver accurate but also more reliable predictions when compared with the LUCAS and SoilGrids results. Thus, INFOSOLO provides the basis for improving soil information in the country and for raising national awareness of the importance of soil resources to the country's development.

1. Introduction

Soils serve as growing medium for feed, food, fibre, and fuel; act as a filter and reservoir for water and nutrients; contribute to climate regulation by acting as a pool for carbon and greenhouse gases (N_20 and CH_4); provide habitat for billions of organisms, contributing to biodiversity; act as a source of raw materials like sand, clay, and wood; support plants, animals, and infrastructures; and act as an aesthetic and cultural resource (Blum, 2005; Dominati et al., 2010; Robinson et al., 2012; Hartemink, 2015). Despite all vital services provided to society, soil degradation processes, which include erosion, organic matter decline, compaction, salinization, landslides, contamination, sealing, and biodiversity decline, remain a critical problem in many regions in the world (Eswaran et al., 2001; Reich et al., 2001; Montanarella, 2007) since stakeholders, policy makers and society at large still fail to perceive the intrinsic relations between soil health and sustainability.

Shifting such paradigm can only be accomplished by raising global awareness of the importance of soils for food security and essential ecosystem functions, and by establishing sound and sustainable land management policies to counteract soil degradation. Global initiatives like the Global Soil Partnership (www.fao.org/global-soil-partnership/en/) can definitely help promoting the importance of soils to human welfare. Also, initiatives such as the soil-net.com website (www.soil-net.com/) from the Cranfield University, UK, and the Soils Challenge Badge from the FAO (2015), where soil services are introduced to kindergarten, primary and high school students can be of extreme

* Corresponding author at: MARETEC, Instituto Superior Técnico, Av. Rovisco Pais, 1, 1049-001 Lisboa, Portugal.
E-mail addresses: tiago_ramos@netcabo.pt, tiagobramos@tecnico.ulisboa.pt (T.B. Ramos).

http://dx.doi.org/10.1016/j.catena.2017.07.020
Received 21 February 2017; Received in revised form 30 June 2017; Accepted 27 July 2017

value. The same can be argued for the work carried out by organisations as ISRIC - World Soil Information, striving to increase awareness and understanding of soils in major global issues (www.isric.org/utilise/global-issues). At the political level, the implementation of policies such as EU Thematic Strategy for Soil Protection (COM(2006)231 final) are essential to trigger actions and plans at the national scale aiming for the protection and sustainable use of soil and promoting its restoration (Ballabio et al., 2016).

However, the definition and assessment of sound and sustainable land management policies remain a difficult challenge, being many times hampered due to the lack of updated, comparable, and reliable soil data (Montanarella, 2007; Tóth et al., 2013a, 2013b). Reliable land management decisions at field or catchment scales require detailed soil information, including knowledge of the spatial variability of soil properties with depth and across geographic areas. However, most conventional soil survey maps present a series of polygons delineated mostly according to qualitative criteria, which are generically capable of portraying soil heterogeneity and to describe structural patterns across the landscape (Lin, 2003, 2010), but which do not adequately express the complexity of soils within the same mapping units (Sanchez et al., 2009). Also, conventional soil survey maps are most times associated with data measured in representative soil profiles, which may be obsolete or out-of-date (Lin et al., 2006; Hartemink et al., 2013; Ramos et al., 2013).

As land management models and decision supporting tools become more sophisticated, there is thus the increasing need of developing modern soil maps based on detailed soil information in order to improve agricultural and environmental planning and monitoring (Sanchez et al., 2009; Panagos et al., 2012; Shangguan et al., 2013). This is the main objective of the GlobalSoilMap.net consortium (Sanchez et al., 2009; Arrouays et al., 2014, 2017), which aims to make a new digital soil map of the world using state-of-the-art and emerging technologies for mapping and predicting soil properties at fine resolution. An example of this effort are the recently released global SoilGrids at 1 km and 250 m resolution (Hengl et al., 2014, 2017), which are produced according to GlobalSoilMap specifications. Also, the Harmonized World Soil Database (HWSD; FAO/IIASA/ISRIC/ISS-CAS/JRC, 2012) shares a similar goal, combining the 1:5 000 000 scale FAO-UNESCO Soil Map of the World with the information from the European Soil Database (Lambert et al., 2003; Tóth et al., 2013b), the Global and National Soils and Terrain Digital Databases (SOTER, van Engelen and Dijkshoorn, 2013), the World Inventory of Soil Emission Potentials (WISE; Batjes, 2009), and the soil map of China at 1:1 million scale (Shi et al., 2004).

Despite all the above mentioned initiatives and advances on soil mapping techniques, the available soil information seems still insufficient for many regions in the world, including Portugal. The WISE dataset (Batjes, 2009), for example, and also the World Soil Information Service (WoSIS) dataset (Batjes et al., 2016), contains information of 10,253 soil profiles collected throughout the world over the last decades, but only 10 of these are actually from Portugal. An improved set with physical and chemical soil data was made available in 2009 through the LUCAS (Land Use/Cover Area frame Statistical Survey) soil database (Tóth et al., 2013b), a project promoted under the EU Thematic Strategy for Soil Protection as a response to the lack of soil quality data. The LUCAS database was the first attempt to construct a harmonized pan-European topsoil (0–20 cm) geodatabase, which could serve as a baseline for EU-wide soil monitoring. LUCAS was the dataset used in SoilGrids (Hengl et al., 2017) to generate soil predictions for Portugal. To conduct LUCAS, the EU' territory was divided using a $2 \times 2\,km^2$ grid whose nodes constituted around 1.1 million points. From this, a sample of 270,000 points were selected on the basis of stratification information, and 465 points were allocated to Portugal. Nonetheless, these numbers of available soil data for Portugal seem surprising small considering that the country has had active soil survey services since the 1940's (first in “Estação Agronómica Nacional”, currently included in “Instituto Nacional de Investigação Agrária e Veterinária”, and later in “Serviço Nacional de Reconhecimento Agrário”, now part of “Direcção Geral de Agricultura e do Desenvolvimento Rural”), and that there was a significant investment on mapping soils in different regions of Portugal during the 1990's and 2000's (e.g., Agroconsultores and Geometral, 1999; Divisão de Solos, 2003; Geometral and Agroconsultores, 2004; DGADR, 2007).

Thus, since data is only valuable when used by technicians, policy makers, and scientists (Lin et al., 2006), why is soil information produced in Portugal not put to a better use? Despite many valid constraints documented in Madeira et al. (2004) and Gonçalves et al. (2005), the main reasons seem to be lack of coordination, with data being scattered among several different institutions, and the non-existence of a modern soil information system capable of storing all soil information produced in Portugal, including that from all Universities, Polytechnics, and State laboratories dedicated to soil science. Only by overcoming these two constraints it will then be possible to assess the quality of the existing data, identify the main gaps on soil information, define the needs for future studies and research, and properly evaluate the impact of national and European policies on land and the environment. One valid example of what should be done with Portuguese soil data is the PROPSOLO database (Gonçalves et al., 2011; Ramos et al., 2013, 2014), specifically developed for storing all reliable information on soil hydraulic properties determined in the country, and which was already included in the European Hydropedological Data Inventory (EU-HYDI; Weynants et al., 2013) in order to develop the new generation of hydraulic pedotransfer functions for Europe (Tóth et al., 2014). The successful example of the PROPSOLO database is now extended to include other soil data and create a unified database named INFOSOLO.

Therefore the objectives of this paper are: (i) to present the INFOSOLO relational database as a first step towards the development of a modern soil information system in Portugal; (ii) to highlight the importance of using the INFOSOLO legacy dataset to understand the spatial and temporal distribution of selected soil properties, namely the organic carbon content – OC (%), pH, and cation exchange capacity – CEC ($cmol_c/kg$); and (iii) to illustrate the capabilities of the new database to characterize national soil spatial patterns by providing 1 km resolution maps of topsoil OC, pH and CEC at the national level. These selected soil properties are included in the minimum dataset established by the GlobalSoilMap (GSM) consortium to produce relevant maps for decision making (Sanchez et al., 2009).

To accomplish objectives ii) and iii), we compared the soil information generated using INFOSOLO with the information obtained using the European dataset LUCAS. INFOSOLO is a legacy dataset with data collected by different institutions, with no specific or comparable sampling design. On the other hand, LUCAS sampling followed a statistical design but the data was collected for a specific year (2009). Therefore, our analysis aimed to understand if there were significant differences in the soil characterization provided by the two datasets. For mapping the soil properties at the national level, we also tested two spatial modelling approaches to evaluate the importance of including environmental covariates to explain the spatial distribution of OC, pH and CEC. The modelling outputs were assessed in terms of their global accuracy and expert knowledge. The final national maps were further compared with the worldwide soil map SoilGrids. This comparison complemented the objectives of this study since SoilGrids is currently the only digital soil map available for Portugal.

Our study follows the commemorations of the International Year of Soil 2015 in Portugal, and provides the basis for improving soil information in the country, for informing a future national soil monitoring program, for updating existing digital soil information, and, ultimately, for raising national awareness of the importance of soil resources to environmental wellbeing and socio-economic development.

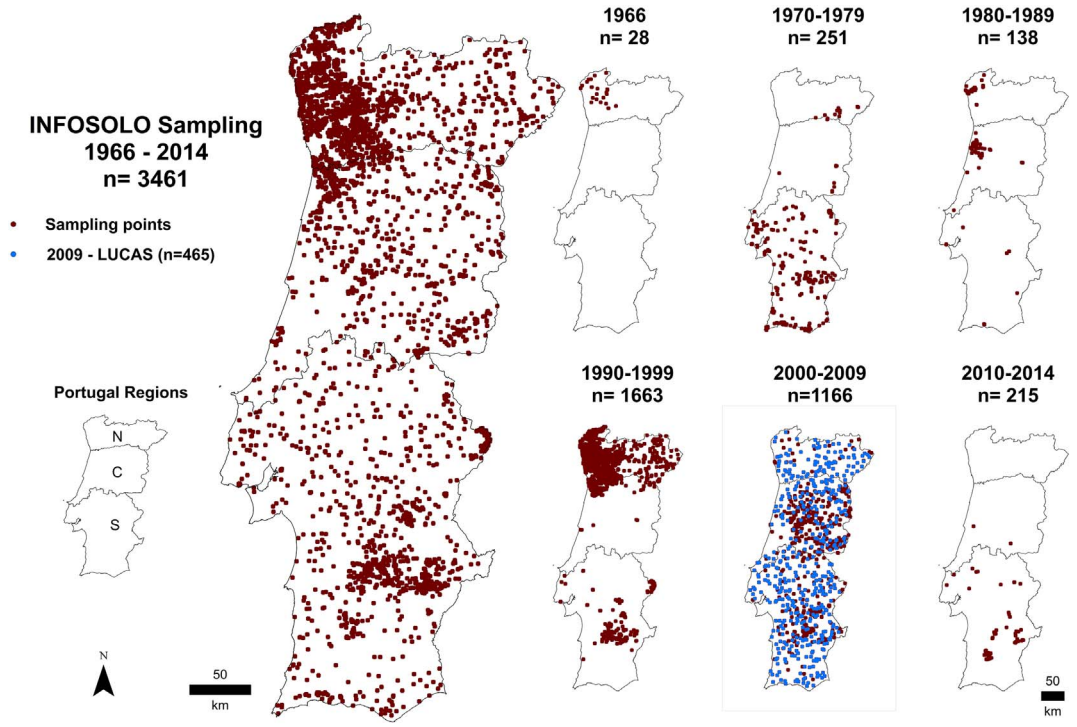

Fig. 1. Location of INFOSOLO and LUCAS sampling points (national distribution and per decade).

2. Material and methods

2.1. The database

The INFOSOLO database is the most comprehensive effort ever made to organize soil information in Portugal. It currently includes a set of 9934 horizons/layers studied in 3461 soil profiles across the country (Fig. 1), from which 570 only contain data for the topsoil horizon/layer. All information of a large number of soil related parameters, namely on physical and chemical properties, were obtained from soil surveys, research projects, and academic studies performed in public Portuguese and other European institutions (Table 1) over the last few decades (1966–2014). These sampling campaigns were not aligned with a national monitoring program, rather they were funded by a range of projects and surveys with different interests, from regional soil mapping to agricultural developments. All data compiled into the database was available in paper reports, thesis, and online. Most were freely accessed, but in some cases data was protected under a license agreement (e.g., Tóth et al., 2013a).

INFOSOLO is a georeferenced relational database developed for the

Table 1
Institutions responsible for the data included in INFOSOLO.

Institutions	Number of soil profiles	Frequency (%)
Direcção Geral de Agricultura e do Desenvolvimento Rural	733	21.2
Direcção Regional de Agricultura e Pescas do Norte	1233	35.6
Instituto da Conservação da Natureza e das Florestas	103	3.0
Instituto Nacional de Investigação Agrária e Veterinária	622	18.0
Instituto Superior de Agronomia	79	2.3
Joint Research Centre	465	13.4
Sociedade Portuguesa da Ciência do Solo	4	0.1
Universidade de Évora	30	0.9
Universidade de Trás-os-Montes e Alto Douro	192	5.5

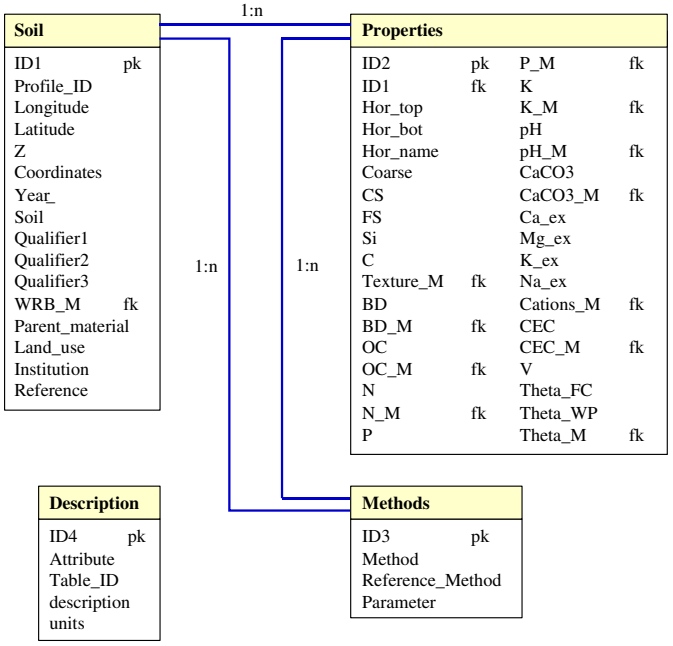

Fig. 2. Structure and attributes included in the INFOSOLO database (pk and fk correspond to primary and foreign keys, respectively).

MYSQL 5.6 (www.mysql.com/) operating system. It is divided into four simple and intuitive tables (Fig. 2): (i) SOIL, which includes the identification of the soil profiles (Profile_ID) and corresponding geographical coordinates (Longitude, Latitude), elevation (Z), year of sampling (Year_), WRB Reference Soil Groups and qualifiers (Soil, Qualifier 1–3; IUSS Working Group, 2006), parent material (parent_material), and land use (Land_use), as well as an indication whether the geographical coordinates were available or had to be estimated (Coordinates; see Section 2.2), the name of the institution responsible for the data (Institution, Table 1), and the reference from which data

was extracted; (ii) PROPERTIES, which includes the top (Hor_top) and bottom (Hor_bot) depths of each horizon/layer, the symbol representing the horizon/layer (Hor_name) (FAO, 2006), and corresponding analytical data, i.e., information on coarse elements (Coarse), coarse sand (CS), fine sand (FS), silt (Si), clay (C), bulk density (BD), organic carbon (OC), total nitrogen (N), extractable phosphorus (P), extractable potassium (K), pH, $CaCO_3$, exchangeable cations (Ca_ex, Mg_ex, K_ex, and Na_ex), cation exchange capacity (CEC), base saturation (V), and soil water content at field capacity (Theta_FC), and at the wilting point (Theta_WP); (iii) METHODS, which holds the information on the methodologies used for analytical data characterization or soil profile classification (Method), their references (Reference_Method), and the soil parameter related to that methodology (Parameter); and (IV) DESCRIPTION, which basically provides the metadata for the database, i.e., it lists the name of the attributes included in the database (Property), the tables where they can be found (Table_ID), their meanings (description), and units of the variables (in the case of the geographical coordinates, elevation, and analytical data).

Tables SOIL and PROPERTIES also contain foreign keys (all attributes ending with _M) that relate a corresponding analytical determination or soil classification to its methodology or soil classification system, respectively, found in table METHODS. It is a simple linkage procedure already used in the Portuguese soil hydraulic properties database (Gonçalves et al., 2011) and recently adopted also in the EU-HYDI database (Weynants et al., 2013). All field attributes can hold null data (i.e., absent data) with the exception of Profile_ID. Similar principles as those adopted in INFOSOLO can also be found, for example, in Leenaars et al. (2014a).

2.2. Quality assurance and harmonization

The INFOSOLO database holds data from many different studies, not always with related objectives. Thus, the level of detail of soil information varied between datasets, with the quality of data being assessed through a series of validation tests. The location of the soil profiles is one of the best examples where the level of detail differed between datasets, being also pointed out as a critical issue in similar studies (e.g., Leenaars et al., 2014b). The geographical coordinates were available for 49% of the soil profiles mainly due to a recent release from the "Direcção Geral de Agricultura e do Desenvolvimento Rural" (Table 1). However, the coordinates of the remaining soil profiles included in INFOSOLO had to be estimated based on broad references given in the reports. This was the case of the representative soil profiles found in most soil survey studies (mainly the oldest data), where the coordinates were never made available and, in most cases, are now lost. A "likely" location was thus included in INFOSOLO after relating each soil profile with the corresponding soil mapping units located closest to the farm, place, village, or civil parish mentioned in those reports. The field attribute Coordinates found in table SOIL refers whether the location of the soil profile was available or had to be estimated (i.e., if it corresponds to a "likely" location). Thus, users can make their own decision on using that information or not. Geographical coordinates were all converted to longitudes and latitudes in decimal degrees (WGS84), but can also be accessed using a local projection system (Lisboa Hayford Gauss IGeoE) if necessary. Elevation data, when not available, was obtained from military maps, in consistency with the geographical coordinates defined for each soil profile.

The soil profiles included in INFOSOLO were originally available under different soil classification systems. This was perhaps the major constraint when handling soil information from Portugal. Most soil profiles studied south of the Tagus River were classified under the Portuguese soil classification system (Cardoso, 1965, 1974), while those studied in the north followed different versions of the FAO classification system. No harmonization of soil classification data was ever attempted. In this study, an effort was made to convert the profiles classified according to the Portuguese soil classification system into the WRB 2006 (IUSS Working Group, 2006) framework. The name of the soil units, soil description (when available), and corresponding analytical data were considered when converting one soil classification into the other. Some expert judgment had also to be considered. The soil profiles classified according to older WRB versions were updated to WRB 2006. The soil profiles already classified according to WRB 2006 were also revised as some were considered questionable. As a result, INFOSOLO now includes 2895 soil profiles classified according to WRB 2006, while the remaining 566 do not have the necessary information for soil classification (Table 2), mostly because only the topsoil horizon/layer was characterized (e.g., Tóth et al., 2013a).

Table 2
Soil reference groups available in the database.

Soil reference groups	Number of soil profiles	Frequency (%)
Acrisols	48	1.4
Alisols	24	0.7
Anthrosols	441	12.7
Arenosols	57	1.6
Calcisols	127	3.7
Cambisols	632	18.3
Ferralsols	25	0.7
Fluvisols	231	6.7
Gleysols	19	0.5
Histosols	3	0.1
Leptosols	185	5.3
Lixisols	6	0.2
Luvisols	288	8.3
Planosols	19	0.5
Plinthosols	3	0.1
Podzols	17	0.5
Regosols	592	17.1
Solonchaks	5	0.1
Solonetz	7	0.2
Stagnosols	2	0.1
Umbrisols	59	1.7
Vertisols	105	3.0
Not classified	566	16.4

Parent material was defined following the nomenclature used in the European Soil Database (Lambert et al., 2003). Only the major class levels (Table 3) were considered since the detail of information varied between datasets. The consistency of the dataset was thus found here to be preferable to detail. Some expert judgment was also considered in order to define the parent material for soil profiles that did not include that information but which could be estimated based on the soil type and description.

Land use information was provided for 2703 soil profiles (Table 4). The level of detail was also found to differ greatly between datasets, but some effort was made to define land use according to the LUCAS 2009 classification scheme (Eurostat, 2009).

The soil properties included in INFOSOLO (Fig. 2) were those more commonly found in the studies from where soil information was obtained. In this database, the particle size distribution (PSD) was defined

Table 3
Nomenclature of parent material.

Major class level	Number of soil profiles	Frequency (%)
Consolidated clastic sedimentary rocks	340	9.8
Sedimentary rocks	168	4.9
Igneous rocks	1027	29.7
Metamorphic rocks	541	15.6
Unconsolidated deposits	800	23.1
Eolian deposits	40	1.2
Organic materials	3	0.1
No information	542	15.7

Table 4
Major land uses.

Land use	Number of soil profiles	Frequency (%)
Arable crop	15	0.4
Irrigated crop	95	2.7
Irrigated arable crop	465	13.4
Rainfed crop	16	0.5
Rainfed arable crop	413	11.9
Rice	6	0.2
Cotton	2	0.1
Horticulture	152	4.4
Melon	3	0.1
Sugar beet	12	0.3
Mixed crops	9	0.3
Forest	145	4.2
Cedars	1	0.0
Chestnuts	11	0.3
Eucalyptus	15	0.4
Pine trees	115	3.3
Oak trees	29	0.8
Pasture	201	5.8
Fruit trees	35	1.0
Olive grove	197	5.7
Vineyard	148	4.3
Woodland	74	2.1
Fallow	541	15.6
Golf course	3	0.1
No information	758	21.9

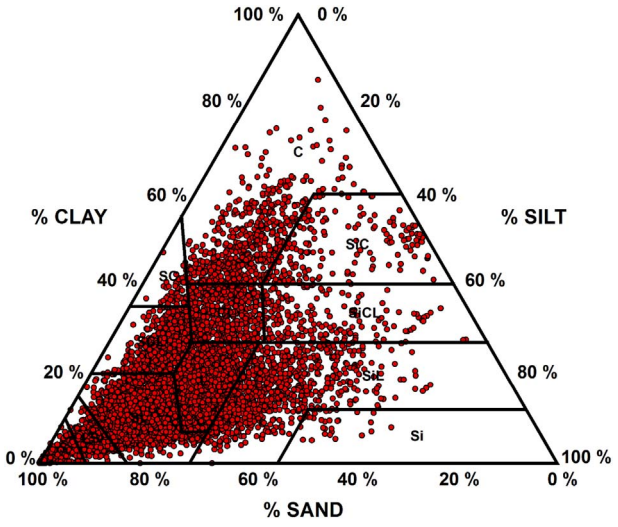

Fig. 3. Textural distribution of the dataset (S, sand; LS, loamy sand; SL, sandy loam; L, loam; SCL, sandy clay loam; CL, clay loam; SC, sandy clay; C, clay; SiC, silty clay; SiCL, silty clay loam; SiL, silty loam; Si, silt).

for particles of diameter < 2 μm (C), 2–20 μm (Si), 20–200 μm (FS), 200–2000 μm (CS), and > 2000 μm (coarse elements) (Fig. 3). These textural classes follow the Portuguese classification system (Gomes and Silva, 1962) and they are based on the International Soil Science Society (ISSS) particle limits (Atterberg scale). As a result, most texture data found for Portugal was already in accordance with these particle size limits. The exceptions were the data from Tóth et al. (2013a) and De Vos and Cools (2011), which used the particle size limit of 63 μm to divide sand from silt conform the standards adopted by FAO (2006), and Batjes (2009), which used the USDA particle size limit of 50 μm to make that same division conform the standards of USDA (Soil Survey Staff, 2014). Smoothing splines (Nemes et al., 1999) were fitted to the cumulative particle size limits of 0, 2, 63 or 50, and 2000 μm to obtain the cumulative percentages of particles at 20 and 200 μm. Smoothing splines were adjusted to the data with the Curve Fitting Toolbox 3.3 available in MATLAB R2012b version 8.0.0.783 (MathWorks Inc., Natick, MA, USA). The resulting RMSE (root mean square error) values were 0.6, 5.2, and 1.2% when comparing estimates with measured data available in Tóth et al. (2013a), De Vos and Cools (2011), and Batjes (2009), respectively. Additional testing was performed to all data to confirm that PSD (< 2000 μm) would sum 100%. Those that did not summed between 97 and 103% were considered to be erroneous and were not included in INFOSOLO, while those that were within that interval were corrected by distributing the error among the various texture classes (when similar weighted) or by adjusting the percentage of the dominant texture classes.

Bulk density data was only considered if determined on undisturbed soil samples (corresponding to the whole earth fraction including coarse fragments). Even so, data from Geometral and Agroconsultores (2004) was not included in INFOSOLO due to the low correlation value ($r = -0.29$) found between bulk density data and water content at field capacity; a value less than half of those computed from Ramos et al. (2014), Divisão de Solos (2003), and DGADR (2007) datasets. Furthermore, values falling outside the range 0.90–1.95 g cm^{-3} were not included as they were considered unrealistic for Portugal's continental area.

When necessary, organic matter (OM) data was converted into organic carbon (OC) content; P_2O_5 was converted into P; and K_2O was converted into K. OC, N, P, and K data were then checked for extreme values. The C/N ratio was computed and corrected when C/N < 7, by adjusting OC or N content (Kristensen et al., 2015). Soil reaction (pH) was checked for values falling outside the possible range (0–14). Nonetheless, four values where pH < 3 were removed as they were considered unrealistic. $CaCO_3$ data was analysed for values falling outside the possible range (0–100%), and to confirm that $CaCO_3 > 0$ would only be possible in horizons/layers with pH > 6.

CEC data was checked for extreme values, while base saturation (V) was tested for values falling outside the possible range (0–100%). Two problems were immediately identified in part of the data. In calcareous soils, base cations often summed > 100%. The main reason was attributed to the use of less adequate methodologies like the acid and neutral pH extractants used, for example, with the ammonium acetate at pH 7.0 method (Schollenberger and Dreibelbis, 1945), which may significantly overestimate Ca^{2+} and to a lesser extent Mg^{2+} (Sumner and Miller, 1996). The dominant cations (Ca^{2+} and sometimes Mg^{2+}) were thus corrected by reducing their concentrations until base saturation summed 100%. In saline soils, base cations occasionally also summed > 100%. The main reason seemed to be that the reported exchangeable cations were actually extractable cations. The former could only be obtained if the soluble cations had also been determined and subtracted from the extractable forms, which was often not the case. In these situations, the error was distributed among the four cations (Ca^{2+}, Mg^{2+}, K^{+}, and Na^{+}) based on the proportion of each element, and concentrations were lowered until base saturation summed 100%.

Values of Theta_FC and Theta_WP were indirectly estimated from sand, silt, and clay content using the ternary diagrams developed in Ramos et al. (2014) with the empirical best linear unbiased predictor algorithm (Lark et al., 2006). These diagrams correspond to kriging surfaces which interpolate existing water content observations in the PROPSOLO database (Gonçalves et al., 2011) across the texture triangle with a relatively low error (RMSE ≤ 0.040 cm^3 cm^{-3}) when compared with other pedotransfer interpolation techniques that require the same data inputs (Ramos et al., 2014). Water content data in INFOSOLO may thus be considered redundant since it is not actually measured data, but something that can be obtained from other soil properties. The main reasons for adopting such procedure were related to the importance of Theta_FC and Theta_WP for computing water and nutrient budgets combined with the lack of information at the national scale; the fact that soil water content information available in most Portuguese soil survey studies may be considered obsolete as it was determined on

disturbed soil samples or simply lacks the quality necessary to be considered in modern hydrological studies; and the existence of the PROPSOLO (Gonçalves et al., 2011) and EU-HYDI (Weynants et al., 2013) databases, already developed specifically for storing reliable soil hydraulic property information.

While some of the soil properties included in INFOSOLO rely on the same principles, others were determined using quite different methodologies. Most studies used the pipette method for determining clay and silt, and sieving for measuring the sand components (Gee and Or, 2002). The main difference between studies was thus the particle size limits adopted to distinguish silt from sand. Also, the methods used for determining bulk density were almost always to dry a volumetric sample in the oven at 105 °C for 48 h, only varying the sample size. On the other hand, soil organic carbon was determined using seven different methodologies. The Walkley-Black method (Nelson and Sommers, 1996) was used in 48% of the available data. No harmonization was yet considered in order to relate results from different methodologies. Some studies used more than one method for determining the same soil property. In the cases where it was not possible to distinguish between methods, all cited methods were related to that soil property using the reference "Not discriminated (method 1, method 2 …)".

2.3. Spatial and temporal distribution of soil properties

To highlight the importance of using the INFOSOLO legacy dataset to understand the spatial and temporal distribution of selected soil properties (OC, pH, and CEC), the INFOSOLO data was divided according to the sampling region – north, central and south (Fig. 1). These three regions are related with the administrative areas proposed by the European Union Statistical Regions (NUTS II) and present distinctive geographic and climate conditions that create different landscapes from north to south (in this study, the south region comprised Lisbon, Alentejo, and Algarve NUTS II). In the north, the landscape is highly diverse due to geomorphology, lithology, land cover, rock outcrops, climate, and a wide variety of land use systems (e.g., Douro vineyards). Precipitation differs significantly, with the greatest values (e.g., 1466.5 mm in Viana do Castelo (1981–2010); www.ipma.pt) being registered in Minho (NW region) due to the topographic characteristics of the region (Fig. 4). The average temperatures also differ substantially, with Trás-os-Montes (NE region) exhibiting larger amplitudes between seasons (e.g., 4.5–21.7 °C in Bragança (1981–2010); www.ipma.pt). In contrast, the south region is characterized by greater average temperature (e.g., 9.7–33.3 °C in Beja (1981–2010); www.ipma.pt), lesser average rainfall (e.g., 557.8 mm in Beja (1981–2010); www.ipma.pt), extensive agriculture areas, and mostly by flat landscapes (Fig. 4). The highest elevations are located in the eastern and NE areas of the country, clearly distinguishing the north and central regions from the south.

For each region, a comprehensive data analysis was conducted to identify spatial and temporal trends observed for measured OC (%), pH, and CEC ($cmol_c/kg$) values in 3397 INFOSOLO topsoil samples, which were mostly (70%) collected from 0 to 25 cm. The same analysis was carried out for the LUCAS dataset (465 topsoil samples) to compare statistically both datasets using exploratory data analysis and spatial continuity analysis. The LUCAS sampling campaign was conducted following a properly designed survey to collect soil data at the national scale for one specific time period (year 2009). Hence, the data analysis aimed to ascertain if the LUCAS statistics and spatial continuity patterns contradicted significantly the ones derived using the legacy dataset INFOSOLO. This was deemed important to justify the use of INFOSOLO to characterize the spatial and temporal distribution of OC, pH and CEC at the national level.

Exploratory data analysis consisted in the description of the spatial distribution of sampled OC, pH and CEC complemented with basic statistics and box-plots. Additionally, data distributions were represented with histograms and compared statistically. This information was relevant for the INFOSOLO and LUCAS comparison and interpretation of the spatial modelling results.

For the spatial continuity analysis, experimental variograms were used to describe the spatial patterns revealed by the INFOSOLO and LUCAS datasets for each selected soil property. The variogram allows to describe how a value measured at a sampling location is expected to be correlated with values close by within a certain range (also known as spatial autocorrelation; Webster and Oliver, 2007). This is measured by calculating the average semivariance among pair of points located a distance (lag) h apart (Isaaks and Srivastava, 1989). In this study, the experimental variogram was built by plotting the semivariance values $\gamma(h)$ against the lags h without considering a preferential direction. These omnidirectional experimental variograms calculated using OC, pH, and CEC point data were used to describe and compare the regional and national spatial continuity within both INFOSOLO and LUCAS.

2.4. Spatial modelling of soil properties at the national level

Digital soil mapping (DSM) is currently adopted in most soil science studies as a framework to characterize the spatial pattern of soil properties. DSM has become widely accepted as a concept and framework after the work presented by McBratney et al. (2003), which recognizes the advantages in producing digital soil information using spatial modelling algorithms incorporating available covariates (predictor variables) related to the expected spatial and temporal distribution of soil properties. The application of DSM to improve the quality and availability of soil information has been extended to the world scale with the GSM project, the first world-wide initiative that aims to encourage the use of state-of-the-art DSM technologies to predict and then map soil properties at a fine (100 m) resolution (Sanchez et al., 2009).

In the context of GSM, the collection of legacy datasets (i.e., "pre-existing, georeferenced field or laboratory measurements"; Sanchez et al., 2009) is an important input for DSM. However, the quality of legacy data obtained through traditional soil surveys might be impacted by poor sampling design (generally empirical and lacking statistical criteria) resulting in sampling bias (Carré et al., 2007). Also, traditional soil sampling is often expensive and time-consuming, which also results in spatial coverage limitations. This particular aspect is conveniently covered in the DSM framework with the use of spatially continuous covariates. With the recent developments in geospatial technologies, soil-related covariates are easily (and often freely) acquired at different resolutions making it possible to include them as predictors in modelling algorithms. As for these, a choice exists between algorithms incorporating spatial correlation (spatial modelling), and those exploring linear and non-linear relationships between variables but not using spatial correlation (machine learning or non-spatial modelling). Although there is no "best" choice, spatial models seem to perform better in terms of achieving higher prediction accuracy (Horta et al., 2013).

Spatial modelling integrating environmental covariates was used by Rawlins et al. (2009), Kempen et al. (2010), Hengl et al. (2014), Aksoy et al. (2016), and Ballabio et al. (2016) to predict soil properties at different scales (from regional to continental). In this work, spatial modelling was performed to predict OC, pH and CEC at the national level using a kriging-based algorithm, namely an empirical best linear predictor (EBLUP) incorporating a linear mixed model built with a set of environmental soil-related covariates (Sections 2.4.1 and 2.4.2). This was done using both INFOSOLO and LUCAS as input data to later compare their mapping outputs (Section 2.4.3).

To evaluate the importance of including environmental covariates in the soil maps, modelling was also performed using raw data only (i.e., without considering the contribution of covariates). This was done using a second kriging-based predictor (ordinary kriging, Section 2.4.2). The results were then compared with the previous modelling using EBLUP to conclude which map output should be chosen to

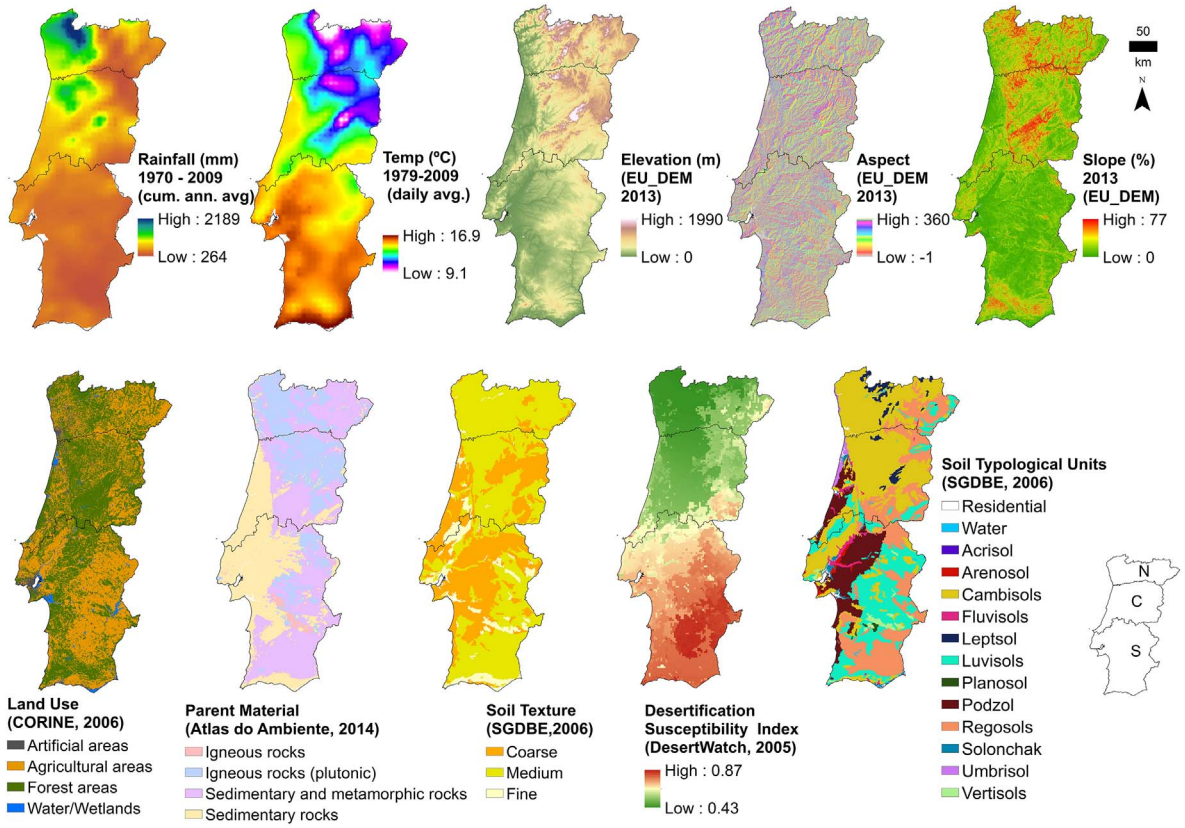

Fig. 4. Covariates for spatial modelling.

describe the spatial patterns expected for OC, pH and CEC at the national level.

Comparison of spatial modelling outputs (Section 2.4.3) was done considering model validation statistics and expert knowledge. The final national maps were further compared with the digital soil maps currently available for Portugal, namely, the SoilGrids product (Hengl et al., 2017).

2.4.1. Environmental covariates

Choosing the right covariates to model the spatial distribution of OC, pH and CEC is to balance their relevance for the case study (i.e., making sure they are directly or indirectly linked to the environmental processes conditioning these soil properties) but also their availability and cost. Following this rationale, thirteen freely available covariates were selected (Fig. 4 and Section 3.3). These included climate (average rainfall and temperature), land use, soil units, soil texture, parent material, and terrain attributes (elevation, slope and aspect). All of these covariates are commonly used for digital soil mapping based on the *scorpan* model proposed by McBratney et al. (2003) (which partly comprises the most dominant factors of soil formation proposed by Jenny, 1941). A desertification susceptibility index was further included with the intention of describing the combined effects of climate, soil and time to evaluate if the process of desertification is likely to occur. The details of the covariates used as continuous and categorical predictor variables are described below:

- The climate variables (rainfall (mm; range: 263.5–2190 mm) and temperature (°C; range: 9–17 °C)) were provided by the "Grupo de Previsão Numérica do Tempo" (Instituto Superior Técnico; meteo.tecnico.ulisboa.pt/) and refer to average climate simulations for mainland Portugal for the period 1979–2009 using the numerical mesoscale model MM5 forced by the initial conditions from "The NCEP Climate Forecast System Reanalysis", at a spatial resolution of 9 km.
- The terrain attributes were obtained using the digital elevation model (DEM; range: 0–1990 m) provided by the European Environment Agency (DEM over Europe, 2013; www.eea.europa.eu/data-and-maps/data/eu-dem) based on the NASA Shuttle Radar Topography Mission and ASTER imagery, with a spatial resolution of 25 m. DEM pre-processing and deriving slope and aspect were carried out using the software ArcGIS 10.2.2 (ESRI, 2014).
- Land use was extracted from the European CORINE Land Cover map produced and validated for 2006 (www.eea.europa.eu/data-and-maps/data/corine-land-cover-2006-raster/#parent-fieldname-title). The CORINE Land Cover inventory is based on automatic or semi-automatic classification of satellite imagery (for 2006, SPOT-4/5 and IRS P6 LISS III dual date) to provide a map comprising 44 land use categories, grouped in 5 broad classes (artificial surfaces, agricultural areas, forest and semi natural areas, wetlands and water bodies) at a spatial resolution of 250 m. The most represented land use classes in Portugal were class 2 (agricultural areas) and class 3 (forest and semi natural areas, namely transitional woodland-shrub).
- Parent material referred to the main geological units provided by the national environment agency (sniamb.apambiente.pt/infos/shpzips/AtAmb/AtAmb_1013111_CLitologica_Cont.zip) within the project "Atlas do Ambiente", which aims to make available geographic information at the national scale (1:1000000), useful for environmental thematic mapping. Parent material data was provided as a vector layer which was converted to a raster with 1 km spatial resolution, using the conversion tools within ArcGIS 10.2.2 (ESRI, 2014). The most represented geological units in mainland Portugal were sedimentary and metamorphic rocks.
- Soil units and soil texture were obtained from the European Soil Database (ESDB) v2.0. (van Liedekerke et al., 2006; Panagos, 2006;

Panagos et al., 2012), which provides raster data for 73 soil attributes with a spatial resolution of 1 km (esdac.jrc.ec.europa.eu/content/european-soil-database-v2-raster-library-1kmx1km). Soil units used in this study corresponded to soils classified according to WRB (IUSS Working Group, 2006). ESDB soil texture refers to the dominant surface textural class. According to this source, soils in Portugal were mostly coarse (< 18% clay and > 65% sand). Soil texture mapped at the national scale was also used as a covariate. Clay, silt and sand contents were predicted using INFOSOLO data and a kriging-based algorithm (ordinary kriging) as explained in Section 2.4.2.

The desertification susceptibility map is the end-product of a European Space Agency (ESA) research project aiming at "developing a user-oriented Information System based on Earth Observation technology to support national and local authorities in responding to the reporting obligations of the United Nations Convention to Combat Desertification (UNCCD) and in monitoring land degradation trends over time" (www.melodiesproject.eu/node/36). Desertification susceptibility aims to capture desertification dynamics using a spatial inference method that combines a climatic component (using ECMWF - European Centre for Medium-Range Weather Forecasts precipitation data) and biophysical components such as land use, NDVI (Normalized Vegetation Index) and soil brightness (derived from SPOT imagery). The index varies between 0 and 1; higher values indicate a dynamic drift towards the advancement of desertification processes, which can be seen as a proxy for soil OC (indication of low OC content). The data available for this study referred to the susceptibility index calculated for 2005.A non-spatial approach was adopted to identify which environmental covariates were likely to be related with the measured OC, pH and CEC values. This approach used the machine-learning algorithm RandomForest (Breiman, 2001; Liaw and Wiener, 2002), as implemented in the R package randomForest (version 4.6–12, Liaw and Wiener, 2002). RandomForest can be described generally as using an ensemble of decision trees to predict new values at unsampled locations by exploring linear and non-linear relationships between variables, but not using spatial autocorrelation. According to Grimm et al. (2008), the RandomForest algorithm is suited for soil modelling due to its flexibility in using both categorical and continuous predictors and its ability to model high dimensional non-linear relationships avoiding overfitting. One of the outputs provided is a measure of how each covariate contributes to prediction accuracy. Covariance importance is quantified in terms of the variance of the out-of-bag predictions, which gives a measure of prediction accuracy (i.e., the mean square error obtained when predicting a subset of original data, not used in the training process, using the corresponding bootstrapping training tree (Svetnik et al., 2003)). This feature was used in this study for a preliminary analysis of relevant environmental covariates.

2.4.2. Spatial modelling

Spatial modelling was performed using two kriging based algorithms, which differ formally in terms of incorporating or not additional data (covariates) to predict new values at unsampled locations.

The first of the modelling approaches followed the DSM framework and used an empirical best linear unbiased predictor (EBLUP) incorporating the covariates through a linear mixed model (Lark et al., 2006; Lark, 2012). This DSM approach is commonly applied in soil science studies (Johnson et al., 2017; Li et al., 2015, 2016; Bishop et al., 2015; Oliver and Webster, 2014; Grunwald, 2009; Chai et al., 2008; Minasny and McBratney, 2007). A comprehensive explanation of linear mixed models (LMM) applied to DSM are well described in the recent work by Karunaratne et al. (2014) and Bishop et al. (2015), and are detailed in Lark et al. (2006) and Nelson et al. (2011).

LMM's can be presented as a model able to predict OC, pH and CEC by building a linear relationship between sampled values and a set of covariates (called fixed effects, which can be interpreted as a physical trend). The values of the covariates at each sampling location were found by spatial coincidence and without changing its spatial resolution (Bishop et al., 2015). The actual covariates used as fixed effects were different for each soil property and for each dataset, and were selected using backward elimination incorporating the Aikaike's Information Criterion (AIC) which determines their statistical significance (p-value < 0.05) to predict the specific attribute.

For continuous covariates, a preliminary analysis looked at the data distribution and the linear correlation among covariates, and between these covariates and OC, pH and CEC. Regression analysis was used to identify values in the distribution that had high leverage, and hence could affect the fitting of a linear model as required in further spatial modelling. Log-transformation was applied to reduce leverage influence when judged necessary. For categorical covariates (land use, soil units, soil texture at the European scale, and parent material), classes were reclassified or merged with a similar class if the number of observations extracted was below 15. When used for spatial modelling, each class in the categorical predictor variable was transformed into dummy variables, i.e., a binary indicator variable which is one (1) when a given class is present at a location and zero (0) otherwise (Samuel-Rosa et al., 2015).

Since the fixed effects contribution might not be enough to explain the variability observed for each property, the residuals (or random effects) from the regression model using the trend (fixed effects) were also calculated for each sampling point. For the spatial model to be worth using instead of solely using the spatial trend model, the residuals must be spatially correlated. The advantage of using the EBLUP approach with LMM's is that it allows to calculate simultaneously the regression relation between the soil property and the fixed effects, and the variogram of the residuals from that regression by residual maximum likelihood (REML) (Lark et al., 2006; Lark, 2012). Hence, the EBLUP predicted value incorporates both the contribution of the trend and the spatially correlated residuals. Moreover, using REML to fit the residuals variogram minimizes the bias in the predictions due to uncertainty in the estimated fixed effects coefficients (Johnson et al., 2017; Lark, 2012).

An important implication of using LMM's is the assumption of data normality. Whenever necessary, data transformation using the logarithmic function was used and predictions were back-transformed (using the lognormal probability density function suggested by Webster and Oliver, 2007) to produce the final predicted map. This was performed for OC and CEC modelling.

The second modelling approach used ordinary kriging (OK) to evaluate the importance of including covariates to predict OC, pH and CEC. This linear unbiased predictor assumes spatial stationarity and uses the raw data to produce local predictions based on spatially weighted neighbourhood data (Isaaks and Srivastava, 1989).

Ordinary kriging was also applied to predict the clay, silt and sand content at the national scale (1 km spatial resolution) using the INFOSOLO texture data. Although potentially introducing uncertainty due to the spatial model adopted to predict clay, silt and sand, we considered it was important to include these as covariates to predict OC, pH and CEC, not only due to their importance in explaining the variability of these soil properties at the national scale but also to complement the available European texture maps which represent broad classes at a global scale.

For both EBLUP and OK modelling, the dataset was firstly separated into calibration and validation subsets chosen randomly from the raw data. The calibration subset was used to create and test the spatial model for prediction whereas the validation subset was used for independent validation of the spatial model predictions (Section 2.4.3). Our validation subset comprised 40% of randomly chosen data values.

After the calibration and validation steps, the full dataset was used as input data for EBLUP and OK prediction of OC, pH and CEC, considering a 1 km spatial resolution grid covering Portugal.

The R software, namely the packages geoR (Ribeiro and Diggle,

2001) and gstat (Pebesma, 2004) were used for modelling, validation, and prediction. The final maps were processed in ArcGIS 10.2.2 (ESRI, 2014).

2.4.3. Model validation and comparison of spatial modelling outputs

Both kriging-based approaches used in this work relied on a spatial model for prediction of soil properties. In the OK modelling approach, this spatial model was derived by fitting a known function to the raw data experimental variogram whereas with EBLUP it was the result of combining the contribution of fixed and random effects. Although formally different, both spatial models predicted a new value for each grid location where no data was available. To test the accuracy of the spatial model in predicting new values, the full dataset was firstly divided into a calibration subset used to create the spatial model, and into a validation subset used for statistically compare the spatial model predictions with true (sampled) value.

Using the calibration dataset, leave-one-out cross-validation (LOOCV) was performed to measure the fitting performance of the spatial model (i.e., to evaluate how similar were the predictions and the correspondent true value used to create the model). LOOCV works simply by removing a data point from the dataset and predicting its value by kriging using the remaining data and the proposed spatial model. The process makes sure that each and every one of the n data points is omitted in turn from the dataset (Oliver and Webster, 2014).

Independent validation was also carried out using the spatial model for predicting soil properties at the locations included in the validation subset and for then measuring the prediction accuracy (i.e., the deviation between the predictions and the true value observed at locations "new" to the model).

The comparison between the predicted and true values can be summarized statistically using the mean error (ME) and the root mean square error (RMSE). Both provide a measure of prediction accuracy at each sampled location although not guaranteeing the same accuracy in the final predicted map. The ME always provides a value close to 0 since kriging-based algorithms are unbiased even when the spatial model is not adequate (Oliver and Webster, 2014). Hence, only the RMSE was used to compare the EBLUP and OK modelling outputs using the INFOSOLO and LUCAS datasets. Kriging should minimize the RMSE; therefore, larger RMSE values indicate less accurate predictions given by the spatial model.

Kriging-based algorithms also aim at minimizing the local error variance. For validation purposes it is important to guarantee that the prediction variance obtained using the spatial model accurately reflects the prediction errors. This can be measured using the squared standardized prediction error (SSPE), defined as (Lark, 2000):

$$\theta(x) = \frac{((z(x) - z'(x))^2}{\sigma^2(x)} \quad (1)$$

where $\theta(x)$ is the SSPE at location x, $z(x)$ and $z'(x)$ are, respectively, the true and predicted value at location x, and $\sigma^2(x)$ is the prediction variance derived by the spatial model at location x. If the correct spatial model is used, the mean $\theta(x)$ will be close to 1, and the median $\theta(x)$ close to 0.455 (Lark, 2000). These $\theta(x)$ reference values were used to compare the EBLUP and OK modelling outputs using the INFOSOLO and LUCAS datasets.

Besides using validation statistics, we mapped the relative difference between INFOSOLO and LUCAS predictions to quantify the deviations between the two maps. A qualitative comparison based on expert knowledge was also included to ascertain map quality given the pedological knowledge of Portuguese soils.

The final national OC, pH and CEC predicted maps were judged based on its accuracy and pedological value, and were further compared with the digital soil maps currently available for Portugal, namely, the SoilGrids product (Hengl et al., 2017). Although SoilGrids was produced at a global scale using a different modelling approach, the comparison with the national maps produced for OC, pH and CEC was considered relevant to discuss the importance of incorporating legacy datasets using local (national) customized modelling approaches.

3. Results and discussion

3.1. Soil information

Fig. 1 shows the sampling effort throughout Portugal, of which a significant part was concentrated in Minho and southern Alentejo regions; the former mostly as a result of soil survey (Agroconsultores and Geometral, 1999), while the latter were studies carried out by different institutions over the last decades, most with the objective of assessing the impact of irrigation on soil properties (e.g., Divisão de Solos, 2003; DGADR, 2007; Gonçalves et al., 2006; Ramos et al., 2011, 2012). During the 2000's, the sampling points were more evenly distributed, covering the entire country. This was mainly due to the sampling campaign performed in 2009 for the European Union (EU) soil monitoring program LUCAS (Tóth et al., 2013b), which is also included in INFOSOLO. In most recent years (from 2010 to 2014), the sampling effort has decreased, with most of the 215 points being located in the south.

Cambisols (18.3%), Regosols (17.1%), Anthrosols (12.7%), and Luvisols (8.3%) were those most represented in the database (Table 2), which is partially explained by the distribution of the soil profiles throughout the country. The large percentage of Cambisols and Regosols reflected the country's orography, namely the mountainous north; Anthrosols reflected mostly the long term anthropogenic activities carried out (e.g., ploughing, terracing, fertilization, liming), for example, in Minho; and Luvisols appeared mostly in the gently sloping landscape in the south, where the climate is dry sub-humid to semi-arid.

The parent material information also reflected the country's diversity, where igneous rocks (29.7%), unconsolidated deposits (23.1%), and metamorphic rocks (15.6%) were the most represented (Table 3). Granite and schist appeared dominantly in the north and central Portugal, while schist, gneiss, and limestone were found mainly in the south.

Most soil profiles were studied in areas under fallow or other non-productive areas (17.7%; Table 4). Areas with irrigated arable crops such as maize (13.4%), and rainfed arable crops such as annual winter cereals (11.9%) also stood out.

Table 5 describes the physical and chemical soil properties of the different horizons/layers included in the database. Soil texture data was more clustered within the coarse and medium texture classes of the ternary diagram, although the fine texture classes were also well represented (Fig. 3). The exceptions were the regions of the texture triangle where Si > 70% and C > 80%.

3.2. Spatial and temporal distribution of soil properties

3.2.1. Exploratory data analysis

The spatial distribution of OC, pH and CEC content recorded in the INFOSOLO database for the topsoil layer is displayed in Fig. 5 and the corresponding histograms are shown in Fig. 6. Accounting for all topsoil values measured from 1966 to 2014, all three soil properties present positive skewed distributions (more pronounced for OC). The mean OC content was 2.2% (considered "high" by Dias, 2000), which could be the result of the highly sampled north region where the OC content tends to be greater compared to the rest of the country (Fig. 5). Overall, OC spatial distribution matched what is expected for Portugal due to its geography and climate: high values located in the NW corner (low temperature and high rainfall) whereas low OC values characterize the south (higher temperatures and mineralization rates) and the NE region (larger temperature ranges). Regarding pH, low values (Costa, 1979) were concentrated in the north and central regions whereas alkaline values were mostly concentrated in the south. For CEC, high values

Table 5
Statistical description of the physical and chemical soil properties included in the database.

Soil property	N	Mean	St. Dev.	Maximum	Minimum	Skewness	Kurtosis
Coarse elements (%)	8353	19.2	16.4	87.7	0.0	0.98	0.84
Coarse sand (%)	9934	33.0	19.2	99.0	0.0	0.48	− 0.02
Fine sand (%)	9934	32.4	12.6	87.4	0.3	0.50	0.44
Silt (%)	9934	18.4	9.7	68.6	0.0	1.12	2.15
Clay (%)	9934	16.2	12.5	85.5	0.0	1.63	2.50
Bulk density (g cm^{-3})	1521	1.51	0.21	1.94	0.91	− 0.45	− 0.21
Organic carbon (%)	8074	1.47	1.57	24.19	0.00	2.61	13.49
N (g kg^{-1})	6386	1.30	1.08	13.18	0.01	1.74	6.28
P (mg kg^{-1})	5883	31.2	69.37	2816.0	0.0	15.5	496.1
K (mg kg^{-1})	5908	84.7	76.78	1019.4	0.0	3.23	20.09
pH (−)	9732	5.8	1.1	9.9	3.4	1.11	0.35
$CaCO_3$ (%)	9658	1.4	7.2	97.6	0.0	6.76	52.50
Exchangeable cations:							
Ca^{2+} ($cmol_c$ kg^{-1})	8310	4.61	7.14	55.9	0.0	2.81	9.09
Mg^{2+} ($cmol_c$ kg^{-1})	8312	1.56	2.82	42.4	0.0	3.36	17.42
K^{+} ($cmol_c$ kg^{-1})	8257	0.14	0.15	2.6	0.0	3.89	28.13
Na^{+} ($cmol_c$ kg^{-1})	8338	0.27	0.89	27.3	0.0	12.37	224.73
CEC ($cmol_c$ kg^{-1})	8880	13.57	8.91	65.3	0.1	1.49	2.95
V (%)	8252	40.8	34.4	100.0	0.0	0.61	− 1.11
Theta_FC (cm^3 cm^{-3})	9934	0.253	0.067	0.475	0.057	− 0.07	1.46
Theta_WP (cm^3 cm^{-3})	9934	0.126	0.062	0.334	0.014	0.95	0.61

CEC, cation exchange capacity; V, base saturation; Theta_FC, soil water content at field capacity; Theta_WP, soil water content at the wilting point.

seemed to be more predominant in the south region matching the spatial cluster observed for high pH.

Comparing the national distribution for the three variables (Fig. 5), the greatest OC content was associated with low pH values mostly in the north, in contrast with the south where low OC values were paired with high pH and CEC values. It is important to note the pattern in the NE corner where low OC contrasted with the high predominant OC values observed in the NW corner. Interestingly, these low OC values also matched low pH and CEC values. This can be attributed to the differences in terms of climate, topography, and land management found between NW and NE regions. In Minho (NW region), air temperature amplitudes are smoother due to the regulating effect of the Atlantic Ocean; mean rainfall is the greatest in the country (ranking among the greatest in Europe) due to the air circulation patterns across the Atlantic and cloud condensation when reaching the local mountainous relief; and land is mostly divided in small holdings which soils have been enriched by agriculture practices carried out over centuries (thus, the large amount of Anthrosols found in this region). On the other end, in Trás-os-Montes (NE region), air temperatures amplitudes between seasons are much wider, with cold winters and hot summers which enhance soil mineralization rates.

Although sampling was not collocated (thus not indicative of a

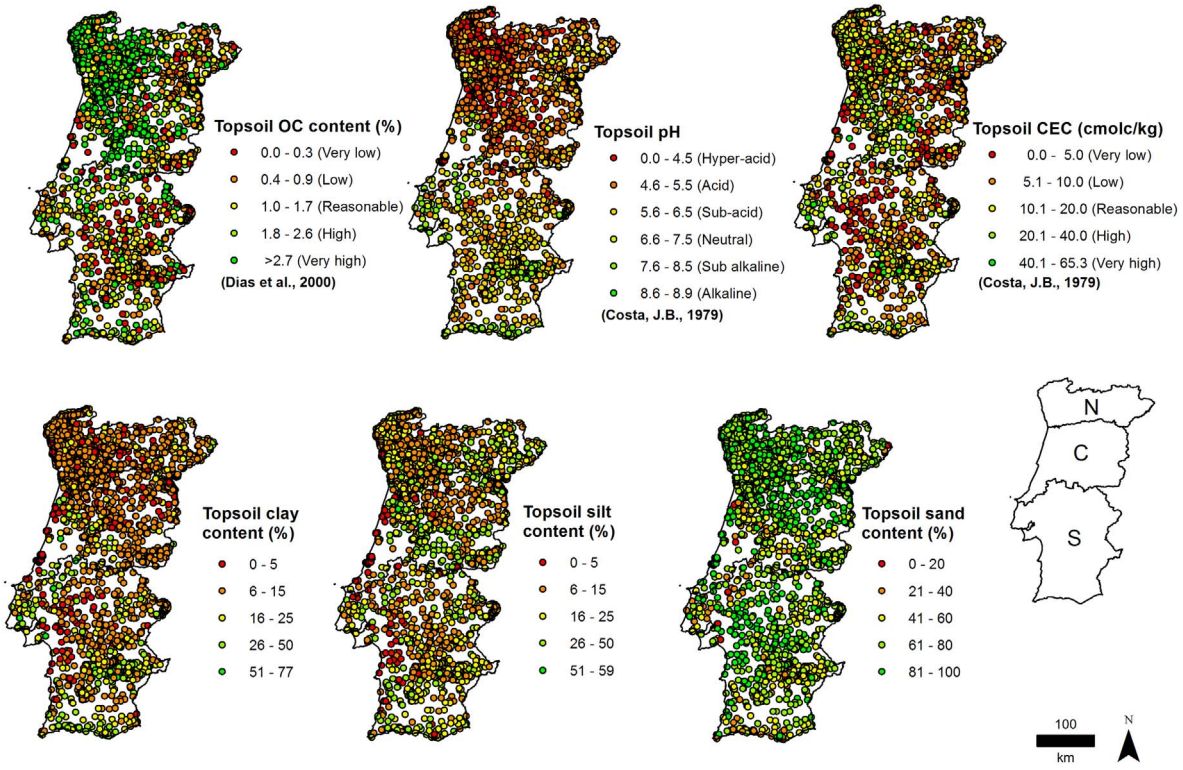

Fig. 5. Spatial distribution of INFOSOLO topsoil data (OC, pH, CEC, clay, silt and sand).

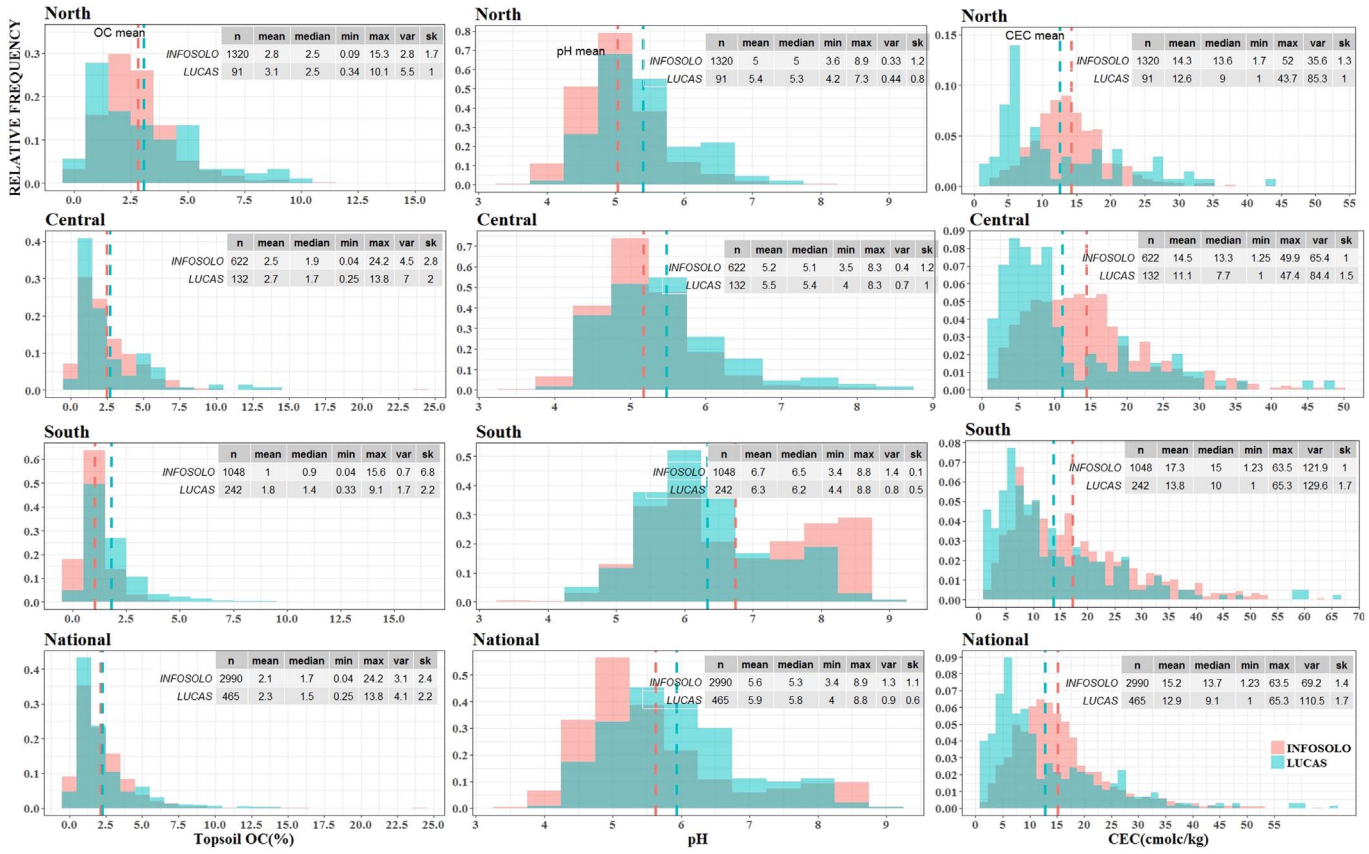

Fig. 6. National and regional histograms for OC, pH and CEC (comparison between INFOSOLO and LUCAS).

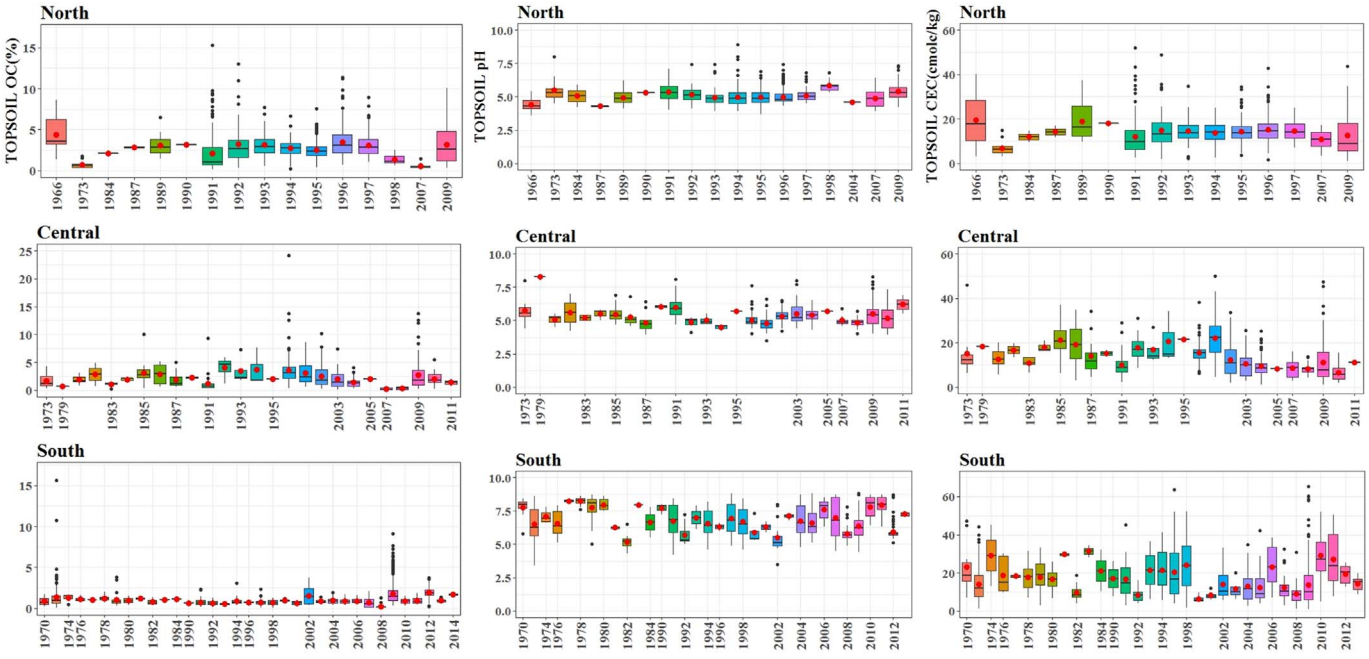

Fig. 7. Regional variability of INFOSOLO OC, pH and CEC.

temporal trend), there was a clear distinction among the range of OC content sampled in each region throughout time (Fig. 7). OC content was greater in the north and lesser in the south (with less variability among years and therefore locations). Regarding pH and CEC (Fig. 7), the highest values and variability characterized the south whereas the north and central regions presented the same pattern (with lower values and less variability).

In terms of how topsoil texture (Fig. 5) correlated with topsoil OC, pH and CEC, no significant global correlation was found between OC and texture for the three regions. pH and CEC were positively correlated with clay, and negatively correlated with sand in the south region. CEC presented a slightly positive correlation with clay in the north region.

Regarding the comparison between INFOSOLO and LUCAS, the boxplots in Fig. 7 show that the 2009 sampling captured the OC and CEC (but not pH) variablity measured for all the other years included in INFOSOLO in all regions.

Fig. 6 further compares the regional and national INFOSOLO and LUCAS histograms and basic statistics for topsoil OC, pH and CEC. The mean and median values characterizing both distributions did not differ greatly except for CEC. For this soil property, the LUCAS dataset included lower values and failed to represent the medium/higher values represented in the INFOSOLO. The predominance of lower CEC values in the LUCAS data was also noticeable in the histogram data distribution (namely for the north and central regions). For OC and pH, the national histograms showed that lower OC and pH values were less represented in the LUCAS dataset. Although this analysis highlighted only CEC distribution to be different for both datasets, the results of the two-sample Kolmogorov–Smirnov indicated that the INFOSOLO and LUCAS data distributions were statistically different for all soil properties under study.

A final note from this initial data analysis of the INFOSOLO dataset to stress the fact that, although sampling was done throughout time (from 1966 to 2014), locations were not revisited which makes the sampling year a proxy for location. This could be a limiting factor when using spatial modelling to characterize OC spatial patterns since OC content is likely to change substantially at the short scale and over time depending, for example, on changes in land use or farming practices (Franzluebbers, 2009).

3.2.2. Spatial continuity analysis

Fig. 8 compares the standardized experimental variograms obtained for the INFOSOLO and LUCAS datasets. Overall, the regional pattern presented by LUCAS was more erratic than the INFOSOLO pattern. The best visual match between experimental variograms was obtained for the central region for the three soil properties. For this specific region, long range continuity was found up to 40 km (OC), 80 km (pH), and 50 km (CEC). Also, there seemed to be a preferential spread in the north-south direction, with consistent high values measured for OC and CEC, and low values measured for pH, which could explain this continuity pattern.

In the north region, the variograms showed a continuous increase of the variance above the sill (more obvious for INFOSOLO). Possibly, spatial autocorrelation was controlled by other factors, for example, topography, rainfall, temperature, and land use (Fig. 4), which clearly defined a separation between the NE and the NW sides in the north region. In the case of OC content, very high to high values dominated the NW side whereas low values were measured in the NE.

In conclusion, no significant differences were found between the two datasets in terms of describing the regional and national spatial continuity patterns. However, the regional experimental variograms obtained with the INFOSOLO dataset presented a clear spatial continuity pattern for OC, pH and CEC when compared with the variograms calculated using LUCAS.

Comparing the regional (local) with the national (global) variograms, we concluded that, for OC, the global variogram was representative of the spatial correlation observed for each region. For pH, the global variogram was affected by an overestimation of the variance due to the cluster of higher values in the south region. This was more evident in the INFOSOLO data (Fig. 5) but it also affected the variogram calculated for LUCAS pH. For this soil property, it was thus necessary to correct the variance to estimate the sill for spatial model fitting when modelling pH using raw data and ordinary kriging. For CEC, although

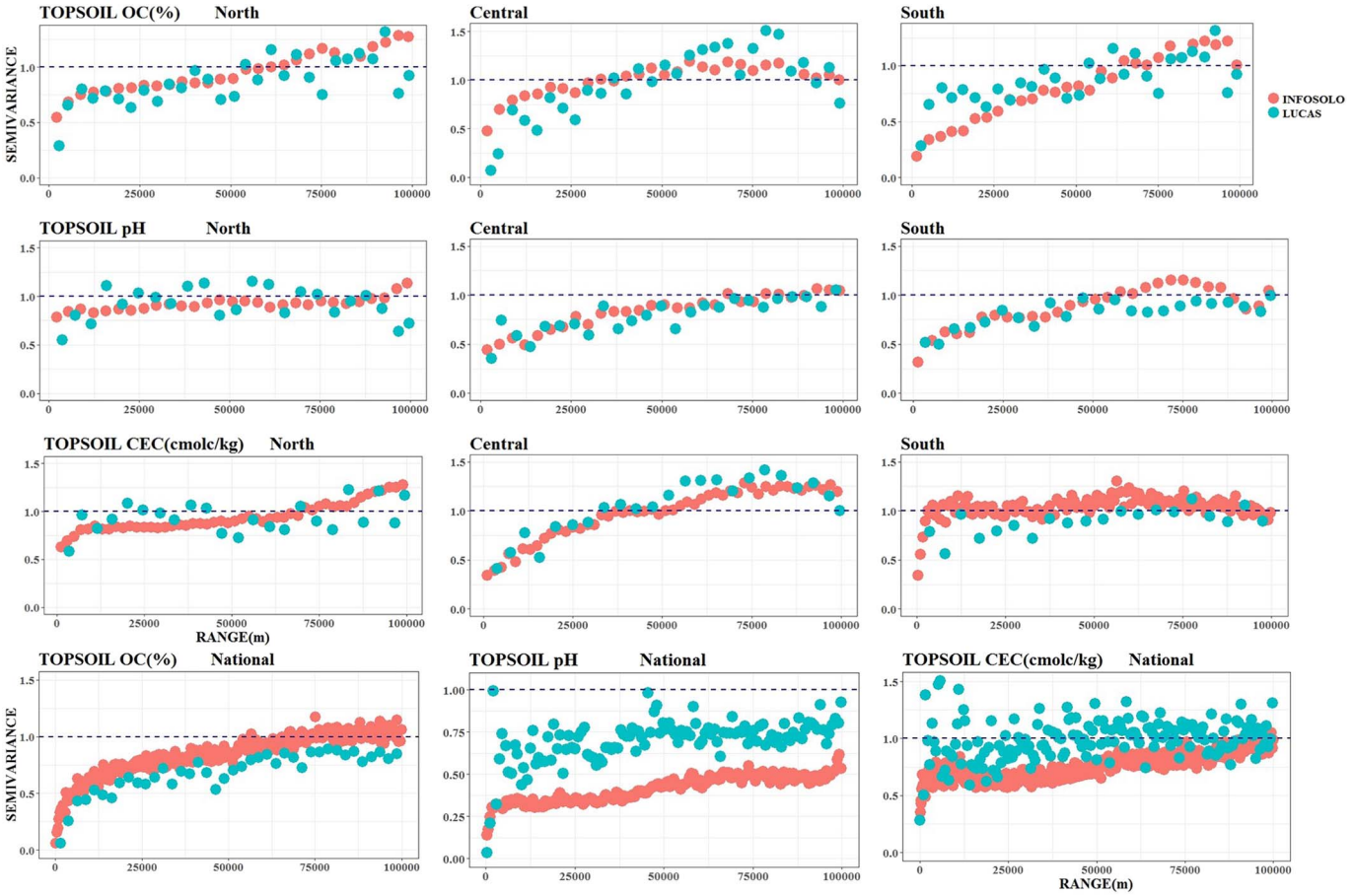

Fig. 8. National and regional spatial continuity analysis for OC, pH and CEC (comparison between INFOSOLO and LUCAS).

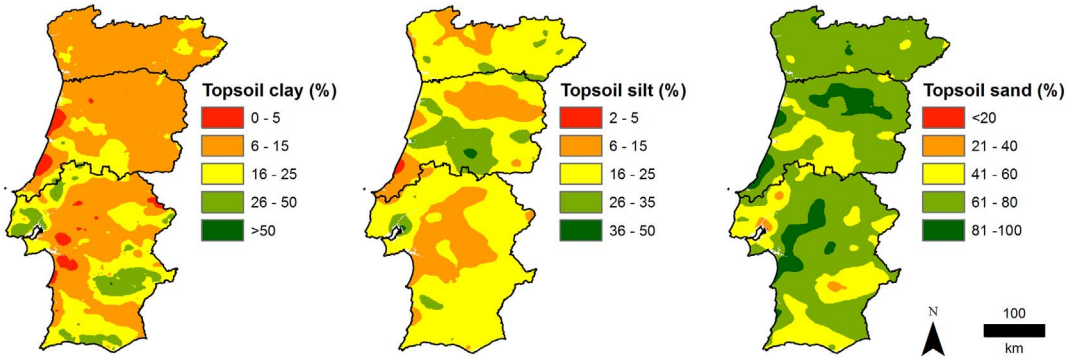

Fig. 9. National soil texture maps (clay, silt and sand).

the global variogram was not as well behaved as the OC one (probably also due to the cluster of high CEC values in the south region) it was considered stationary and thus used to describe CEC continuity at the national scale.

3.3. Mapping of additional covariates (national soil texture)

A decision was made to include the soil texture data (clay, silt, and sand content) available in the INFOSOLO dataset, sampled for the same locations as OC, pH and CEC (Fig. 5), as part of the covariates used for modelling. As these texture data needed to be available as a continuous grid, ordinary kriging was used to interpolate clay, silt and sand to create 1 km resolution maps (Fig. 9).

Finding the appropriate spatial model for prediction proved to be straightforward for silt and sand (both fitted with a spherical model), resulting in continuity ranges of 125 km and 85 km, respectively. For clay, log-transformation was necessary to avoid a pure nugget effect model. The experimental variogram was then fitted with a Matérn model which estimated a 25 km continuity range. All estimated ranges are adequate to explain the spatial distribution of these texture properties considering the scale involved in this study. Independent validation results (obtained as explained in Section 2.4.3) were similar for all spatial models and showed small RMSE values, median θ(x) value close to 0.45, and the mean θ(x) value slightly above 1. Based on these results, we considered the spatial models adequate for prediction of clay, silt and sand.

The predicted maps obtained for clay, silt and sand also captured the trend expected for Portugal, namely, the clay and sand distribution in the south, and the spatial distribution of silt areas generally in the south region and locally in areas in the left margin of the Tagus River and in the Mondego river catchment. These spatial trends generally aligned with the results presented in Ballabio et al. (2016), which used the LUCAS dataset to predict topsoil texture at the continental scale although more detail was added in our maps most likely due to the nature of the data and the spatial model used.

3.4. Spatial modelling of soil properties at the national level

3.4.1. Spatial modelling integrating environmental covariates

3.4.1.1. Covariates importance. The environmental covariates presented in Section 2.4.1 were used to create the trend component for predicting OC, pH and CEC using the EBLUP spatial modelling approach. Covariance importance obtained using RandomForest determined average rainfall as the most important variable to explain OC variability using both INFOSOLO and LUCAS. Additionally, elevation and silt were also highlighted for the INFOSOLO dataset. Clay was the most important covariate explaining pH and CEC variability using both datasets. For these soil properties, other covariates were indicated as relevant when using the LUCAS dataset, namely, parent material and desertification for pH, and sand for CEC. The analysis of covariance importance suggested that, from the initial set of covariates, only few of them were likely to contribute to explain OC, pH and CEC variability from which soil texture at the national scale was likely to be the most relevant.

3.4.1.2. Model calibration. The environmental covariates selected to predict OC, pH and CEC were similar to the covariance importance results. The final predictors determined for INFOSOLO OC were clay, silt, average rainfall and elevation (with a residual contribution) whereas for LUCAS OC the statistically significant predictors were average temperature, silt, average rainfall and elevation (also with a residual contribution). For INFOSOLO pH, the predictors with a larger contribution to the model were fine soil texture and soil type with silt, clay, average temperature and average rainfall also selected. For LUCAS pH, the final covariates included the parent material (sedimentary rocks), sand and silt, with climate covariates contributing residually. The results for INFOSOLO CEC indicated clay as the covariate with the largest contribution to the model followed by sand, silt, average temperature, and rainfall (with residual contributions). These results were similar for LUCAS CEC, with clay being again the covariate with the largest contribution followed by sand and average rainfall. The estimated fixed effect terms for the EBLUP spatial models used to predict OC, pH and CEC using both datasets are depicted in Table 6.

Comparing both datasets, the predictors selected for LUCAS OC (namely average temperature and silt) were statistically more significant than the ones contributing for modelling INFOSOLO OC, which indicates that the predictions based on this spatial model will rely on the contribution of the spatially correlated residuals. For pH, there were no differences in terms of the magnitude of the modelled regression coefficients but the significant predictors were soil texture and soil type for INFOSOLO, and parent material for LUCAS. A similar analysis was done for CEC but, for this soil property, clay was the most significant predictor for both datasets.

The significant predictors for each soil property were then used to fit omnidirectional variograms to the calibration subsets. The Matérn function (a generalization of several theoretical variogram functions incorporating a smoothness parameter; Minasny and McBratney, 2005) was chosen as the suitable model for REML fitting. The estimated variogram parameters are presented in Table 7. Results showed that, regardless the dataset considered, spatial correlation existed up to 25 km for OC, and 15 km for pH and CEC. The Nugget to Sill ratio (NSR) included in Table 7 was below 25%, which indicates that the model was able to capture spatial correlation reasonably well even for shorter distances (Cambardella et al., 1994).

3.4.1.3. Model validation. Table 9 summarizes the independent validation statistics obtained when using the EBLUP spatial model. These results showed small RMSE values with similar magnitudes for both INFOSOLO and LUCAS. Regarding the median θ(x) results, the spatial model used to predict OC performed poorly regardless the

Table 6
Summary of estimated fixed effect terms.

Model	Regression coefficients					
	OC		pH		CEC	
	INFOSOLO	LUCAS	INFOSOLO	LUCAS	INFOSOLO	LUCAS
Intercept	− 1.6	− 5.2	5.3	10.4	− 4.3	0.7
Avg. rainfall	0.001	0.09	− 0.0009	− 0.001	0.0006	0.001
Avg. temperature	–	0.3	0.05	–	0.0004	–
Elevation	0.0009	0.002	–	− 0.0008	–	–
Clay	0.04	–	0.06	–	0.1	0.6
Silt	0.02	0.3	− 0.03	− 0.03	0.07	–
Sand	–	–	–	− 0.05	0.05	0.04
Soil texture (Fine)	–	–	0.4	–	–	–
Soil type (Fluvisol)	–	–	0.5	–	–	–
Soil type (Umbrisol)	–	–	0.6	–	–	–
Parent material (sedimentary rocks)	–	–	–	− 0.4	–	–

Table 7
Summary of the estimated variogram parameters for the spatially correlated residuals.

Parameter	OC		pH		CEC	
Dataset	INFOSOLO	LUCAS	INFOSOLO	LUCAS	INFOSOLO	LUCAS
NSR (%)	22	0	8	0	22	4
Distance (km)	25	25	15	15	15	15

dataset used. Nevertheless, mean θ(x) was very close to 1.0, which can be interpreted as a relatively good model fit for OC (Johnson et al., 2017). For the other soil properties, the mean θ(x) was also close to 1 and the median θ(x) close to 0.455, with the best results obtained for independent validation using the LUCAS spatial model.

3.4.1.4. Predicted maps. Following the validation results, the spatial model was used to obtain the EBLUP OC, pH and CEC, using the INFOSOLO and LUCAS datasets. The national 1 km resolution maps are presented in Figs. 10, 11, and 12 (referred to as "Covariates"). A summary of the EBLUP values is presented in Table 10. Compared with the data distribution statistics presented before for the sampled OC (Fig. 6), the statistics for the INFOSOLO EBLUP predictions were closer to the experimental data, in terms of the mean, median and standard deviation. For pH and CEC, the statistics describing the distribution of the EBLUP predicted values were very close to the statistics for the experimental data, regardless the dataset considered.

3.4.2. Spatial modelling without environmental covariates

3.4.2.1. Model calibration. As part of the process to present the most accurate map for OC, pH and CEC spatial distribution, a second modelling approach which excludes the contribution of environmental covariates was tested. For this approach, the spatial model was derived from the raw data and the predictions were obtained using ordinary kriging (OK).

The experimental omnidirectional variogram obtained for the calibration subset (not substantially different from the national variograms presented in Fig. 8) was fitted with an exponential model using a weighted least squares criteria (Webster and Oliver, 2007). This spatial model was deemed the most parsimonious to describe spatial continuity. The corresponding parameters of the variogram models are presented in Table 8. Results showed that OC and CEC models determined greater continuity ranges for the INFOSOLO dataset when compared to the ranges obtained for LUCAS (around 15 km difference between ranges). The nugget effect was significant in the CEC model, but negligent for OC. In spite of the modelling efforts, the spatial model for pH delivered considerably low continuity ranges using both INFOSOLO and LUCAS, thus not being able to provide a model for distances above 2 km.

3.4.2.2. Model validation. Table 9 summarizes the validation statistics obtained when using the OK spatial model with the calibration and independent validation subsets. The results for independent validation showed that the RMSE values had the same magnitude for both INFOSOLO and LUCAS subsets, and were greater for OC validation but acceptable for pH and CEC. Overall the mean and median θ(x) results diverted from the reference values and indicated that the spatial model used to predict OC, pH and CEC performed poorly regardless the dataset used. This was expected for pH but also confirmed for OC and CEC in spite of reasonable model fitting performance based on the median θ(x) results obtained for the calibration dataset.

3.4.2.3. Predicted maps. The national 1 km resolution maps displaying OK predicted OC, pH and CEC values are presented in Figs. 10, 11, and 12 (referred to as "Without Covariates"). A summary of the OK predicted values is presented in Table 10. Compared with the data distribution statistics presented before for the sampled OC (Fig. 6), the statistics for the INFOSOLO OK predictions were closer to the experimental data in terms of the mean, median, and standard deviation. For pH and CEC, the variability in the predicted data distribution was lower compared to the experimental data for both datasets.

3.4.3. Comparison of spatial modelling outputs

3.4.3.1. Comparison of EBLUP and OK approaches. Based on the independent validation statistics, the EBLUP spatial model described more accurately OC, pH, and CEC variability. The differences between EBLUP and OK were more evident for pH and CEC likely due to the fact that the spatial model incorporating the covariates was able to better capture pH and CEC variability at the national level compared to the spatial model using raw data. As noted before for the spatial continuity analysis, the global variogram obtained with pH and CEC data was affected by the high values sampled for these properties in the south region.

We have also calculated the prediction variances and, as expected, EBLUP variances were generally lower than the ones obtained for OK. Additionally, the variance pattern for OK matched the sampling configuration, with smaller kriging variances close to the sampling points and larger variances in the areas where no data existed. These results and the importance of incorporating soil-related covariates in prediction were reported in other soil science studies (e.g., Chai et al., 2008; Minasny and McBratney, 2007).

Hence, to further compare the INFOSOLO and LUCAS modelling outputs we focused on the EBLUP maps presented on Figs. 10, 11, and 12 (referred to as "Covariates").

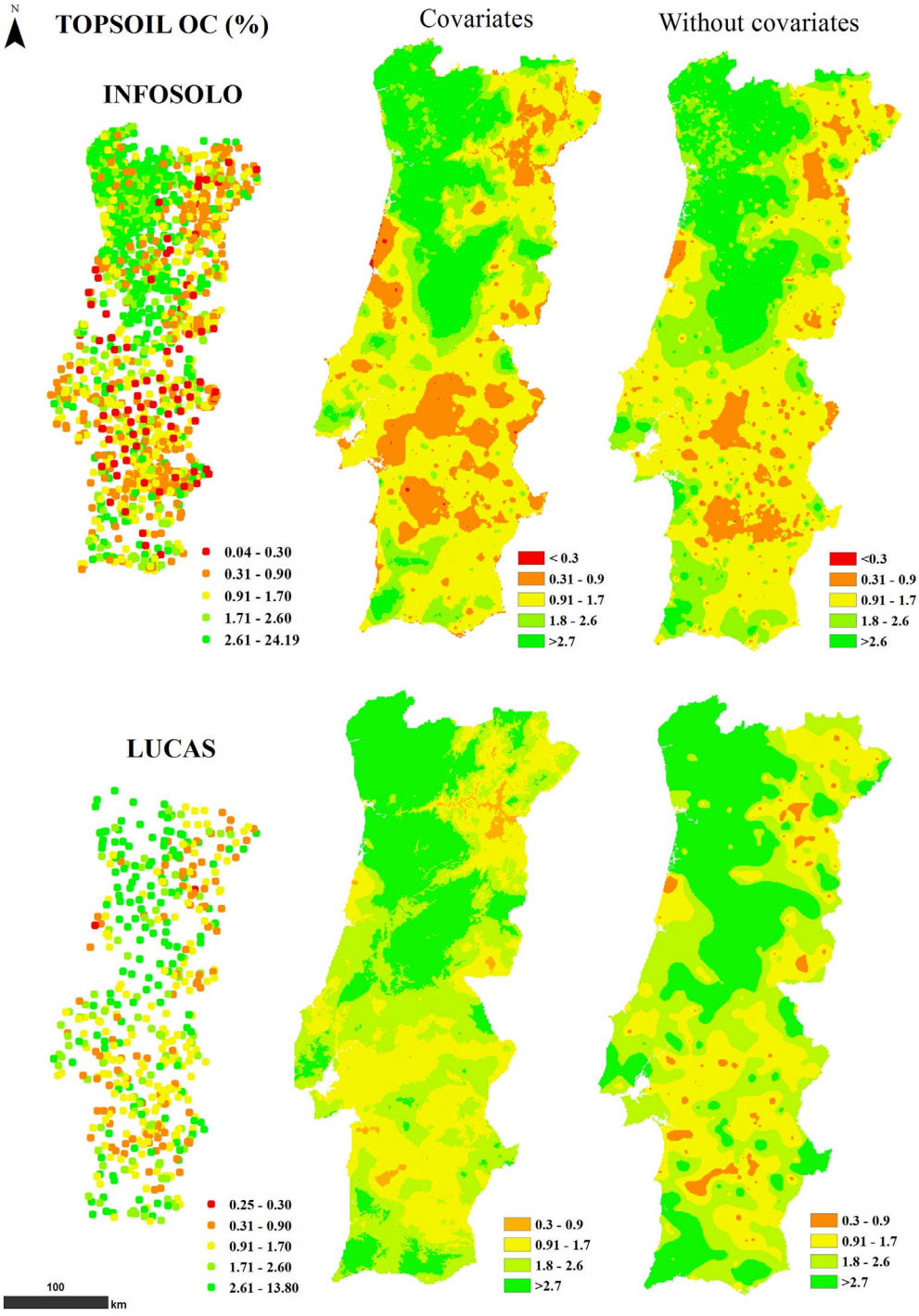

Fig. 10. OC national soil map (comparison between INFOSOLO and LUCAS, and modelling approaches – EBLUP + covariates prediction under "Covariates", kriging prediction under "Without covariates").

3.4.3.2. Comparison of INFOSOLO and LUCAS. To evaluate which dataset delivered a reliable representation of OC, pH, and CEC spatial patterns, we firstly analysed the covariates used in the spatial model and compared the distribution of high and low values.

Although the EBLUP covariates selected for modelling OC were different for INFOSOLO and LUCAS, they still related with the same biophysical variables, namely, climate and soil texture which were found relevant to explain OC variability in Portugal. As shown in Fig. 10, the patterns depicted in both INFOSOLO and LUCAS maps followed the distribution of the sampled values. The broad patterns were represented the same way with both datasets, namely the contrast between high and low OC content in the NW and NE regions, and in the north and south regions. Both maps depicted areas with higher OC content in the SW corner, in the central region and in the west coast.

However, in the INFOSOLO map, low OC areas (0.3% - 0.9%) had greater spatial extent namely in the south since low OC values were less represented in the LUCAS sampling campaign for this region. These areas of low OC are of great importance since they relate with soil quality degradation in the south, specifically in the left margin of the Guadiana River (Rosário, 2004). Also, the spatial patterns for areas with OC content > 0.9% were contoured differently in the two maps probably due to the contribution of the covariates (e.g., the patch of high values stretching from the north to the central region and in the south end appeared smoother in the LUCAS map).

To quantify the dissimilarity between INFOSOLO and LUCAS predictions as well as to identify where it occurred, we mapped the relative

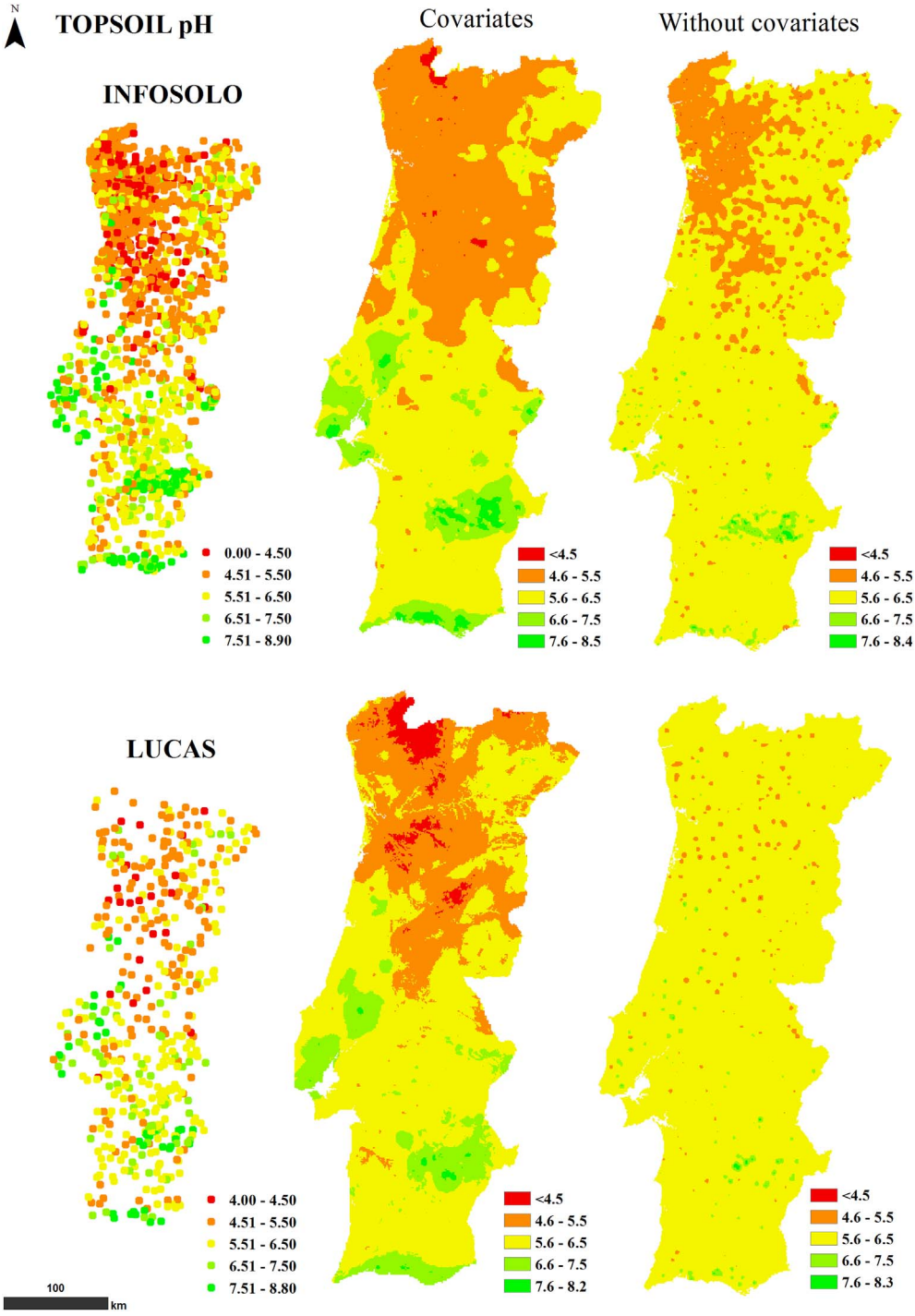

Fig. 11. pH national soil map (comparison between INFOSOLO and LUCAS, and modelling approaches – EBLUP + covariates prediction under "Covariates", kriging prediction under "Without covariates).

difference between the two maps as shown in Fig. 13. The values shown represent the magnitude of deviations between the LUCAS predicted value and the INFOSOLO value for that location in the grid. Negative values indicate greater predicted values for LUCAS whereas a positive value is obtained when LUCAS predictions are lower than INFOSOLO. Deviations between 10 and 20% were considered relevant. To display theses deviations, we have classified the relative difference maps using the 5, 25, 50, 75 and 95 percentile values.

The differences mapped for topsoil OC showed the highest deviations in the central and north regions (namely in the NW region where INFOSOLO had a high sampling density). In the south region, deviations were also significant namely the negative deviation in the left margin of the Guadiana River for which the LUCAS predictions varied considerably from INFOSOLO. Overall, 80% of the prediction grid area showed significant deviations between LUCAS and INFOSOLO predictions.

For pH, the selected covariates were also different for INFOSOLO and LUCAS EBLUP. However, they also represent the same biophysical variables related to pH variability in the soil (namely, soil texture and parent material which may be directly related to soil type). Comparing INFOSOLO and LUCAS maps, it can be seen that the same patterns were identified representing the changes from low to high values, from north to south, as displayed in the original sampled data. The cluster of high values in the south region was well represented with both datasets. However, the low pH areas had lesser spatial extent in the LUCAS map, with the patterns depicted with more detail. The relative difference

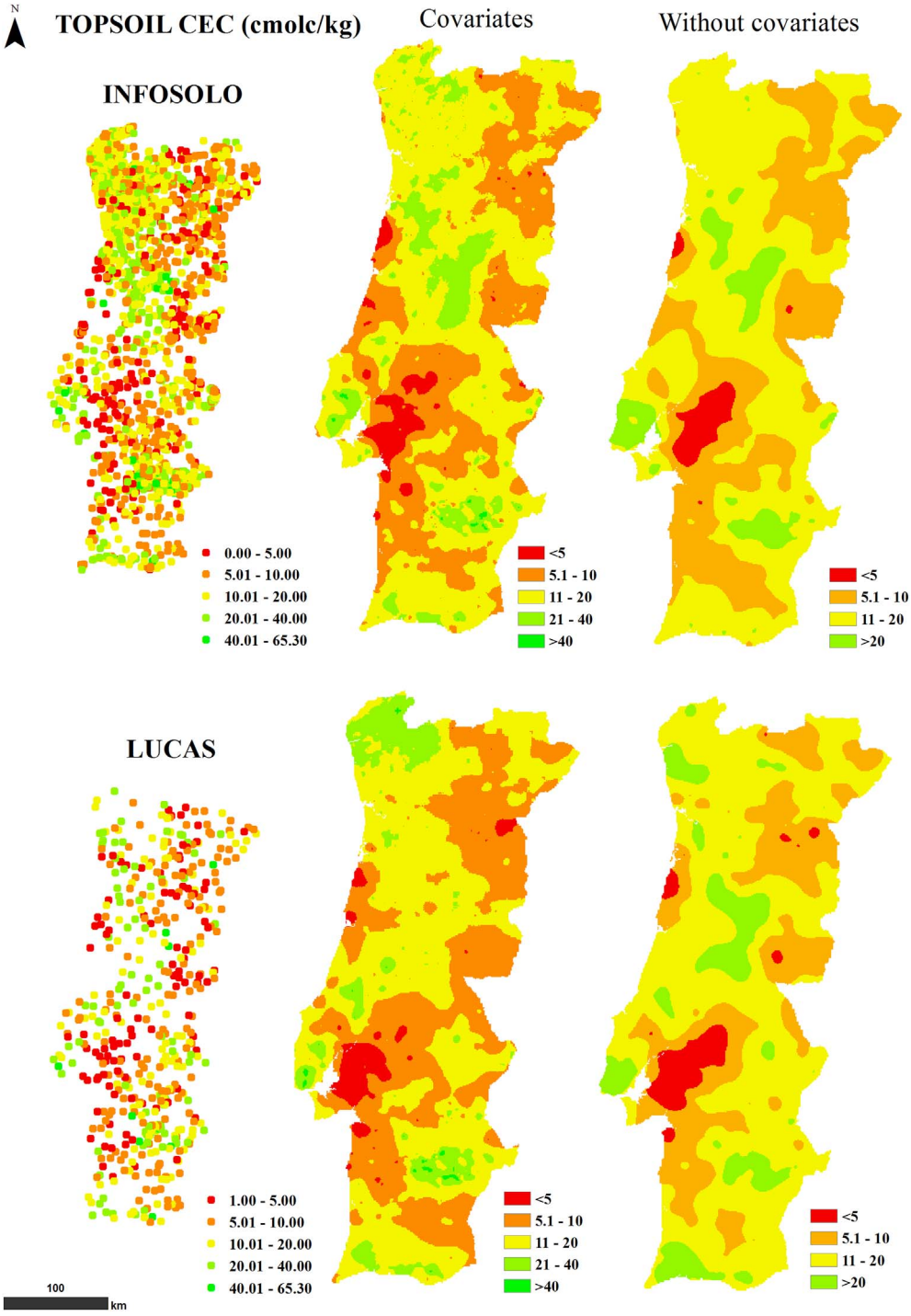

Fig. 12. CEC national soil map (comparison between INFOSOLO and LUCAS, and modelling approaches – EBLUP + covariates prediction under "Covariates", kriging prediction under "Without covariates).

Table 8
Summary of the variogram parameters (raw data).

Parameter	OC		pH		CEC	
Dataset	INFOSOLO	LUCAS	INFOSOLO	LUCAS	INFOSOLO	LUCAS
NSR (%)	0.6	0	30	0	44	40
Distance (km)	25.5	10	1.8	1.9	35	18

map displayed in Fig. 13 for topsoil pH showed a fairly good agreement between INFOSOLO and LUCAS predictions (only 25% of the prediction grid area showed relevant deviations). Pockets of relevant deviations were observed in the western coastal area, the top north, and also in the south, specifically in the area identified as a cluster of high values of clay and carbonate content and with a higher sampling density in INFOSOLO.

Clay was the covariate that most contributed to explain CEC variability for both INFOSOLO and LUCAS EBLUP, which makes sense considering that CEC mostly refers to the ionic exchange capacity of clay minerals. Both maps delivered the same general pattern for CEC, with low values located preferentially in the west and a cluster of higher values in the south (in agreement with the sampled data). One of the main differences in the predicted maps was the patch of high CEC values depicted only in the central region of the INFOSOLO map. This was evident in the differences map shown in Fig. 13, which also showed

Table 9
Independent validation (EBLUP and OK).

	OC		pH		CEC	
	INFOSOLO	LUCAS	INFOSOLO	LUCAS	INFOSOLO	LUCAS
EBLUP:						
RMSE	0.63	0.59	0.65	0.68	0.47	0.64
θ(x) (mean)	0.98	0.91	1.10	0.99	1.00	1.00
θ(x) (median)	0.29	0.29	0.34	0.45	0.39	0.46
OK:						
RMSE	1.5	1.8	0.83	0.94	0.52	0.77
θ(x) (mean)	1.7	1.3	1.68	1.46	1.5	0.9
θ(x) (median)	0.23	0.26	0.61	0.80	0.52	0.38

Table 10
Summary of the EBLUP and OK predictions.

Model	Minimum	Maximum	Median	Mean	St. Deviation
EBLUP:					
OC (INFOSOLO)	0.08	13.37	1.40	1.96	1.45
OC (LUCAS)	0.54	87.66	3.08	2.01	4.08
pH (INFOSOLO)	4.1	8.5	5.6	5.7	0.68
pH (LUCAS)	3.0	8.3	5.8	5.8	0.67
CEC (INFOSOLO)	2.4	62.7	11.4	12.5	5.65
CEC (LUCAS)	1.3	74.4	11.0	12.3	6.00
OK:					
OC (INFOSOLO)	0.16	16.1	1.6	2.0	1.2
OC (LUCAS)	0.3	12.9	2.1	2.6	1.4
pH (INFOSOLO)	4.3	8.4	5.6	5.6	0.3
pH (LUCAS)	4.1	8.5	5.9	5.9	0.17
CEC (INFOSOLO)	2.8	34.2	11.4	12.4	4.9
CEC (LUCAS)	2.1	40.0	12.2	13.0	5.3

an area of relevant deviations in the NW region and western coastal area due to lower INFOSOLO predictions. These areas presented different sampling densities for LUCAS and INFOSOLO, and the western area was less sampled in both datasets which explains the consistent deviations found for all soil properties.

3.4.3.3. Expert evaluation. It is also relevant to interpret the INFOSOLO and LUCAS EBLUP maps using expert knowledge to evaluate, from a pedological perspective, the quality of the predicted OC, pH, and CEC spatial patterns.

Based on expert interpretation of the maps presented in Figs. 10, 11, and 12, we consider that the INFOSOLO EBLUP maps reproduced better what is expected for the spatial distribution of OC, pH, and CEC in Portugal. This is particularly true for OC, where the LUCAS maps produced higher values than the expected for areas in the south, NE, and western coastal regions. For example, the left margin of the Guadiana River exhibited high OC values ($> 1.8\%$) in the LUCAS map which are not in accordance with an area where summer average temperatures are normally very high, average rainfall is the lowest, and where many desertification indexes have been pointing out the region as the most threatened in the country (Fig. 4; Perez-Trejo, 1992; Kosmas et al., 1999; Rosário, 2004). Furthermore, the INFOSOLO topsoil OC map was also more in agreement with the ones produced by Jones et al. (2005) and Lugato et al. (2014), following different modelling techniques. For pH, the INFOSOLO map represented the lower values ($pH < 4.5$) more realistically than the LUCAS map, especially in the north region, but also in the western coastal area, in agreement with the map produced by Freitas (1984) (Fig. 14), which grouped soils (20,000 soil samples) into different pH classes based on their reaction and on the soil associations of the Soil Map of Portugal at 1:1000000 scale (Cardoso et al., 1973). For CEC, the LUCAS and INFOSOLO maps exhibited many similarities. The main difference was the overestimation of LUCAS CEC topsoil values in the NW region and underestimation in the central region relatively to INFOSOLO.

3.4.3.4. Comparison with SoilGrids. Finally, Fig. 15 displays the predicted INFOSOLO EBLUP maps for OC, pH and CEC with the results presented in Hengl et al. (2017), which provided 250 m resolution maps for these soil properties but at the continental scale (SoilGrids product). SoilGrids is currently the only digital soil map available for Portugal. Hengl et al. (2017) improved the modelling approach carried out in Hengl et al. (2014) and included additional soil databases for data analyses, particularly the LUCAS topsoil survey dataset, which for Portugal provided more coverage than the WISE dataset used previously. The modelling approach also used the contribution of covariates to explain the variability of soil properties and the results indicated climatic and biomass indices (extracted from satellite imagery), topography, lithology, land cover, and soil type as the most relevant for global modelling.

Fig. 13 further displays the comparison between INFOSOLO and SoilGrids (at 30 cm depth) by quantifying the relative differences between predictions obtained for both maps (using the same procedure applied to compare INFOSOLO and LUCAS maps). The greatest deviations were obtained for topsoil OC and CEC (SoilGrids predictions varied considerably from INFOSOLO in 75% and 55% of the prediction grid area, respectively). Interestingly, SoilGrids deviations from INFOSOLO were opposite to the ones encountered for LUCAS. For example, OC deviations were positive in the NW region and in the south. Differences in OC predictions in the left margin of the Guadiana River were less pronounced, hinting higher predicted INFOSOLO values. The deviation pattern in the western coastal area, central region and central south (matching the cluster of sampled values) was highlighted for both SoilGrids and LUCAS.

The deviations observed for topsoil pH were scattered along the country, with small patches of negative deviations from the north to the central region, and noticeable areas of positive deviations in the south (namely in in the cluster of high values). For topsoil CEC, the most significant patch of negative deviations was found in the western coastal area and the left margin of the Tagus River (which was classified as a positive deviation in the LUCAS comparison). There was a significant deviation patch in the central region, also identified for LUCAS.

The spatial resolution of SoilGrids, but also the modelling approach and the covariates used, explain why the comparison of SoilGrids with INFOSOLO was not completely in agreement with the results obtained when comparing INFOSOLO and LUCAS. However, the INFOSOLO comparison with LUCAS and SoilGrids highlighted consistently deviations in the top NW region and the western coastal region. Both these areas were either not sufficiently covered by the latter products, with the INFOSOLO sampling density being here considerably higher than LUCAS. There was also a central area in the central region (with low temperatures and high elevation and slope) which showed differences, namely for OC and CEC. The sampling cluster in the south also produced mostly positive deviations (hence higher INFOSOLO predictions) for the three soil properties, which were evident when comparing INFOSOLO, LUCAS and SoilGrids maps.

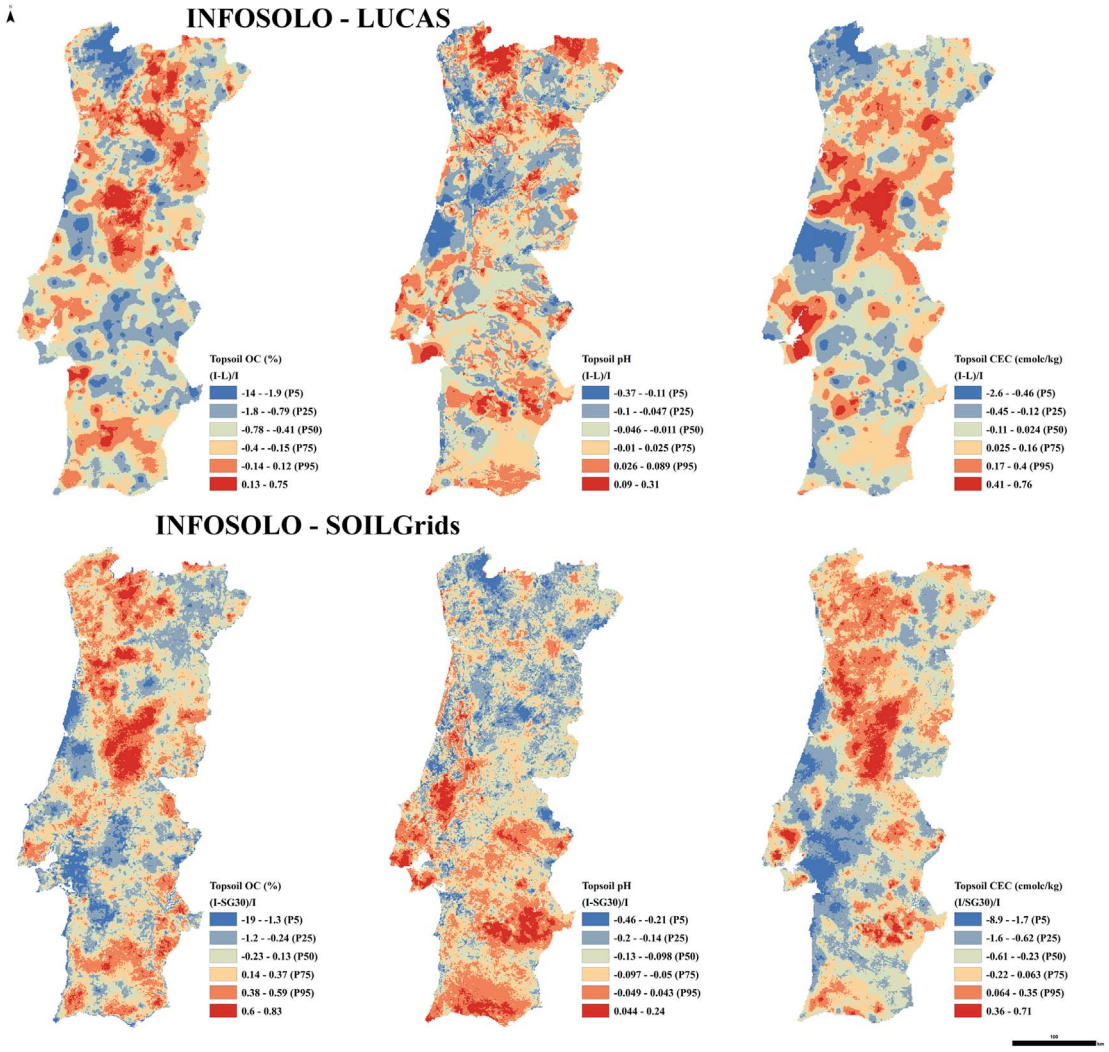

Fig. 13. Comparison of INFOSOLO, LUCAS and SoilGrids (at 30 cm depth) predictions for OC, pH and CEC.

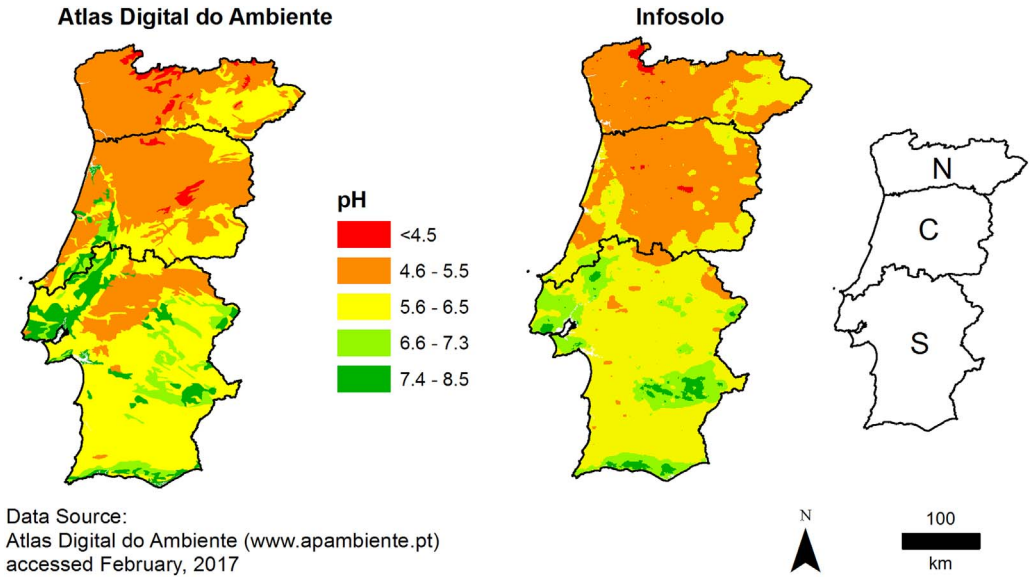

Fig. 14. Comparison of pH maps (validation using expert knowledge).

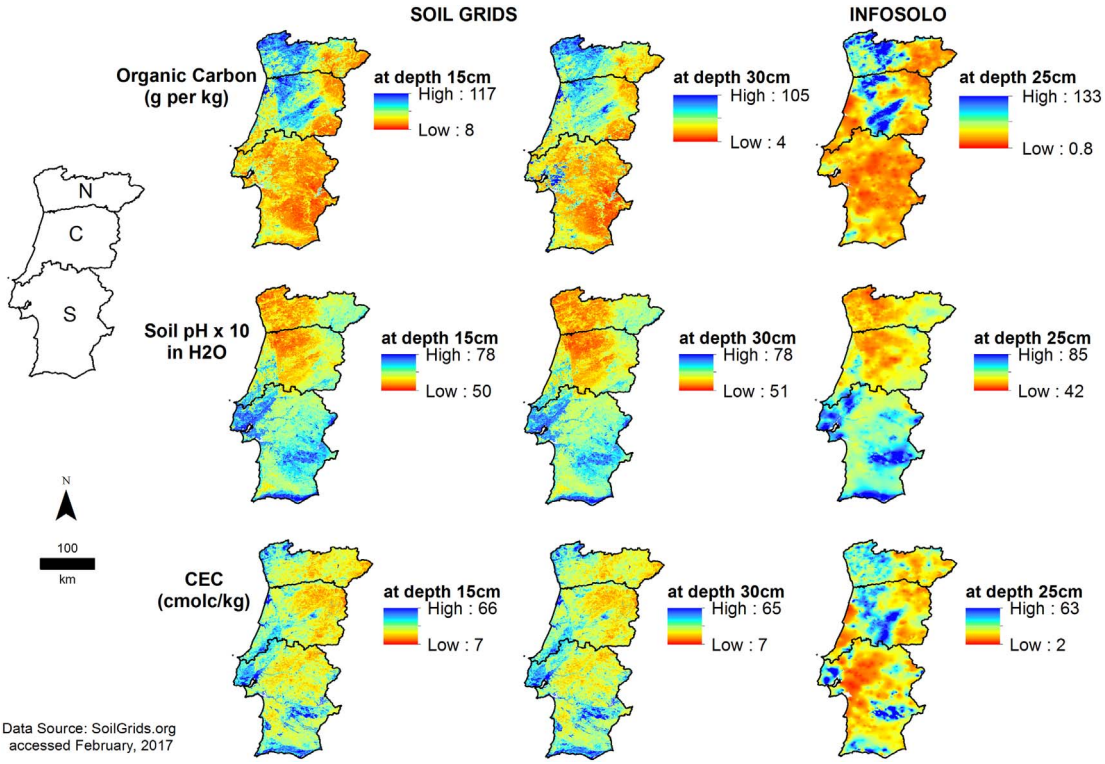

Fig. 15. SoilGrids digital maps for Portugal.

4. Conclusions and future developments

This work presents a comprehensive overview of the INFOSOLO dataset and highlights the importance of compiling Portuguese soil data into an unique database to be used as a baseline for future soil management policies. This work also represents the most comprehensive effort ever made to organize soil information in Portugal, with the database currently including physical and chemical characteristics of 9934 horizons/layers studied in 3461 soil profiles across the country. INFOSOLO was thoroughly validated using a sound process of quality assurance and harmonization to deliver the most valuable repository of soil data collected in the past six decades in Portugal.

The number of sampling points included in INFOSOLO is superior to any other dataset currently available for Portugal, namely, the EU-wide LUCAS, adding 2932 locations to the 465 LUCAS points sampled in the country. But to consider INFOSOLO a better alternative to LUCAS, it is important to compare these datasets starting with the fact that INFOSOLO sampling did not comply with any statistical design which was done for LUCAS. This explains the existence of a highly sampled area in the NW region of Portugal, and of the cluster located in the central part of the south region (Fig. 1). As a consequence, INFOSOLO data is characterized by a higher OC content and lower pH in the north region. Also, in the south, INFOSOLO displays lower OC values. This specifically is a significant difference between the two datasets since low OC content characterizes most of the southern Portugal due to its edapho-climatic characteristics. The INFOSOLO cluster located in the south also explains the high pH and CEC values not represented in LUCAS.

Besides sampling density, it is the configuration of the sampling points that can determine distinctive continuity patterns and this is, ultimately, the most important aspect to understand the spatial variability of OC, pH, and CEC. The continuity patterns obtained for INFOSOLO and LUCAS data are very similar and allow us to conclude that INFOSOLO and LUCAS datasets are not substantially different. Hence, in spite of not being the product of a carefully designed campaign, INFOSOLO can be used to produce quality soil information. This was tested further in this work by conducting spatial modelling using kriging-based prediction to map the spatial distribution of OC, pH and CEC at the national level.

The spatial continuity analysis results were an early indication that the use of soil related environmental covariates could improve spatial prediction, especially for pH and CEC. Indeed, based on independent validation, we have concluded that the spatial modelling approach incorporating covariates produced a spatial model able to deliver accurate predictions. This conclusion was valid for both datasets and for all soil properties. Hence, we cannot differentiate INFOSOLO and LUCAS based on the accuracy of the spatial model used for prediction. But we can evaluate the similarity between their maps to conclude if using INFOSOLO is indeed relevant to characterize OC, pH and CEC at the national level.

Both INFOSOLO and LUCAS maps can reproduce broadly OC, pH, and CEC variability and, depending on the purpose and the scale of the study, it wouldn't be relevant which one to use. But the predictions provided by both maps can be significantly different. Significant deviations between LUCAS and INFOSOLO predictions were found to occur consistently in the north (particularly in the NW region), in the western coastal area, in the central part of the central region, and in the south (namely in the area of clustered sampling). In the north, west coast and in the south cluster, INFOSOLO sampling density is considerably higher than LUCAS which we consider to improve the predictions, namely, for pH and CEC in the south region. The western coastal area is unevenly covered by both datasets therefore differences in predictions are the result of the spatial model used. The central area in the central region presents specific climate and terrain conditions (low temperatures and high elevation and slope) likely incorporated differently in the spatial models, resulting in differences in predictions, namely for OC and CEC. It is also important to note the deviations between LUCAS and INFOSOLO in terms of OC prediction. Since INFOSOLO incorporates lower OC values for the south, the differences between LUCAS and INFOSOLO predictions are evident, quantitatively and qualitatively.

Finally, to complete this evaluation, we have also compared our

INFOSOLO maps with SoilGrids, which maps soil properties at a global scale using available datasets from different countries. The predictions for Portugal were derived using LUCAS. Overall, the deviations observed between the two maps were similar to the ones noted for LUCAS.

Hence, based on our evaluation, INFOSOLO is a database capable to characterize with accuracy the spatial distribution of soil properties and thus to provide reliable soil information to inform future soil management and planning policies. It represents an important step towards the development of a soil information system in Portugal. Despite that, there is still much to be done:

- The database must be integrated in a Web-based digital platform that will make georeferenced soil information and derived soil properties maps freely available for land managers, scientists, policy decision makers, and students;
- Other soil properties that may be considered relevant can be easily included in the database. However, currently such information will hardly show a reasonable distribution throughout Portugal since the database holds already the most common soil properties found in the available literature;
- Soil profiles descriptions can also be included in the database. But, like for soil data, the quality of soil descriptions varies between studies. Also, most of the existing descriptions show subjective and qualitative criteria which are difficult to harmonize;
- The soil information obtained using different methodologies need most likely also to go through a harmonization process. Weynants et al. (2013) described the harmonization process carried out in the EU-HYDI. A similar process may prove also to be necessary in INFOSOLO, particularly for soil organic content which was determined using seven different methodologies. However, the conversion factors used to harmonize organic carbon content in Weynants et al. (2013) should be carefully revised since they were developed for very different edapho-climatic regions;
- The considerable amount of soil information still available in Portuguese Universities and Polytechnics dedicated to soil science should also be included in the database. Currently, data from these institutions reaches merely 2.6% of the database content, with most being found online;
- The covariates used for modelling show that it is possible to build a spatial model to be used in the future with minimum requirements for sampling. This, together with the regular updating of the INFOSOLO dataset will make it possible to provide up-to-date and accurate soil information. A soil monitoring program aiming to validate future predicted maps should be incorporated in the modelling effort;
- There is also the need to produce maps for the subsoil horizons/layers, for the remaining soil properties included in the database (e.g., total N, extractable P, extractable K, and $CaCO_3$ content), as well as derived properties (e.g., available water capacity);
- Finally, INFOSOLO is just a first step into valuing and preserving national legacy soil data. This should be an on-going effort at the research and governmental level to ensure the preservation of Portuguese soils.

The database offers a large potential to provide solutions for different regional and national environmental challenges, namely for counteracting soil degradation at different scales, for improving watershed management, for assessing the role of soils in climate change mitigation, and for valuing soil ecosystem services. It is also a convenient mean of educating students, providing the opportunity of handling a large quantity of new and reliable soil information, and studying the relationships between soil quality, climate, topography, and land management in different regions in the country.

While the database represents a first step towards the development of a much needed modern soil information system in Portugal, it will require the combined effort of many more soil scientist in order to improve the quality of the data available, its distribution throughout the country, and the quality of related outputs, namely, the soil maps produced. Nonetheless, the database is an important tool for raising awareness of the general public, land users, stakeholders, and policy decision makers about the importance of soils as natural resources and their relations to human welfare and sustainability.

Acknowledgments

A special thanks is due to “Direcção Geral de Agricultura e do Desenvolvimento Rural”, namely to Miguel Pereira and Manuel Frazão, for supplying the coordinates of many soil profiles studied in that institution. Authors would like to acknowledge also Raquel Mano from “Instituto Nacional de Investigação Agrária e Veterinária” for providing the BioSoil dataset for Portugal. The LUCAS topsoil dataset used in this work was made available by the European Commission through the European Soil Data Centre managed by the Joint Research Centre (JRC), http://esdac.jrc.ec.europa.eu/. The digital terrain model used in this work was produced using Copernicus data and information funded by the European Union - EU-DEM layers. MARETEC acknowledges the national funds from the Foundation for Science and Technology (FCT) (Project UID/EEA/50009/2013). T. B. Ramos was supported by the FCT grant SFRH/BPD/110655/2015.

References

Agroconsultores and Geometral, 1999. Carta de solos e carta de aptidão da terra para a agricultura (1:25.000) em Entre Douro e Minho. Direcção Regional de Agricultura de Entre Douro e Minho. Ministério da Agricultura, Pescas e Florestas, Braga.

Aksoy, E., Yigini, Y., Montanarella, L., 2016. Combining soil databases for topsoil organic carbon mapping in Europe. PLoS ONE 11 (3), e0152098. http://dx.doi.org/10.1371/journal.pone.0152098.

Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B.M., Hong, S.Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J., Mendonça-Santos, M.D.L., Minasny, B., Montanarella, L., Odeh, I.O.A., Sanchez, P.A., Thompson, J.A., Zhang, G.L., 2014. GlobalSoilMap: toward a fine-resolution global grid of soil properties. Adv. Agron. 125, 93–134.

Arrouays, D., Leenaars, J.G.B., Richer-de-Forges, A.C., Adhikari, K., Ballabio, C., Greve, M., et al., 2017. Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. GeoResJ 14, 1–19.

Ballabio, C., Panagos, P., Montanarella, L., 2016. Mapping topsoil physical properties at European scale using the LUCAS database. Geoderma 261, 110–123.

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.

Batjes, N.H., Ribeiro, E., van Oostrum, A., Leenaars, J.G.B., Hengl, T., Mendes De Jesus, J., 2016. WoSIS: providing standardised soil profile data for the world. Earth Syst. Sci. Data 9, 1–14. http://dx.doi.org/10.5194/essd-9-1-2017.

Bishop, T.F.A., Horta, A., Karunaratne, S.B., 2015. Validation of digital soil maps at different spatial supports. Geoderma 241, 238–249.

Blum, W.E.H., 2005. Functions of soil for society and the environment. Rev. Environ. Sci. Biotechnol. 4, 75–79.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Cambardella, C.A., Moorman, T.B., Parkin, T.B., Karlen, D.L., Novak, J.M., Turco, R.F., Konopka, A.E., 1994. Field-scale variability of soil properties in central Iowa soils. Soil Sci. Soc. Am. J. 58, 1501–1511.

Cardoso, J.C., 1965. Os solos de Portugal. Sua classificação, caracterização e génese. I – A sul do Rio Tejo. Direcção Geral dos Serviços Agrícolas, Lisboa.

Cardoso, J.C., 1974. A classificação dos solos de Portugal – nova versão. In: Boletim de Solos. 17. pp. 14–46.

Cardoso, J.C., Bessa, M.T., Marado, M.B., 1973. Carta de solos de Portugal (1/1000000). Agron. Lusit. 33, 481–602.

Carré, F., McBratney, A., Minasny, B., 2007. Estimation and potential improvement of the quality of legacy soil samples for digital soil mapping. Geoderma 141, 1–14.

Chai, X., Shen, C., Yuan, X., Huang, Y., 2008. Spatial prediction of soil organic matter in the presence of different external trends with REML-EBLUP. Geoderma 148, 159–166.

Costa, J.B., 1979. Caracterização e constituição do Solo, 7th edition. Fundação Calouste Gulbenkian, Lisboa.

De Vos, B., Cools, N., 2011. Second European forest soil condition report. In: Results of the BioSoil Soil Survey. INBO.R.2011.35. vol. 1 Research Institute for Nature and Forest, Brussel, Belgium.

DGADR, 2007. Extensão do estudo de caracterização dos solos e esboço de aptidão das terras para o regadio à escala 1:25.000 aos blocos de Serpa e Enxoé da área a beneficiar com o Empreendimento de Fins Múltiplos de Alqueva. Direcção Geral de Agricultura e Desenvolvimento Rural. Ministério da Agricultura, do Desenvolvimento Rural e das Pescas, Lisboa.

Dias, J.C.S., 2000. Manual de fertilização das culturas. Instituto Nacional de Investigação Agrária. Ministério da Agricultura e do Desenvolvimento Rural e das Pescas, Lisboa.

Divisão de Solos, 2003. Estudo de caracterização dos solos e esboço de aptidão das terras para o regadio à escala 1:25.000 na área a beneficiar com o Empreendimento de Fins Múltiplos de Alqueva. Relatório Final. Instituto de Desenvolvimento Rural e Hidráulica, Ministério da Agricultura, do Desenvolvimento Rural e das Pescas, Lisboa.
Dominati, E., Patterson, M., Mackay, A., 2010. A framework for classifying and quantifying the natural capital and ecosystem services of soils. Ecol. Econ. 69, 1858–1868.
ESRI, 2014. ArcGIS Desktop: Release 10.2.2. Environmental Systems Research Institute, Redlands, CA.
Eswaran, H., Lal, R., Reich, P.F., 2001. Land degradation: an overview. In: Bridges, E.M., Hannam, I.D., Oldeman, L.R., Pening de Vries, F.W.T., Scherr, S.J., Sompatpanit, S. (Eds.), Responses to Land Degradation. Proc. 2nd. International Conference on Land Degradation and Desertification, Khon Kaen, Thailand. Oxford Press, New Delhi, India.
EUROSTAT, 2009. LUCAS 2009 (land use/cover area frame survey). In: Technical Reference Document C-3: Land Use and Land Cover: Nomenclature. European Commission.
FAO, 2006. Guidelines for Soil Description, 4th edition. (Rome).
FAO, 2015. Soils Badge Challenge. Food and Agriculture Organization of the United Nations, Rome.
FAO/IIASA/ISRIC/ISS-CAS/JRC, 2012. Harmonized World Soil Database (version 1.2). FAO, Rome, Italy.
Franzluebbers, A.J., 2009. linking soil organic carbon and environmental quality through conservation tillage and residue management. In: Lal, R., Follett, R.F. (Eds.), Soil Carbon Sequestration and the Greenhouse Effect, SSSA Spec. Publ. 57. SSSA, Madison, WI, pp. 263–289.
Freitas, F.C., 1984. Acidez e alcalinidade dos solos. Notícia explicativa III.2. Atlas do Ambiente. Comissão Nacional do Ambiente, Lisboa.
Gee, G.W., Or, D., 2002. Particle-size analysis. In: Dane, J.H., Topp, G.C. (Eds.), Methods of Soil Analysis. Part 4. Physical Methods. SSSA Book Ser. 5. SSSA, Madison, WI, pp. 255–294.
Geometral and Agroconsultores, 2004. Elaboração da carta de solos e de aptidão das terras da Zona Interior Centro. Instituto de Desenvolvimento Rural e Hidráulica, Ministério da Agricultura, Pescas e Florestas, Lisboa.
Gomes, M.P., Silva, A.A., 1962. Um novo diagrama triangular para a classificação básica da textura do solo. In: Garcia da Orta. 10. pp. 171–179.
Gonçalves, M.C., Reis, L.C.L., Pereira, M.V., 2005. Progress of soil survey in Portugal. In: Jones, R.J.A., Houšková, B., Bullock, P., Montanarella, L. (Eds.), European Soil Bureau Research Report No. 9. Office for Official Publications of the European Communities, Luxembourg, pp. 275–279.
Gonçalves, M.C., Šimůnek, J., Ramos, T.B., Martins, J.C., Neves, M.J., Pires, F.P., 2006. Multicomponent solute transport in soil lysimeters irrigated with waters of different quality. Water Resour. Res. 42 (17), W08401. http://dx.doi.org/10.1029/2005WR004802.
Gonçalves, M.C., Ramos, T.B., Pires, F.P., 2011. Base de dados georreferenciada das propriedades do solo. In: Coelho, P.S., Reis, P. (Eds.), Agrorrural. Contributos Científicos. Instituto Nacional dos Recursos Biológicos, Oeiras, Portugal, pp. 564–574.
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.
Grunwald, S., 2009. Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma 152, 195–207.
Hartemink, A.E., 2015. On global soil science and regional solutions. Geoderma Reg. 5, 1–3.
Hartemink, A.E., Krasilnikov, P., Bockheim, J.G., 2013. Soil maps of the world. Geoderma 207-208, 256–267.
Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Gonzalez, M.G., 2014. SoilGrids1 km — global soil information based on automated mapping. PLoS ONE 9 (8), e105992. http://dx.doi.org/10.1371/journal.pone.0105992.
Hengl, T., Mendes de Jesus, J., Heuvelink, G.B.M., Gonzalez, M.R., 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. SoilGrids250 m: global gridded soil information based on machine learning. PLoS ONE 12 (2), e0169748. http://dx.doi.org/10.1371/journal.pone.0169748.
Horta, A., Bishop, T.F.A., Karunaratne, S.B., 2013. Model-based Geostatistics and machine-learning methods: a comparison in terms of estimates of prediction uncertainty. In: 10th Biennial Meeting of Commission 1.5 Pedometrics Division 1 of the International Union of Soil Science (IUSS). ICRAF & CIAT, Kenya. https://sites.google.com/a/cgxchange.org/pedometrics2013/.
Isaaks, E., Srivastava, R.M., 1989. Applied Geostatistics. Oxford University Press, New York.
IUSS Working Group, 2006. World reference base for soil resources 2006. A framework for international classification, correlation and communication. In: World Soil Resources Reports 103. Food and Agriculture Organization of the United Nations, Rome, Italy.
Jenny, H., 1941. Factors of soil formation: a system of quantitative pedology. In: Dover Books on Earth Sciences. Dover Publications.
Johnson, L.E., Bishop, T.F.A., Birch, G.F., 2017. Modelling drivers and distribution of lead and zinc concentrations in soils of an urban catchment (Sydney estuary, Australia). Sci. Total Environ. 598, 168–178.
Jones, R.J.A., Hiederer, R., Rusco, E., Montanarella, L., 2005. Estimating organic carbon in the soils of Europe for policy support. Eur. J. Soil Sci. 56, 655–671.
Karunaratne, S.B., Bishop, T.F.A., Baldock, J., Odeh, I.O.A., 2014. Catchment scale mapping of measureable soil organic carbon fractions. Geoderma 219, 14–23.
Kempen, B., Heuvelink, G.B.M., Brus, D.J., Stoorvogel, J.J., 2010. Pedometric mapping of soil organic matter using a soil map with quantified uncertainty. Eur. J. Soil Sci. 61, 333–347.
Kosmas, C., Kirkby, M., Geeson, N., 1999. The Medalus project Mediterranean desertification and land use. In: Manual on Key Indicators of Desertification and Mapping Environmentally Sensitive Areas to Desertification. EUR 18882. Science, Research and Development, Directorate-General, European Commission.
Kristensen, J.A., Breuning-Madsen, H., Balstrøm, T., 2015. SPADE-14. Suggested Corrections Based on the SPADE-8 Evaluation. European Commission Joint Research Centre.
Lambert, J.J., Darousin, J., Eimberck, M., Le Bas, C., Jamagne, M., King, D., Montanarella, L., 2003. Soil Geographical Database for Eurasia and the Mediterranean: Instruction Guide for Elaboration at Scale 1:1,000,000, Version 4.0. Tech. Rep. EUR 20422. European Commission Joint Research Centre, Italy.
Lark, R.M., 2000. Estimating variograms of soil properties by the method-of-moments and maximum likelihood. Eur. J. Soil Sci. 53 (51), 717–728.
Lark, R.M., 2012. Towards soil geostatistics. Spat. Stat. 1, 92–99.
Lark, R.M., Cullis, B.R., Welham, S.J., 2006. On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. Eur. J. Soil Sci. 57, 787–799.
Leenaars, J.G.B., Kempen, B., van Oostrum, A.J.M., Batjes, N.H., 2014a. Africa soil profiles database: a compilation of georeferenced and standardised legacy soil profile data for Sub-Saharan Africa. In: Arrouays, D., McKenzie, N., Hempel, J., Richer de Forges, A., McBratney, A.B. (Eds.), GlobalSoilMap - Basis of the Global Spatial Information System. Taylor & Francis group, London, pp. 51–57.
Leenaars, J.G.B., van Oostrum, A.J.M., Ruiperez Gonzalez, M., 2014b. Africa Soil Profiles Database, Version 1.2. A compilation of georeferenced and standardised legacy soil profile data for Sub-Saharan Africa (with dataset). In: ISRIC Report 2014/01. Africa Soil Information Service (AfSIS) Project. ISRIC - World Soil Information, Wageningen (162 pp.).
Li, H.Y., Webster, R., Shi, Z., 2015. Mapping soil salinity in the Yangtze delta: REML and universal kriging (E-BLUP) revisited. Geoderma 237–238, 71–77.
Li, H.Y., Marchant, B.P., Webster, R., 2016. Modelling the electrical conductivity of soil in the Yangtze delta in three dimensions. Geoderma 269, 119–125.
Liaw, A., Wiener, M., 2002. Classification and regression by random forests. In: R News. 2/3. pp. 18–22.
Lin, H., 2003. Hydropedology: bridging disciplines, scales, and data. Vadose Zone J. 2, 1–11.
Lin, H., 2010. Earth's critical zone and hydropedology: concepts, characteristics, and advances. Hydrol. Earth Syst. Sci. 14, 25–45. http://dx.doi.org/10.5194/hess-14-25-2010.
Lin, H., Bouma, J., Pachepsky, Y., Western, A., Thompson, J., van Genuchten, M.Th., Vogel, H.-J., Lilly, A., 2006. Hydropedology: synergistic integration of pedology and hydrology. Water Resour. Res. 42, W05301. http://dx.doi.org/10.1029/2005WR004085.
Lugato, E., Panagos, P., Bamba, F., Jones, A., Montanrella, L., 2014. A new baseline of organic carbon stock in European agricultural soils using a modelling approach. Glob. Chang. Biol. 20, 313–326.
Madeira, M., Constantino, A.T., Réffega, A.G., Martins, A.A., Alexandre, C.J., Sousa, E., Monteiro, F.G., Pinheiro, J.F., Cardoso, J.C., Silva, J.V., Ricardo, R.P., 2004. Bases para a revisão e actualização da classificação dos solos de Portugal. Sociedade Portuguesa de Ciência do Solo, Lisbon, Portugal.
McBratney, A., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117, 3–52.
Minasny, B., McBratney, A.B., 2005. The Matérn function as a general model for soil variograms. Geoderma 128, 192–207.
Minasny, B., McBratney, A.B., 2007. Spatial prediction of soil properties using EBLUP with the Matérn covariance function. Geoderma 140, 324–336.
Montanarella, L., 2007. Trends in land degradation in Europe. In: Sivakumar, M.V.K., Ndiang'ui, N. (Eds.), Climate and Land Degradation. Springer, pp. 83–104.
Nelson, D.W., Sommers, L.E., 1996. Total carbon, organic carbon, and organic matter. In: Sparks, D.L., Page, A.L., Helmke, P.A., Loeppert, R.H., Soltanpour, P.N., Tabatabai, M.A., Johnston, C.T., Sumner, M.E. (Eds.), Methods of Soil Analysis, Part 3. Chemical Methods. Soil Sci. Soc. Am. Inc., American Society of Agronomy Inc., Madison, WI, pp. 961–1010.
Nelson, M.A., Bishop, T.F.A., Triantafilis, J., Odeh, I.O.A., 2011. An error budget for different sources of error in digital soil mapping. Eur. J. Soil Sci. 62, 417–430.
Nemes, A., Wösten, J.H.M., Lilly, A., Oude Voshaar, J.H., 1999. Evaluation of different procedures to interpolate the cumulative particle-size distribution to achieve compatibility within a soil database. Geoderma 90, 187–202.
Oliver, M., Webster, R., 2014. A tutorial guide to geostatistics: computing and modelling variograms and kriging. Catena 113, 56–69.
Panagos, P., 2006. The European Soil Database. 5(7). GEO: connexionpp. 32–33.
Panagos, P., Van Liedekerke, M., Jones, A., Montanarella, A., 2012. European soil data centre: response to European policy support and public data requirements. Land Use Policy 29, 329–338.
Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30 (7), 683–691.
Perez-Trejo, F., 1992. Desertification and Land Degradation in the European Mediterranean. Report EUR 14850. Science, Research and Development, Directorate-General XII, European Commission (ISSN 1018-5593).
Ramos, T.B., Šimůnek, J., Gonçalves, M.C., Martins, J.C., Prazeres, A., Castanheira, N.L., Pereira, L.S., 2011. Field evaluation of a multicomponent solute transport model in soils irrigated with saline waters. J. Hydrol. 407, 129–144.
Ramos, T.B., Šimůnek, J., Gonçalves, M.C., Martins, J.C., Prazeres, A., Pereira, L.S., 2012.

Two-dimensional modeling of water and nitrogen fate from sweet sorghum irrigated with fresh and blended saline waters. Agric. Water Manag. 111, 87–104.
Ramos, T.B., Gonçalves, M.C., Brito, D., Martins, J.C., Pereira, L.S., 2013. Development of class pedotransfer functions for integrating water retention properties into Portuguese soil maps. Soil Res. 51, 262–277.
Ramos, T.B., Horta, A., Gonçalves, M.C., Martins, J.C., Pereira, L.S., 2014. Development of ternary diagrams for estimating water retention properties using geostatistical approaches. Geoderma 230, 229–242.
Rawlins, B.G., Marchant, B.P., Smyth, D., Scheib, C., Lark, R.M., Jordan, C., 2009. Airborne radiometric survey data and a DTM as covariates for regional scale mapping of soil organic carbon across Northern Ireland. Eur. J. Soil Sci. 60, 44–54.
Reich, P.F., Numbem, S.T., Almaraz, R.A., Eswaran, H., 2001. Land resource stresses and desertification in Africa. In: Bridges, E.M., Hannam, I.D., Oldeman, L.R., Pening de Vries, F.W.T., Scherr, S.J., Sompatpanit, S. (Eds.), Responses to Land Degradation. Proc. 2nd. International Conference on Land Degradation and Desertification, Khon Kaen, Thailand. Oxford Press, New Delhi, India.
Ribeiro, P.J., Diggle, P.J., 2001. geoR: a package for geostatistical analysis. In: R-News. 1 (2). pp. 15–18.
Robinson, D.A., Hockley, N., Dominati, E., Lebron, I., Scow, K.M., Reynolds, B., Emmet, B.A., Keith, A.M., de Jonge, L.W., Schjønning, P., Moldrup, P., Jones, S.B., Tuller, M., 2012. Natural capital, ecosystems, and soil change: why soil science must embrace and ecosystems approach. Vadose Zone J. 11. http://dx.doi.org/10.2136/vzj2011.0051.
Rosário, L., 2004. Indicadores de desertificação para Portugal Continental. Direcção Geral dos Recursos Florestais. Ministério da Agricultura, Desenvolvimento Rural e Pescas, Lisboa (56 pp.).
Samuel-Rosa, A., Heuvelink, G.B.M., Vasques, G.M., Anjos, L.H.C., 2015. Do more detailed environmental covariates deliver more accurate soil maps? Geoderma 243–244, 214–227.
Sanchez, P.A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendonça-Santos, M.L., Minasmy, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vägen, T.-G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., Zhang, G.-L., 2009. Digital soil map of the world. Science 325, 680–681.
Schollenberger, C.J., Dreibelbis, F.R., 1945. Determination of the exchange capacity and exchangeable bases in soils. Soil Sci. 59, 13–14.
Shangguan, W., Dai, Y., Liu, B., Zhu, A., Duan, Q., Wu, L., Ji, D., Ye, A., Yuan, H., Zhang, Q., Chen, D., Chen, M., Chu, J., Dou, Y., Guo, J., Li, H., Li, J., Liang, L., Liang, X., Liu, H., Liu, S., Miao, C., Zhang, Y., 2013. A China data set of soil properties for land surface modelling. J. Adv. Model. Earth Syst. 5, 212–224.
Shi, X., Yu, D., Warner, E.D., Pan, X., Petersen, G.W., Gong, Z.G., Weindorf, D.C., 2004. Soil database of 1:1,000,000 digital soil survey and reference system of the Chinese genetic soil classification system. Soil Surv. Horiz. 45, 129–136.
Soil Survey Staff, 2014. Keys to Soil Taxonomy, 12th edition. USDA-Natural Resources Conservation Service, Washington, DC.
Sumner, M.E., Miller, W.P., 1996. Cation exchange capacity and exchange coefficients. In: Sparks, D.L. (Ed.), Methods of Soil Analysis, Part 3. Chemical Methods. SSSA Book Series No. 5, pp. 1201–1229 (Madison, Wisconsin, USA).
Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P., 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci. 2003 (43), 1947–1958.
Tóth, G., Jones, A., Montanarella, L., 2013a. LUCAS topsoil survey. In: Methodology, Data and Results. JRC Technical Reports. Institute for Environment and Sustainability, Joint Research Centre, European Commission, Ispra, Italy.
Tóth, G., Jones, A., Montanarella, L., 2013b. The Lucas topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union. Environ. Monit. Assess. 185, 7409–7425.
Tóth, B., Weynants, M., Nemes, A., Makó, A., Bilas, G., Tóth, G., 2014. New generation of hydraulic pedotransfer functions for Europe. Eur. J. Soil Sci. 66, 226–238.
van Engelen, V.W.P., Dijkshoorn, J.A. (Eds.), 2013. Global and National Soils and Terrain Digital Databases (SOTER). Procedures Manual, Version 2.0. ISRIC – World Soil Information, Wageningen (198 pp.).
van Liedekerke, M., Jones, A., Panagos, P., 2006. ESDBv2 Raster Library - A Set of Rasters Derived From the European Soil Database Distribution v2.0 (Published by the European Commission and the European Soil Bureau Network), CD-ROM, EUR 19945 EN.
Webster, R., Oliver, M., 2007. Geostatistics for Environmental Scientists, 2nd ed. John Wiley & Sons, Chichester.
Weynants, M., Montanarella, L., Tóth, G., Strauss, P., Feichtinger, F., Cornelis, W., et al., 2013. European Hydropedological Data Inventory (EU-HYDI). EUR – Scientific and Technical Research Series. Publications Office of the European Union, Luxembourg.