Geoderma 261 (2016) 110–123

Contents lists available at ScienceDirect

Geoderma

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

Mapping topsoil physical properties at European scale using the LUCAS database

Cristiano Ballabio *, Panos Panagos, Luca Monatanarella

European Commission, Joint Research Centre, Institute for Environment and Sustainability, Via E. Fermi 2749, I-21027 Ispra, VA, Italy

ARTICLE INFO

Article history:
Received 12 September 2014
Received in revised form 11 May 2015
Accepted 12 July 2015
Available online 30 July 2015

Keywords:
Soil texture
Maps
Europe
LUCAS survey
Multivariate Additive Regression Splines
Bulk density
Available Water Capacity
USDA texture classes

ABSTRACT

The Land Use and Cover Area frame Statistical survey (LUCAS) aimed at the collecting harmonised data about the state of land use/cover over the extent of European Union (EU). Among these $2 \cdot 10^5$ land use/cover observations selected for validation, a topsoil survey was conducted at about 10% of these sites. Topsoil sampling locations were selected as to be representative of European landscape using a Latin hypercube stratified random sampling, taking into account CORINE land cover 2000, the Shuttle Radar Topography Mission (SRTM) DEM and its derived slope, aspect and curvature.

In this study we will discuss how the LUCAS topsoil database can be used to map soil properties at continental scale over the geographical extent of Europe. Several soil properties were predicted using hybrid approaches like regression kriging. In this paper we describe the prediction of topsoil texture and related derived physical properties. Regression models were fitted using, along other variables, remotely sensed data coming from the MODIS sensor. The high temporal resolution of MODIS allowed detecting changes in the vegetative response due to soil properties, which can then be used to map soil features distribution. We will also discuss the prediction of intrinsically collinear variables like soil texture which required the use of models capable of dealing with multivariate constrained dependent variables like Multivariate Adaptive Regression Splines (MARS).

Cross validation of the fitted models proved that the LUCAS dataset constitutes a good sample for mapping purposes leading to cross-validation R^2 between 0.47 and 0.50 for soil texture and normalized errors between 4 and 10%.

1. Introduction

Soils are increasingly recognised as major contributors to ecosystem services in terrestrial environment (Palm et al., 2007). Services such as food production, prevention of land degradation, water quality and carbon sequestration, just to name a few, are provided by soils (Lal, 2004). The importance of these ecosystem services has increased the relevance of soils in the socio-political agenda, resulting in an increased need of worldwide soil information. Spatial resolutions of soil maps for most parts of the world are too low to help with practical land management (Sanchez et al., 2009). Other earth sciences (e.g., climatology, geology) have taken advantage of the digital revolution and data availability. However, the conventional soil mapping still delineates space mostly according to qualitative criteria and renders maps using a series of polygons, which limits resolution.

The need for higher resolution soil datasets was already recognised by soil scientists, particularly by the International Union of Soil Sciences (IUSS) working group on digital soil mapping when proposing the launch of the *GlobalSoilMap* (GSM) project (Hempel et al., 2014). The GSM project aims at establishing standards for and eventually produce a global digital coverage of soil properties at 100 m resolution. The *GlobalSoilMap* provides the framework for supplying soil information in a format and resolution compatible with other fundamental data sets from remote sensing, terrain analysis, and other systems for mapping, monitoring, and forecasting biophysical processes (Arrouays et al., 2014).

In Europe, concern about soil conservation, resulted in the EU Thematic Strategy for Soil Protection (COM(2006)231 final) setting out a series of actions aimed at defining a comprehensive approach with the overall objective of the protection and sustainable use of soil by preventing further soil degradation, preserving its functions and restoring degraded soils. In the context of the European Union' Soil Thematic Strategy, policy makers require easy access to soil data and information of various types and scales to assess the state of soils at European level. In this context, the European Commission has decided for the establishment of the European Soil Data Centre (ESDAC) (Panagos et al., 2012). However, the most requested ESDAC dataset is the European Soil Database (King et al., 1994) which dates back to the 1990.

One of the key attributes of the European Soil Database is the soil texture along with soil coarse fragments content. It is determined by the proportion of sand, silt and clay (%) and it is expressed as a texture class (Jones et al., 2005). The spatial mapping units of

* Corresponding author.
E-mail address: cristiano.ballabio@jrc.ec.europa.eu (C. Ballabio).

http://dx.doi.org/10.1016/j.geoderma.2015.07.006

European Soil Database (ESDB) have been transferred to a raster format to map the dominant soil typological unit (Panagos, 2006). However, the textural classes are of limited use for modelling activities. The soil characteristics of ESDB were combined with the data in the Harmonised World Soil Database (HWSD) (Nachtergaele and Batjes, 2012) in order to produce soil properties maps such as clay, silt and sand (%) (Hiederer, 2013). Those texture-related datasets are accessible in the European Soil Data Centre.

According to the log statistics of the European Soil Data Centre, those spatial layers are highly requested for modelling activities in erosion by water and wind, biodiversity modelling, water capacity, crop growth, vegetation, soil conservation, moisture, land use, ecological analysis, groundwater vulnerability and hydrology.

The recent modelling developments and the increasing number of policy information requests on soil data call for more precise and updated soil information. The information in ESDB and HWSD is outdated as those datasets are based on the 1960's soil surveys and the spatial resolution is coarse for the modelling applications. Usable information on soil status and trends is a pre-requisite for developing a convincing rationale for investment in soil protection and for evaluating the effectiveness of protective actions, but the costs of soil monitoring are substantial (Kibblewhite et al., 2012).

In this context, the European Commission has performed the LUCAS soil data collection exercise (Tóth et al., 2013). The new developments both in digital soil mapping (McBratney et al., 2003) and the recent data collection exercise of LUCAS soil in 2009 allow estimating soil properties with more updated (and detailed) input datasets. LUCAS data has been already used to validate legacy maps (Panagos et al., 2013), where the estimation of soil organic carbon at NUTS2 level based on LUCAS soil data proved that the current dataset OCTOP (Jones et al., 2005) showed a clear underestimation of SOC in Southern Europe whilst in central and Eastern Europe a net overestimation is visible. Moreover LUCAS was used to produce an updated map of SOC (de Brogniez et al. 2014).

The objectives of this paper are related to the development of physical soil properties datasets based on measured soil profile data using digital soil mapping recent developments. Specifically, the study intends to:

- Introduce and apply digital soil research on a large spatial dataset of 19,857 points covering 25 Member States of the European Union
- Develop clay, silt and sand (soil texture), and coarse fragments datasets and improving the current available soil classes datasets
- Develop the first derived products (e.g. bulk density) based on the soil texture datasets
- Demonstrate how the physical property maps are converging towards *GlobalSoilMap* specifications

2. Materials and methods

2.1. LUCAS

2.1.1. LUCAS survey

The Land Use and Cover Area frame Statistical survey (LUCAS) is a project, initiated by Eurostat, aimed at the collection of harmonised data about the state of land use/land cover over the extent of European Union (EU). The survey initiated in 2006 started with the classification, through photo-interpretation, of ten million georeferenced points placed at the nodes of a 2 km grid covering EU. Among these $2 \cdot 10^5$ were selected and visited in field during the summer of 2009 to validate the survey, during this field campaign, a topsoil survey was conducted simultaneously at about 10% of the sites in 25 of the EU Member States (thus excluding Bulgaria, Romania and Croatia). Topsoil sampling locations were selected as to be representative of European landscape feature, thus a Latin hypercube stratified random sampling was applied to design the survey (Carré and Jacobson, 2009). The features taken into account for the stratified random sampling were CORINE land cover 2000, the Shuttle Radar Topography Mission (SRTM) DEM and its derived slope, aspect and curvature (Montanarella, 2011). The aim of the LUCAS survey is to establish a fully harmonised database within the European Union on land use/cover and to document changes over time. Areas above 1000 m were excluded from the survey for reasons related to the difficulties in reaching and sampling these locations. The density of LUCAS topsoil sample points is around 1 per 199 km^2, which would, in principle, allow a grid cell size of around 14 km (Hengl, 2006).

2.1.2. Soil sampling

The 19,969 topsoil samples were taken following a composite sample approach. A first subsample was taken at the location of the sampling point by sampling the topsoil (removing the litter) to a depth of 20 cm. Four other similar samples were then taken at a distance of 2 m from the original one following the four cardinal directions. The five subsamples were subsequently mixed together and 500 g of the mixture was taken as the final sample. This approach is effectively a physical averaging (De Gruijter et al., 2006) with the advantage of reducing the number of samples to analyse. About 112 composite samples were lost for lack of proper tagging, georeferencing or mismanagement reducing the total number of samples to 19,857.

2.1.3. Laboratory analysis

Dried soil samples were analysed for pH in H_2O and $CaCl_2$ solution, coarse fragments and Particle Size Distribution (PSD), $CaCO_3$ content, Cation Exchange Capacity (CEC), extractable phosphorus content, total nitrogen content, organic carbon content, extractable potassium, and visible and near infra-red diffuse reflectance. The analysis of soil parameters and in specific the particle size distribution followed standard procedures (ISO 11277).

2.1.4. LUCAS topsoil dataset

The results of the analysis are stored in the LUCAS topsoil database (Toth et al., 2013), which includes (among others) the particle size distribution expressed as percentages of clay (0–0.002 mm), silt (0.002–0.05 mm), sand (0.05–0.2 mm) as well as the coarse fragments expressed as a percentage (%) of coarse material (>2.0 mm) present in soils. The significant advantage of the unique ISO certified laboratory used to derive the attributes of LUCAS topsoil database is that discrepancies arising from inter-laboratory differences (Cools et al., 2004) have been avoided.

2.2. Environmental covariates

2.2.1. Remotely sensed data

A series of Moderate-resolution Imaging Spectroradiometer (MODIS) image products for year 2009 was collected; in particular, the MODIS Global vegetation indices (NASA Land Processes Distributed Active Archive Center (LP DAAC); ASTER L1B. USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota 2001). They are characterised by a spatial resolution between 250 and 500 m and a temporal resolution of 16 days. These MODIS products include blue, red and near- and mid-infrared reflectance, cantered at 469 nm, 645 nm, and 858 nm. The reflectance is used to determine the MODIS daily vegetation indices, such as the Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI). NDVI is defined as $\mathrm{NDVI} = \frac{(\mathrm{NIR}-\mathrm{RED})}{(\mathrm{NIR}+\mathrm{RED})}$, where VIS and RED stand for the spectral reflectance measurements acquired in the visible (red) and near-infrared regions, respectively. NDVI has been used to estimate a large number of vegetation properties from its value, such as biomass, chlorophyll concentration in leaves, plant productivity, fractional vegetation cover and accumulated rainfall. The EVI index has the form:

$$\mathrm{EVI} = g\frac{(\mathrm{NIR}-\mathrm{RED})}{(\mathrm{NIR} + c_1\mathrm{RED} - c_2\mathrm{BLUE} + l)}$$

Table 1
Prediction performances for texture and coarse fragments mapping from the LUCAS database using multivariate MARS.

	CV-RMSE	R^2	*k*-CV R^2	CV R^2	CV R^2 ESDB
Clay	7.70	0.93	0.65	0.50	0.51
Silt	12.60	0.92	0.62	0.47	0.49
Sand	17.30	0.93	0.60	0.49	0.48
Coarse f.	19.22	0.73	0.52	0.40	0.39

where NIR, RED, and BLUE are the respective surface reflectance, l is the canopy background adjustment, and c_1 and c_2 are coefficients for the aerosol resistance term, which uses the blue band to correct for aerosol influences on the red band. The coefficients adopted by the MODIS-EVI algorithm are; $l = 1$, $c_1 = 6$, $c_2 = 7.5$, and g (gain factor) $= 2.5$.

The mapping resolution of 500 m was chosen since all the selected covariates, except climate, were available at a finer or equal resolution. Doing this, we avoided the generation of artefacts by downscaling. Using the full series of 16 days MODIS products for year 2009 as covariates was deemed impractical, thus we transformed each product time series into a smaller set of images using Principal Component Analysis.

2.2.2. Land surface parameters

The NASA-Shuttle Radar Topography Mission (SRTM) digital elevation model was used to derive land features at a resolution of 100 m for all Europe. Both the DEM and the derived surface parameters were then rescaled at 500 m. The derivation of land surface parameters was made using the SAGA GIS software. Amount the various parameters derived and tested, the most useful ones were the Multi-resolution Valley Bottom Flatness (MRVBF) and the Multi-resolution Ridge Top Flatness (MRRTF) (Gallant and Dowling, 2003), slope, slope height and vertical distance to channel network (CNBL).

2.2.3. Land cover

The CORINE (CORdinate INformation on the Environment) is a raster format land cover database comprising 44 classes. CORINE is derived from aerial photographs using computer aided photointerpretation. CORINE nominal scale is 1:100,000 with a minimum mapping unit (MMU) of 25 ha and a change detection threshold of 5 ha. The CORINE

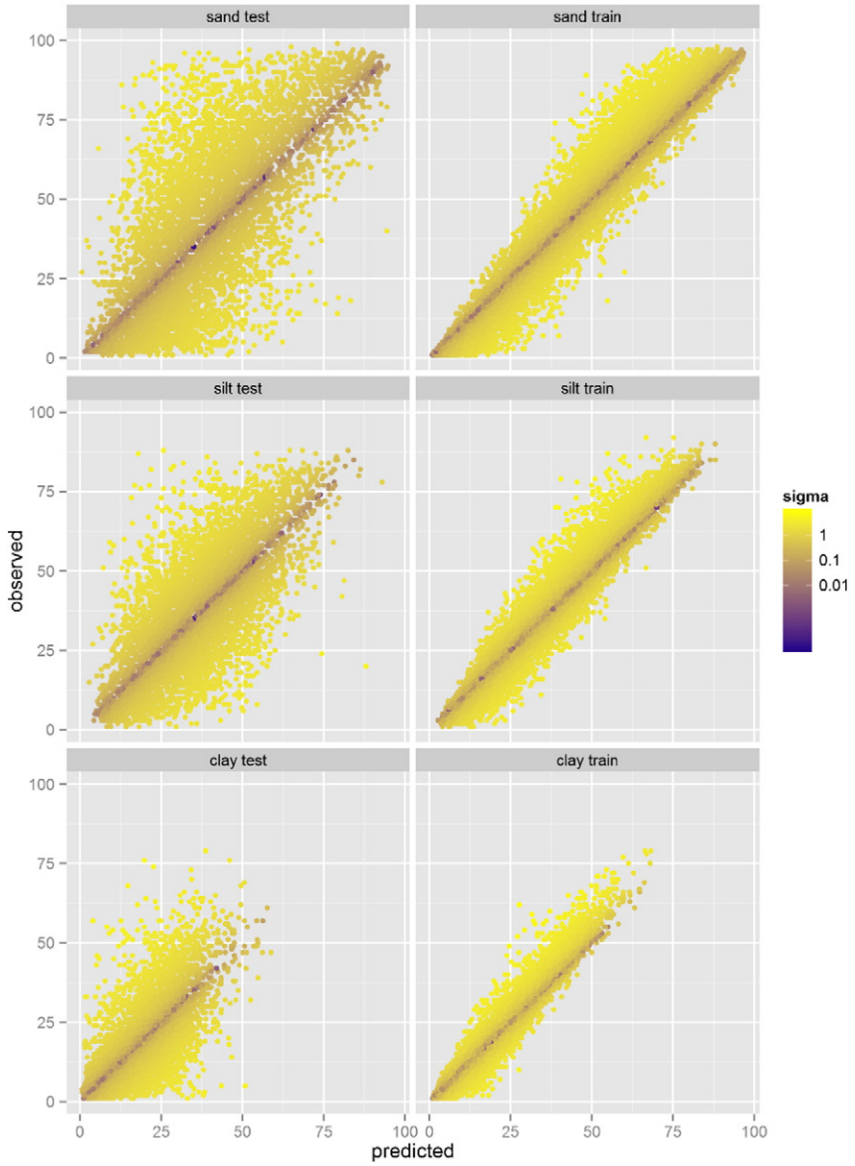

Fig. 1. Predicted-observed scatterplots for training and the validation sets of the three textural components.

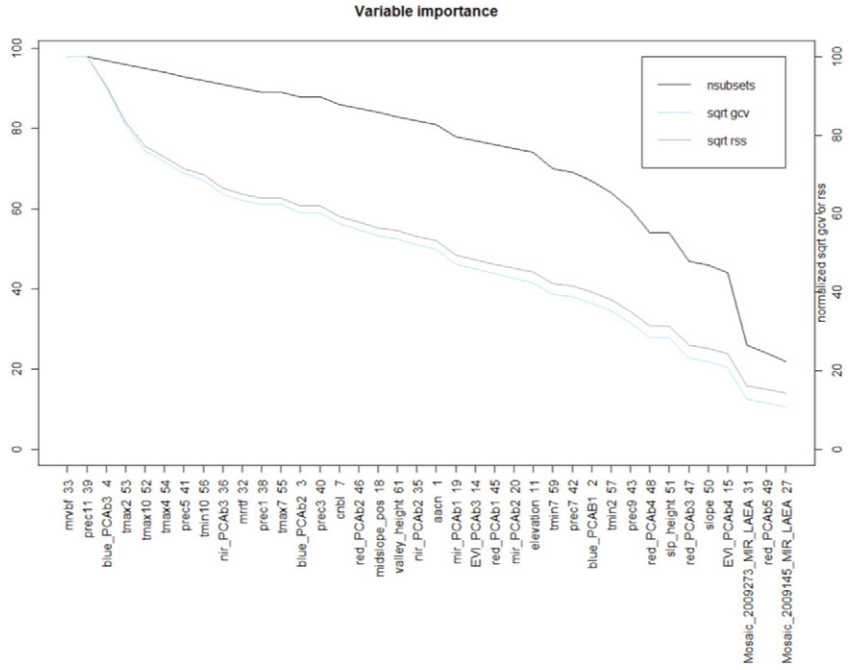

Fig. 2. Relative importance of covariates used in the Multivariate Additive Regression Splines model (mrvbf, multiresolution valley bottom flatness; prec*x*, precipitation of month *x*; *colour*_PCA*by*, component *y* of the PCA of given *colour*; temp*x*, temperature of month *x*; mrrft, multiresolution ridge top flatness; cnbl, channel network base level; acnbl, altitude above channel network base level).

was used to represent the spatial distribution of land use/land cover. The reliability of CORINE 2000 version at 95% confidence level is 87.0 ± 0.7%, according to the independent interpretation performed on the LUCAS (Land Use/Cover Area Frame Survey) data (Büttner et al., 2011).

2.2.4. Climatic data

Monthly temperature averages and extremes and monthly precipitations were obtained from the WorldClim (http://www.worldclim.org/) dataset at a spatial resolution of 1 km^2. These data layers are the interpolated values of average monthly climate data collected from numerous weather stations. The approach uses a thin plate smoothing spline with latitude, longitude and elevation as independent variables to locally interpolate data (Hijmans et al., 2005). Climatic data was included in the model after PCA projection of monthly inputs. Temperature and rainfall were treated as two independent variables, thus performing two separate PCAs.

2.2.5. Soil data

The European Soil Database (ESDB) was considered in the first stages of this study as a possible covariate. The ESDB was used first as an ordered categorical variable by attributing the relative texture classes to each mapping unit. In the following steps the ESDB was utilised as a multinomial variable just by identifying and labelling similar soil units. In both cases the use of the ESDB was found to provide little improvement to the model outcome and was then dropped from the analysis.

2.3. Statistical analysis

2.3.1. Additive log-ratio transform

Texture is expressed as the relative percentage of sand (>0.05 mm), silt (0.002–0.05 mm) and clay (<0.002 mm), as such the sum of the three components always equals 100. In this sense the three variables are effectively two as any third component can be inferred by the difference between the other two. This means that these three dependent variables are both highly correlated and constrained. Modelling each one of them separately would result in inconsistent results, like sum values above 100.

The values of sand, silt and clay content in a soil are a composition of three elements which does not belong to a $\mathbb{R}^3$ real space, but comes from a two-dimensional simplex plane embedded in 3-d space. This simplex plane is commonly shown as the triangular ternary diagram of soil textural classes.

As such sand, silt and clay are not only reciprocally bounded, but also spuriously negatively correlated (Aitchison, 1982).

As suggested by Aitchison (1986), compositional variables should be transformed into log ratios. Given a composition of D elements $\mathbf{z}$

$$\mathbf{z} = [z_1, \ldots, z_D].$$

Such that $z_i > 0 \; \forall \; i = 1, \ldots, D$ and $\sum_{i=1}^{D} z_i = 1$, the additive log-ratio (ALR) transform is defined as

$$\mathbf{x} = ALR(\mathbf{z}) = \left(\ln \frac{z_1}{z_D}, \ldots, \ln \frac{z_{D-1}}{z_D} \right).$$

The new variate $\mathbf{x}$ belongs to a space of dimension D-1. By defining a vector $\mathbf{w} = [\mathbf{x}^T, 0]^T$, the inverse of the ALR transform can be defined as

$$\mathbf{z} = \frac{\exp(\mathbf{w})}{\mathbf{j}^T \exp(\mathbf{w})}.$$

Where $\mathbf{j}$ is a vector of length D with all elements equal to one. Whilst there is no unbiased simple back transform of the Addictive Log-Ratio (ALR), numerical approximation is possible by Gauss–Hermite quadrature (Aitchison, 1986).

Lark and Bishop (2007) applied the ALR to the cokriging of soil texture. They evidenced that the ALR transformed variables preserve information on the spatial correlation whilst avoiding the occurrence of singular matrices. As such ALR was applied on LUCAS texture data before further analysis.

2.3.2. Prediction of soil properties using Multivariate Adaptive Regression Splines

Multivariate Adaptive Regression Splines (MARS) (Friedman, 1991) is an adaptive procedure for regression. MARS makes use of basis expansion using piecewise linear functions set as

$$(x-t) = \begin{cases} x-t, & \text{if } x>t \\ 0, & \text{otherwise} \end{cases} \quad (t-x) = \begin{cases} t-x, & \text{if } x>t \\ 0, & \text{otherwise} \end{cases}. \tag{1}$$

These functions are linear splines with a knot at value t, the two functions are reflected pairs, meaning that they are symmetric in respect to t. The MARS approach aims at building reflected pairs for each input X_j, with knots placed at X_{ij}. This means building a collection of basis functions definable as

$$C = \left\{ (X_j - t)_+, (t - X_j)_+ \right\} \tag{2}$$

where $t \in x_{1j}, \cdots, x_{Nj}$ are the locations of the knots and $j = 1, \cdots, p$ are the inputs.

Model building is done using an approach similar to forward stepwise linear regression, but using functions from a collection of basis functions

$$f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X) \tag{3}$$

where $h_m(X)$ is a function in C or a product of two of such functions. Starting from a constant model $h_m(X) = 1$ new basis functions are added, pairing the products of a function h_m with one of the reflected pairs in C, the product which results in the largest reduction of training error is retained. The model thus fitted is typically over-fitted, so a backward selection procedure is applied, usually aiming at minimizing Generalized Cross-Validation (GCV) error:

$$GCV(\lambda) = \frac{\sum_{i=i}^{N} \left(y_i - \hat{f}_\lambda(x_i) \right)^2}{\left(1 - \frac{M(\lambda)}{N} \right)^2} \tag{4}$$

where $\hat{f}_\lambda$ is the best model of size λ and $M(\lambda)$ is the number of model parameters. This pruning procedure also acts as a feature selection procedure removing the uninformative inputs from the model. In MARS, overfitting is controlled by the number of basis functions. MARS also has additional controls, like the degree of interaction allowed, and the number of data points required between knots. The model was tested with different sets of parameters using cross-validation on an independent sample whilst aiming at minimizing GCV. In this study the allowed degree of interaction was set at 2 and the number of knots was chosen as $M(\lambda) + \frac{M(\lambda)-1}{2}$ with $=$ 2.137.

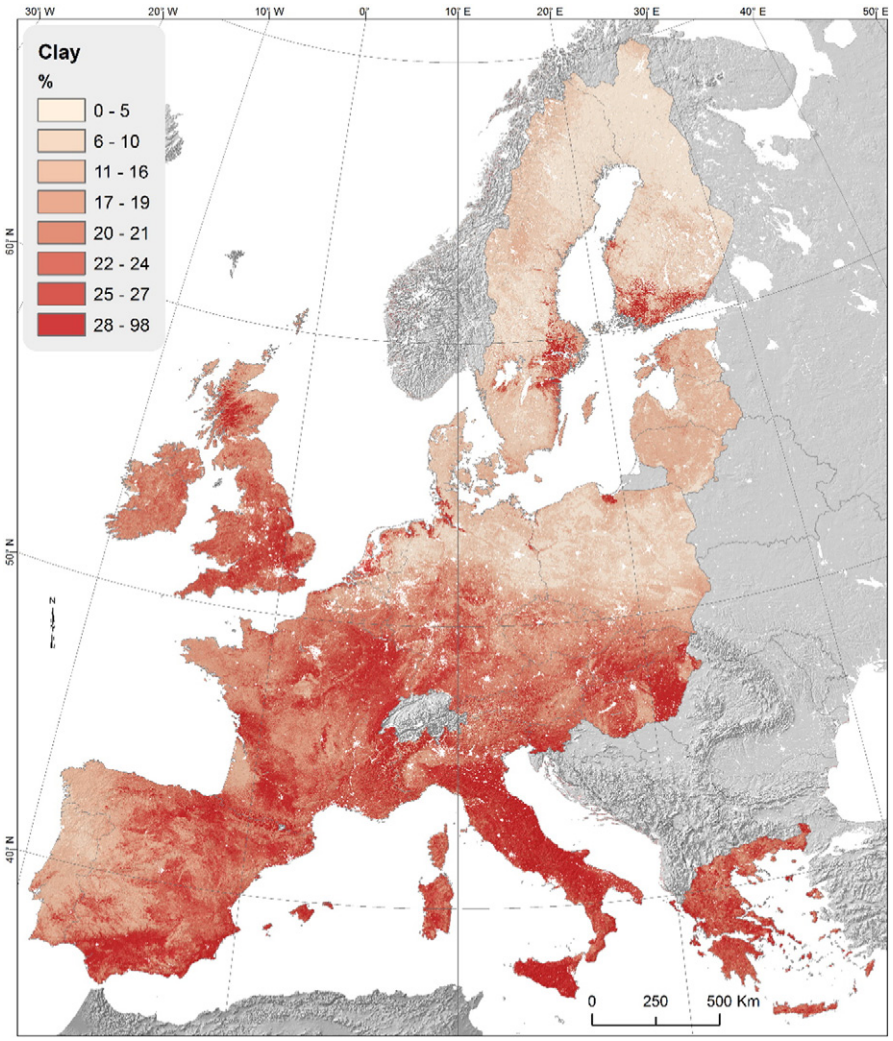

Fig. 3. Topsoil (0–20 cm) clay content (%) modelled by Multivariate Additive Regression Splines.

Different metrics were used to estimate model performance, the first one being the adjusted coefficient of determination ($\overline{R}^2$) defined as

$$\overline{R}^2 = R^2 - \left(1-R^2\right)\frac{p}{n-p-1}$$

where R^2 is the coefficient of determination, p is the total number of regressors in the model, and n is the sample size. Another metric utilised, is the Root Mean Square Error (RMSE) defined as $RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$ where $\hat{y}$ is the predicted value of the dependent variable y.

2.3.3. Multivariate prediction of texture using MARS

A feasible approach is to treat the parameters as a multivariate outcome in a multiple response model. MARS models can also deal with multiple responses, like multivariate dependent variables. In this case, k simultaneous models are built using the same set of basis functions, but different coefficients. The models are subsequently pruned considering summed GCV across all the k responses. This usually results in a fitting which is slightly worse than one of models fitted on a single response. However this procedure is more meaningful for compositional variables. Given the ALR transformed data a multivariate MARS approach was applied on the components of the ALR transformed variate.

2.3.4. Mars standard deviation estimation

A variance function estimation (Davidian and Carroll, 1987) was used to calculate prediction intervals. The procedure aims at building a residual model using Iteratively Reweighted Least Squares (IRLS). The residual model is the regression of the absolute residuals of the MARS model, this absolute residuals are defined as

$$\sqrt{\hat{\epsilon}_{i_f}^2} = \frac{(y_i - \hat{y}_i)^2}{1-h_{ii}} + modvar_i$$

where $\hat{y}_i$ is the predicted value of a given instance y_i, h_{ii} is the point leverage defined as the diagonal of the hat matrix from the linear fit of the response in MARS basis matrix, and $modvar_i$ is the estimated model variance at the point.

The residual model is then used to estimate prediction intervals for predictions as follows. The mean absolute error is estimated from the residual model, then the error is rescaled as $\sigma = \sqrt{\pi/2}\,\mathrm{mean}(\mathrm{asb}(\mathrm{error}))$, the next step is to convert the standard deviation into an estimated prediction interval. In this study we decided to show a map of the

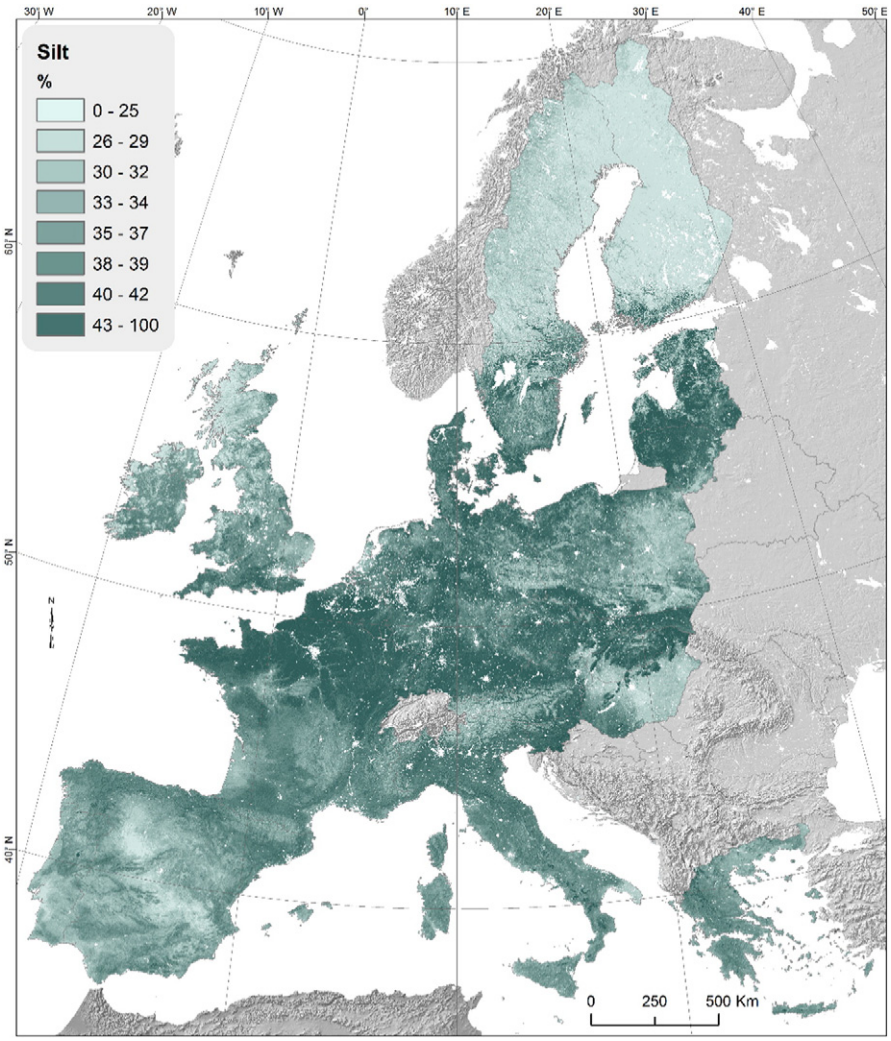

Fig. 4. Topsoil (0–20 cm) silt content (%) modelled by Multivariate Additive Regression Splines.

standard deviation, so we did not explicitly calculate the prediction intervals.

3. Results and discussion

3.1. Model fitting

Texture was predicted using a multivariate MARS; this procedure constrains the prediction of every single property. Texture data was transformed using the additive log-ratio transform (Aitchson, 1986; Lark and Bishop, 2007). Table 1 shows prediction performance for model fitting (R^2), *k*-fold cross validation (*k*-CV R^2) and independent sample validation (CV R^2). Independent sample validation was performed by selecting 5000 random samples (by a stratified random sampling) and using them to validate the model fitted on the remaining ~15,000 samples; in this case the metrics used to evaluate model performance is RMSE. The *k*-fold cross-validation was performed for a $k = 5$ and repeated 100 times using different random splits in order to obtain more stable estimates by averaging.

The best predicted variable was the clay content, whilst silt content was less well predictable. However the differences are substantially negligible. Coarse fragments were treated as an independent variable and predicted by a different MARS model, as such the metrics for coarse fragments are presented in a different line of Table 1. Model fitting resulted in very good performance metrics both in fitting and cross-validation (Table 1), with only the prediction of coarse fragments performing quite differently from the others.

Table 1 also depicts the change of CV R^2 when ESDB units are added as dummy variables (CV R^2 ESDB), it should be noted that being the GCV term in MARS comparable to Akaike Information Criterion (Barron and Xiao, 1991) the fitting procedure of the model already selects the most efficient model. It is thus the model that selects the most informative variables or excludes the least informative. In this context we found that MARS models consistently rejected data from soil units. We will discuss this aspect below.

Fig. 1 depicts the *k*-fold cross validation results by plotting the predicted versus observed values for the three variables for both the fitting and the validation sets. The variable colour scale in the same plot depicts the normalized standard deviation for a given observation as estimated through the 100 repetitions. From Fig. 1 we can see that the fitted values present a quite low dispersion with most of the values within the value of the standard deviation. In general the errors are homoscedastic, this contributes to the high R^2 values of Table 1. However it is possible to notice a slight bias as the values are consistently over predicted for high observed values and under predicted for the lower ones. *k*-Fold errors are more dispersed as usual with some quite large deviation, this is expected as cross validation tests the generalization capacity of the model on new samples. Nevertheless model performance is still quite good with most of the samples falling within the value of the standard deviation.

The prediction performances we obtained in this study are quite high compared to similar studies. We believe that having a harmonised

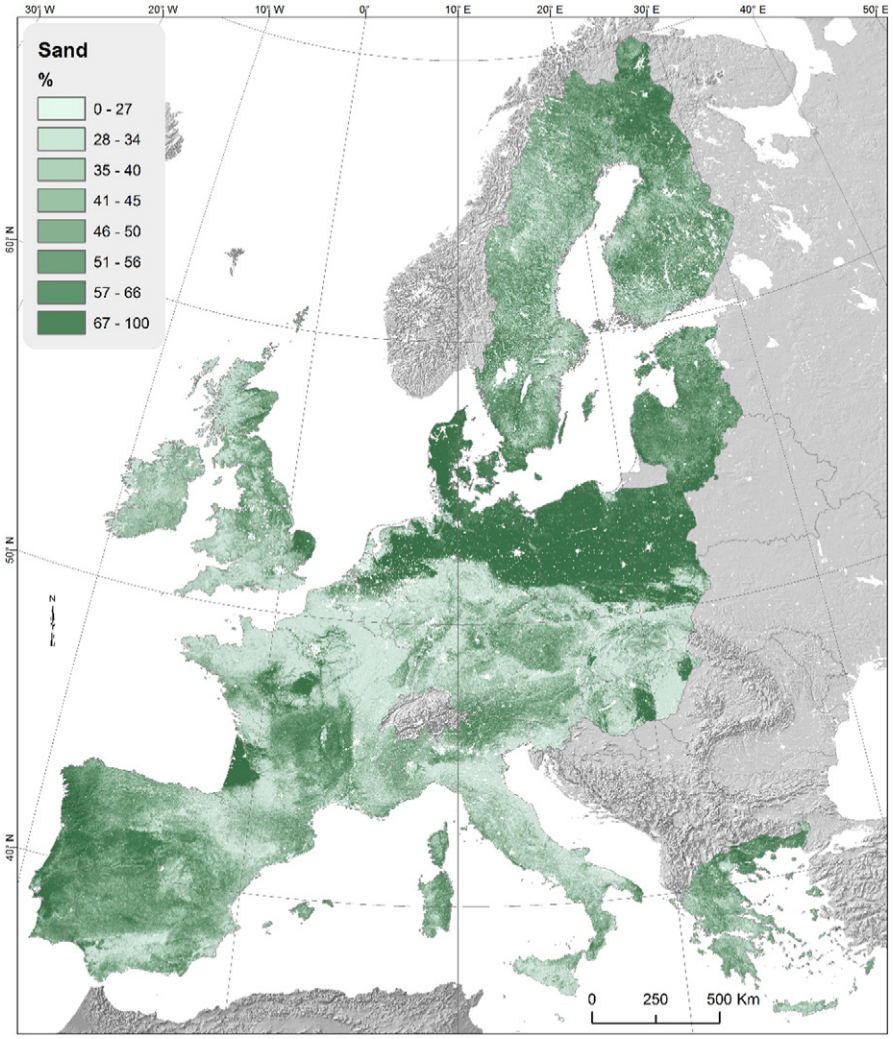

Fig. 5. Topsoil (0–20 cm) sand content (%) modelled by Multivariate Additive Regression Splines.

sample taken with a standardized, a well defined procedure with a consistent support is what makes the difference in this study. Commonly DSM studies at this scale use quite heterogeneous data sources resulting in added bias and noise as different surveys are not meant to be comparable or measuring the same thing at the same scale.

Fig. 2 shows variable importance according to the difference in Residual Sum of Squares (RSS) and generalized cross-validation (GCV) that the removal of a given variable produces on the model outcome.

The most important variable is the Multiresolution Valley Bottom Flatness (mrvbf) followed by November average precipitation (prec11), PCA decomposition of the MODIS blue band yearly series and the maximum temperatures of October and April. The following variables show a mix of terrain features, climate and MODIS data.

Whilst the interpretation of the influence of terrain variables might be interpreted in terms of different sediments (fluvial, marine, fluvioglacial, etc.), belonging to different land-forms, the effect of climate is less clear. Although climate influences leaching rates, its influence on the model is probably due to the interaction with vegetation cover response and soil wetness due to different permeability and water retention. As vegetation under similar climates will react differently due to soil properties (water retention, nutrients holding capacity, etc.), the combination of climate and vegetation response through the year helps discriminate different soils.

The soil data (ESDB) failed to significantly increase the performance of the model. The first approach was to use the ESBD textural classes as an ordinal categorical variable, after the model failed to improve its performance (in terms of cross validation RMSE). In the second approach the ESDB was used as a multinomial variable to stratify the other covariates. Nevertheless, the adjusted R^2 did not increase in a significant way. The fact that the use of the ESDB do not raise or lowers model performance can be interpreted as the information from the soil map being successfully provided by the other covariates. However, this issue can arise for different reasons. For instance, the ESDB map can lack the necessary spatial detail to describe soil variation at finer scales (like the scale of topographic variation as described by DEM). Another possibility is that ESDB is too spatially heterogeneous; being the result of the merging and harmonization of national maps, there is the possibility that the map is not geographically harmonised, meaning that similar soil units were designed using different criteria in different countries (something that is quite obvious from the ESDB map itself). The last reason we hypothesise here is that soil units defined for the purpose of soil classification might not reflect in significant changes in topsoil properties, especially those considered in this study. No matter the reason, we consider the finding of a lack of relevance of the ESDB in the mapping process as an interesting result of this study, as it tells some cautionary tale about the use of legacy soil maps (especially those resulting from the aggregation of other legacy soil maps).

After this we decided to drop the ESDB from the model for the rest of the study. Whilst this choice might be criticized, we argue that not using legacy soil data can be seen as having also a positive impact on the study. First, it proves that a broad range of covariates can act as a proxy for soil–

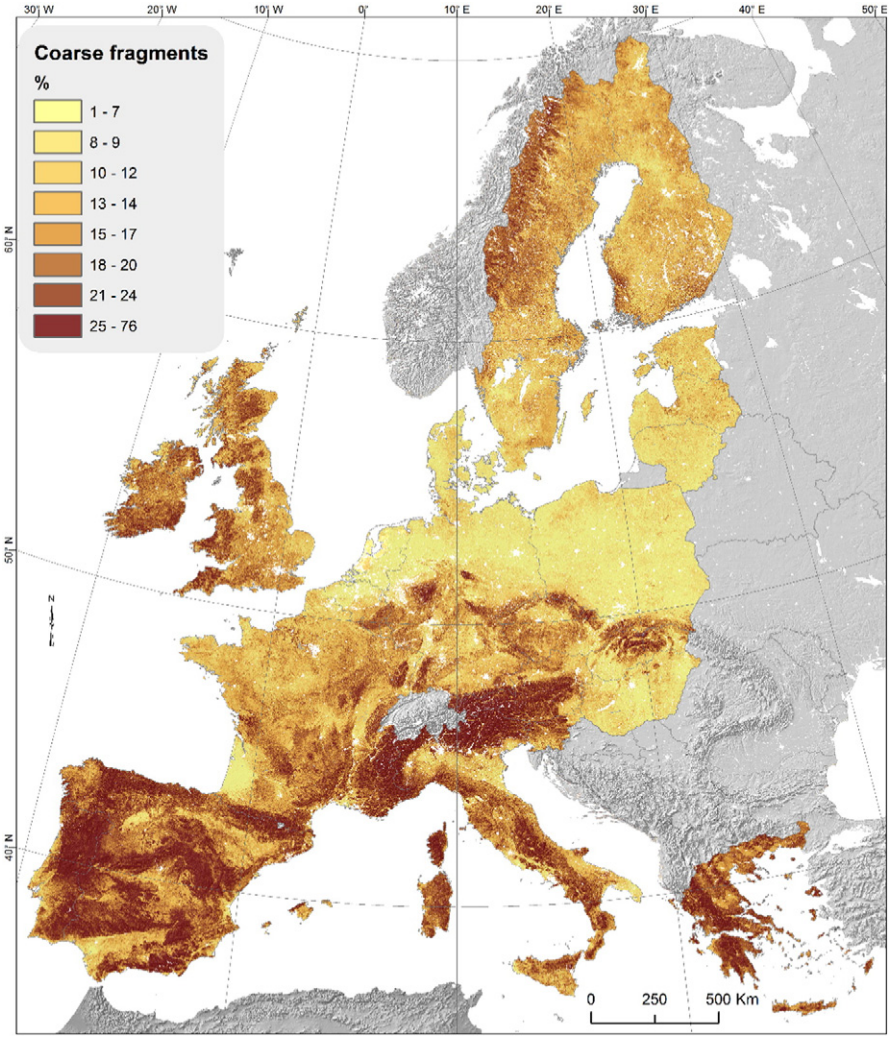

Fig. 6. Topsoil (0–20 cm) coarse fragments (%) content modelled by Multivariate Additive Regression Splines.

landscape relation and can, at least at this working scale and sampling density, substitute soil data. This can be seen as a positive finding as it allows us to perform predictions in areas where legacy soil maps are missing or outdated. Moreover our finding is supported by studies using terrain attributes and remotely sensed data to spatially disaggregate soil maps (Odgers et al., 2014) showing the close relation between them.

3.2. Prediction of soil properties

Out of twelve soil properties listed by the GlobalSoilMap.net (GSM) specifications, seven can potentially be directly predicted from LUCAS samples using MARS, namely:

1. organic carbon (g Kg^{-1})
2. pH (in water and $CaCl_2$ solution)
3. sand (%)
4. silt (%)
5. clay (%)
6. gravel (% vol)
7. ECEC.

As referred in the Introduction section, the present study will focus only on the physical attributes: sand, silt, clay and coarse fragments. Moreover, the application of pedotranfer rules will result in the most common derived datasets from those textural parametres such as soil textural classes (USDA classification), bulk density and Available Water Capacity (AWC). These physical properties are generally stable in a relatively long timespan (decades). Transformations of physical features such as texture and mineralogical composition will only occur over decades whilst properties such as pH, organic matter content or microbial activity will show a more rapid reaction (Jones et al., 2012).

Figs. 3 to 5 show the maps of topsoil clay, silt, sand and coarse fragments, areas where soil is not present, like water bodies and built areas were masked in the final maps.

Fig. 3 shows the clay content in the topsoil, showing a relatively high clay content in the Mediterranean region whilst low clay content is noticed in the Scandinavian countries and Baltic States.

The most striking features of the maps can be described by referring at the geological history of Europe. It is beyond the aim of this study to give a full reference to all the map features, but we try to give a short interpretation of the most striking ones. The silt content spatial distribution (Fig. 4) seems to reflect the influence of Late Glacial loess deposition (Semmel and Terhorst, 2010). Sandy soils are found in Nordic countries and Baltic States whilst the Mediterranean basin has a low sand content (Fig. 5). These maps seem to reflect the late geological history of Europe. Particularly striking is the clear definition of the extent of the Last Glacial Maximum (LGM) (Svendsen et al., 2004) ice sheet over Scandinavia and northern Europe, which is marked by a clear passage from sandy soil to finer textures and follows the subdivision proposed by Plant et al. (2003). Essentially, Scandinavia underwent a net loss of material by glacial scouring, whilst much of northwest and central Europe,

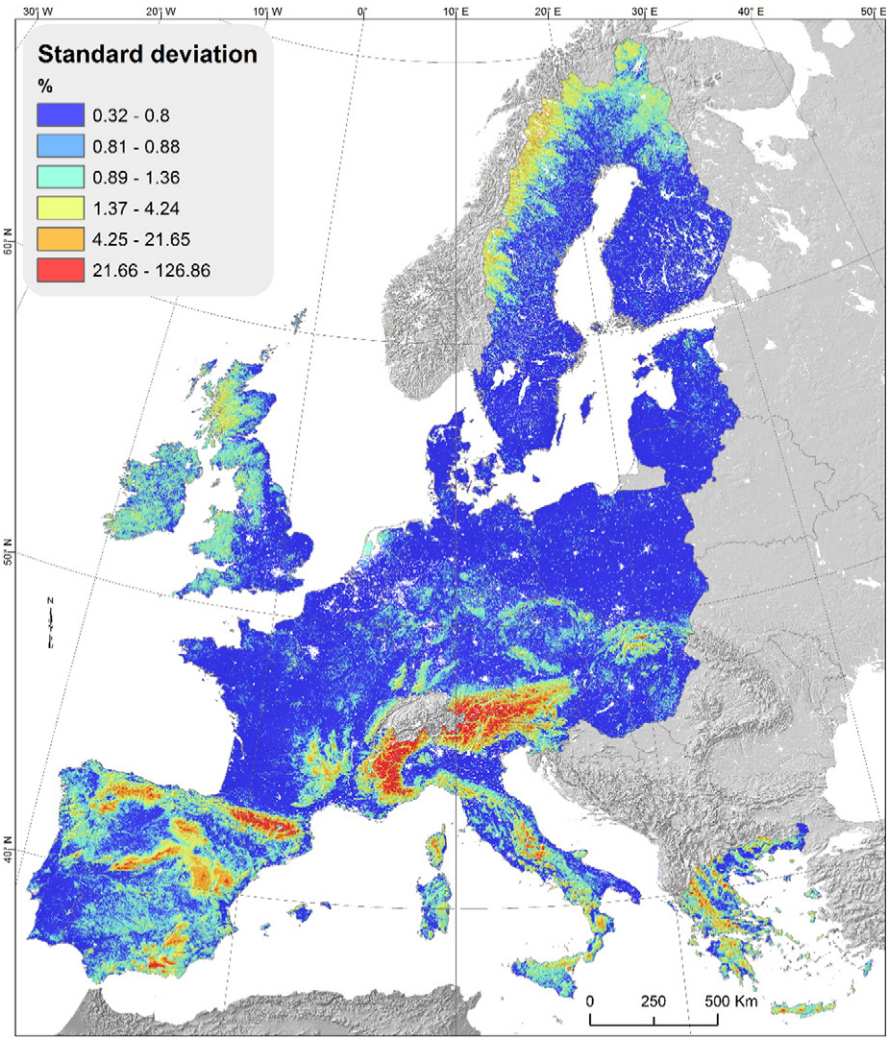

Fig. 7. Averaged standard deviation of the Multivariate Additive Regression Splines model.

south to 50N, experienced a net gain of a complex suite sand, gravel and loess forming the North European Plain. South of this latitude, towards the present-day Mediterranean, fluvial outwash sediments and loess, cover the underlying pre-Quaternary strata. Also due to the LGM the silt content is increasing in correspondence with aeolian loess deposits, which follows the same spatial pattern of the Loess map of Europe 1:2,500,000 (Haase et al., 2007).

Clay seems more related to the presence of sedimentary rocks in the substrate, especially in areas where limestone and clay stone are common, such as central and southern Italy, France and Spain (Asch, 2003).

The coarse fragments map (Fig. 6) depicts the distribution of coarse fragments as following the spatial distribution of main mountain ranges in Europe. The highest concentration of coarse fragments can be found in the Alps, Pyrenees, Iberian mesas, Apennines and the Balkans. The presence of coarse fragments is minimal in areas of shallow water deposition along the Baltic and North Seas southern perimeter. This is not surprising as mountain soils are characterised by high percentages of coarse fragments both due to increased soil erosion (which removes the finer soil fraction) and due young soil rich in parent material coming from the mechanical alteration of the bedrock.

3.3. Map of uncertainties

A map of model standard deviation (Fig. 7) was also produced. As the MARS models the variables as an ensemble, the resulting standard deviation map was obtained as an averaged composite of the standard error of the three variables. Areas above 1000 m evidence the high uncertainties and evidence the difficulty in predicting unstamped areas.

In general the map depicts a quite low model standard deviation in relatively homogeneous areas such as plains. Regions with a more diverse morphology are in general less well predicted (western Scotland, Pyrenees, Apennines, western Greece, etc.). In this case topography seems to be the main controlling factor in determining model performance. In general the worst performance is obtained in mountain and hilly areas, this can be explained by the fact that these areas have a high diversity in terms of terrain, land cover and substrate, whilst being sampled with the same density as the rest of Europe, resulting in a larger model deviation. Areas above 1000 m of altitude show the highest uncertainties which are of the same order of the predicted values (up to and above 100%).

3.4. Residual spatial correlation

Residuals from the MARS model were checked for the presence of spatial correlation. Variograms and cross-variograms (Fig. 8) were computed for all the textural components. Only the silt model residuals' variogram shows a negligible spatial correlation at relatively short distances (~25 km), whilst the sand residuals appear to have dubious spatial correlation and clay's variogram is decreasing up to ~10 km and then flattens. Probably the absence of spatial correlation among residuals is due to the strong influence topography exerts on texture as well as the good performance of the model which succeeds in explaining most of the texture variability. Given the very little evidence

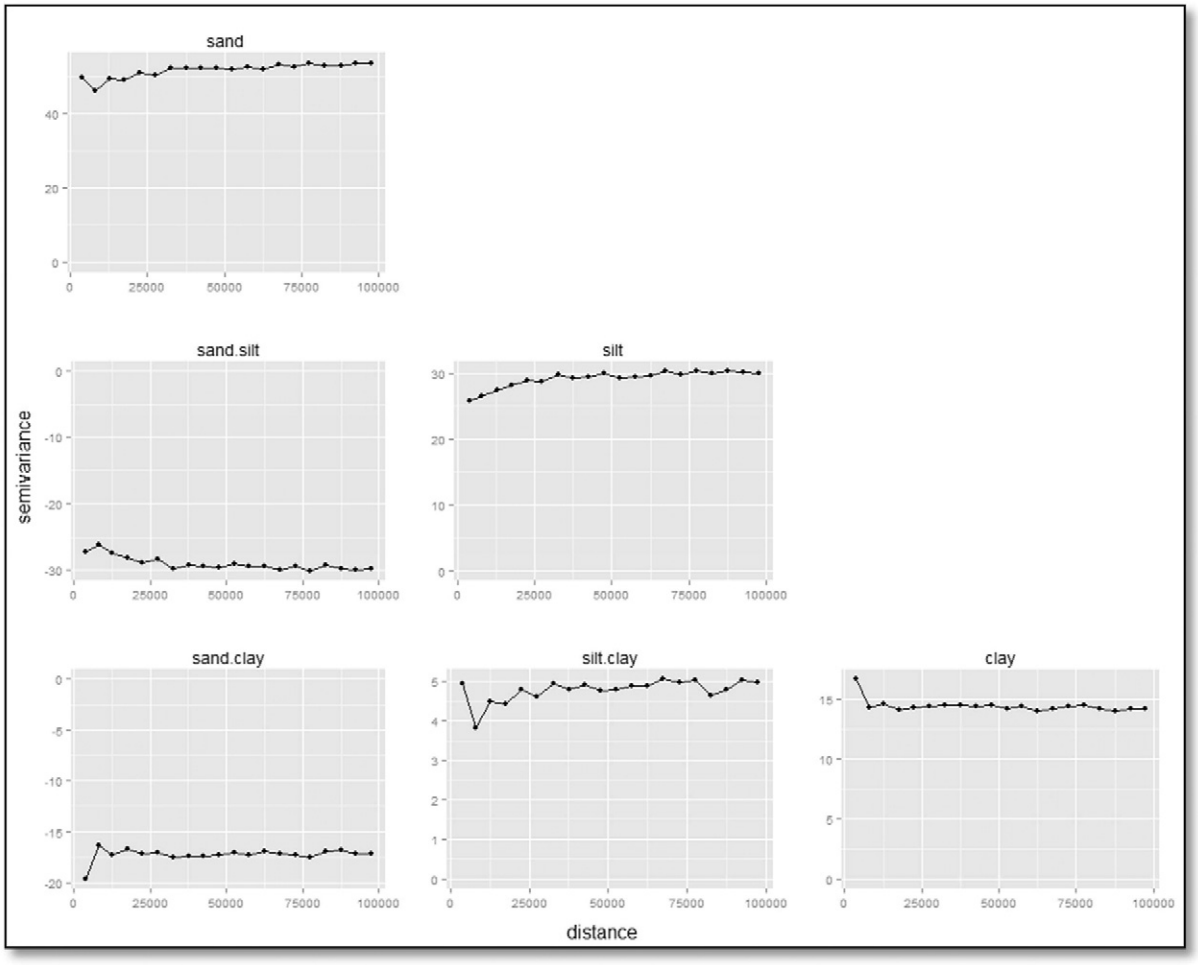

Fig. 8. Variograms and cross-variograms of Multivariate Additive Regression Splines residuals (distances in metres).

of residual spatial correlation, the use of hybrid techniques, like regression kriging was ruled out from the present study.

3.5. Derived products

Soil–water content data such as hydraulic conductivity, their relationships and Available Water Capacity are needed for many plant and soil–water studies. Measurement of those data and possible relationships is costly, difficult, and often impractical. Research studies have developed statistical correlations between soil texture and selected soil potentials using a large database, and also between selected soil textures and hydraulic conductivity (Saxton et al., 1986). In this section, some of the potential applications on how to use the physical attributes data (e.g. texture) is demonstrated; bulk density, USDA soil texture classes and AWC were derived from texture and legacy data.

Bulk density has application to nearly a large number of soil studies and analyses. The current activity in soil quality, soil sufficiency, and sequestration of C has increased interest in bulk density, particularly of surface layers (Dane et al., 2002). The bulk density (Fig. 9) was obtained from the packing density and the mapped clay content (Fig. 3) following the equation of Jones et al. (2003)

$$\rho_B = \rho_P - 0.009\ C$$

where ρ_P the packing density and C is the clay content.

The productivity of agricultural soils is mainly driven by the soil water properties which are largely dependent on soil texture. Soil texture is the main driver for nutrient dynamics and soil resilience (McLauchlan, 2006).

USDA soil texture classes are widely used in Europe for estimation of other physical properties (compaction) or hydraulic properties (Wösten et al., 2001). Using the combination of three textural maps (clay, silt and sand), the USDA classes map has been developed (Fig. 10).

The identification of fine and coarse soils (Fig. 10) allows policy makers to develop soil management techniques. For example, they can propose tillage technologies to save water or fertilizer practices to ensure the long-term soil sustainability and agricultural productivity. Moreover, soil erodibility (known as K-factor in RUSLE family models) largely depends on soil texture (Panagos et al., 2014), as well as the sensitivity of soils to compaction (which also depends on several external factors such as climate and land use) (Horn et al., 1995).

Compared with past attempts to propose the particle size distribution at the European level (Hiederer, 2013), the new presented clay, silt and sand datasets have the advantage of being based on pan-European harmonised field data.

A useful indirect way to investigate the effects of different soil textures on soil hydraulic properties is to use pedo-transfer functions. These are regression equations that enable the soil hydraulic parametres to be estimated from soil texture. Wösten et al. (2001) proposed a set of pedo-transfer functions based on texture data. Soil water

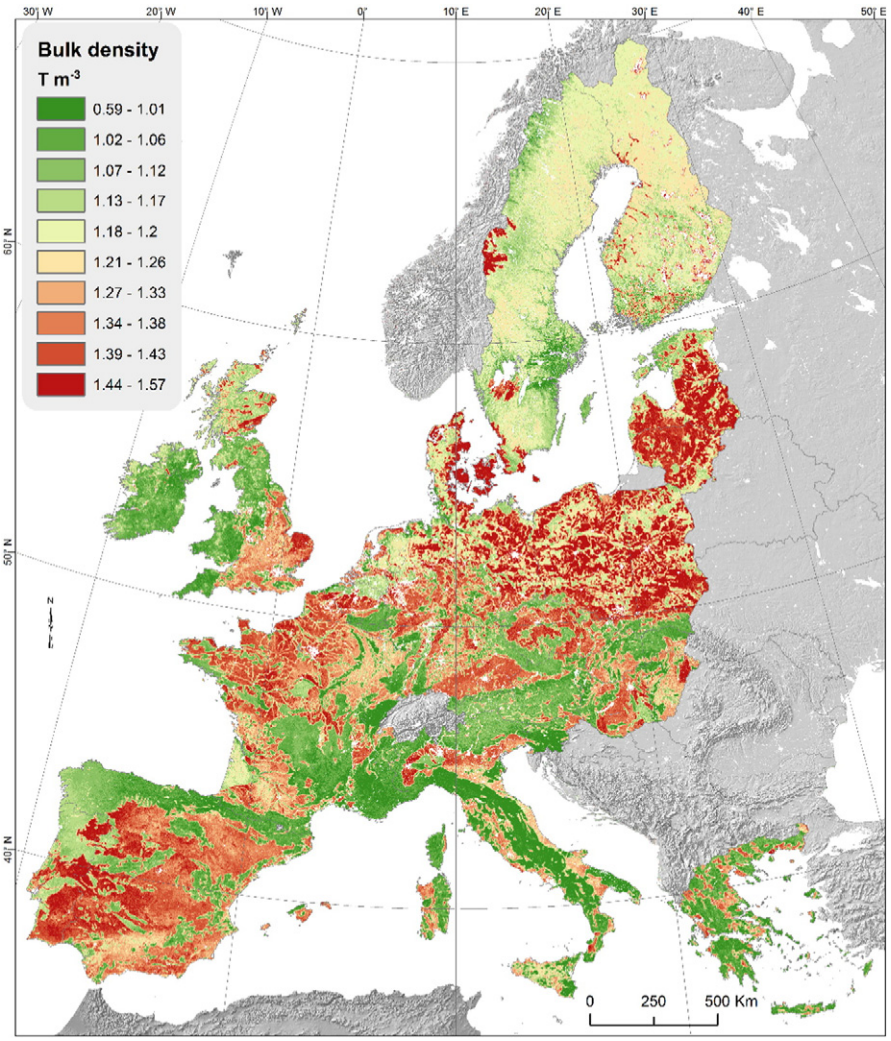

Fig. 9. Bulk density derived from soil texture datasets.

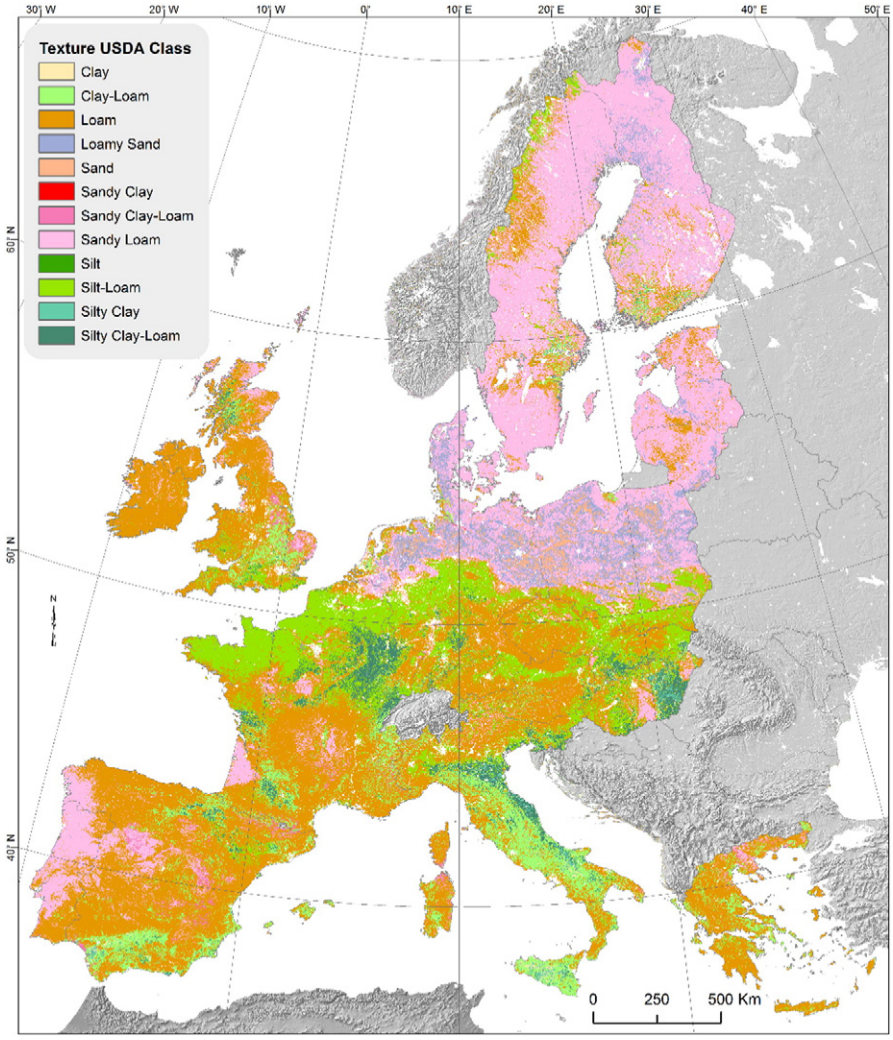

Fig. 10. USDA soil textural classes derived from clay, silt and sand maps.

retention characteristics depend largely on texture, the amount of SOM and climate. Variations in any of these three variables will affect soil water retention characteristics and ultimately, soil functions (e.g. agriculture, water storage).

The relation between soil–water tension h and water content θ was calculated using the van Genuchten water retention function (Van Genuchten, 1980). The water retention function has the form

$$\theta(\psi) = \theta_r + \frac{\theta_s - \theta_r}{((1 + (\alpha|\psi|)^n)^{1-1/n}}$$

where $\theta(\psi)$ is the water retention curve, $|\psi|$ is the is suction pressure, θ_s the saturated water content, θ_r the residual water content; and α and n are parametres related to the air entry suction and to pore-size distribution respectively.

The parametres for the van Genuchten equation were derived from texture using the continuous pedotransfer functions of Wösten et al. (2001) to predict the saturated water content θ_s of a soil after its clay and silt content, bulk density, organic matter content and topsoil or subsoil qualifier. Organic matter content was estimated using the map of organic carbon produced by de Brogniez et al. (2014). Finally, the Available Water Capacity (AWC) was derived as the difference between the −33 kPa and the −1500 kPa water content (expressed as volume fraction). Following the textural distribution, the Available Water Capacity map (Fig. 11) has a major north/south gradient following the higher sand content found in northern European soils. In general higher AWC values can be found in areas where soil texture is more favourable corresponding to western-central Europe, southern Europe and part of the British Isles.

3.6. Data availability

The data availability of LUCAS topsoil physical properties is a key issue for modellers who have no access to high spatial resolution data. With the publication of this study, modellers and in general scientists will be able to download the soil texture (clay, silt and sand) and coarse fragment datasets and the derived products (Bulk density, AWC) from the European Soil Data Centre (ESDAC).

4. Conclusions

The three layers of soil texture (clay, silt and sand) plus coarse fragments were mapped at 500 m grid cell resolution for the European Union applying the Multivariate Adaptive Regression Splines (MARS) model. The spatial interpolation model showed a good performance (cross validation $R^2 = 0.65$, 0.62, and 0.60 corresponding to the clay, silt and sand prediction), and high prediction uncertainty was limited to relatively few areas. This study also contributed to the process towards producing the first components of the *GlobalSoilMap* deliverables

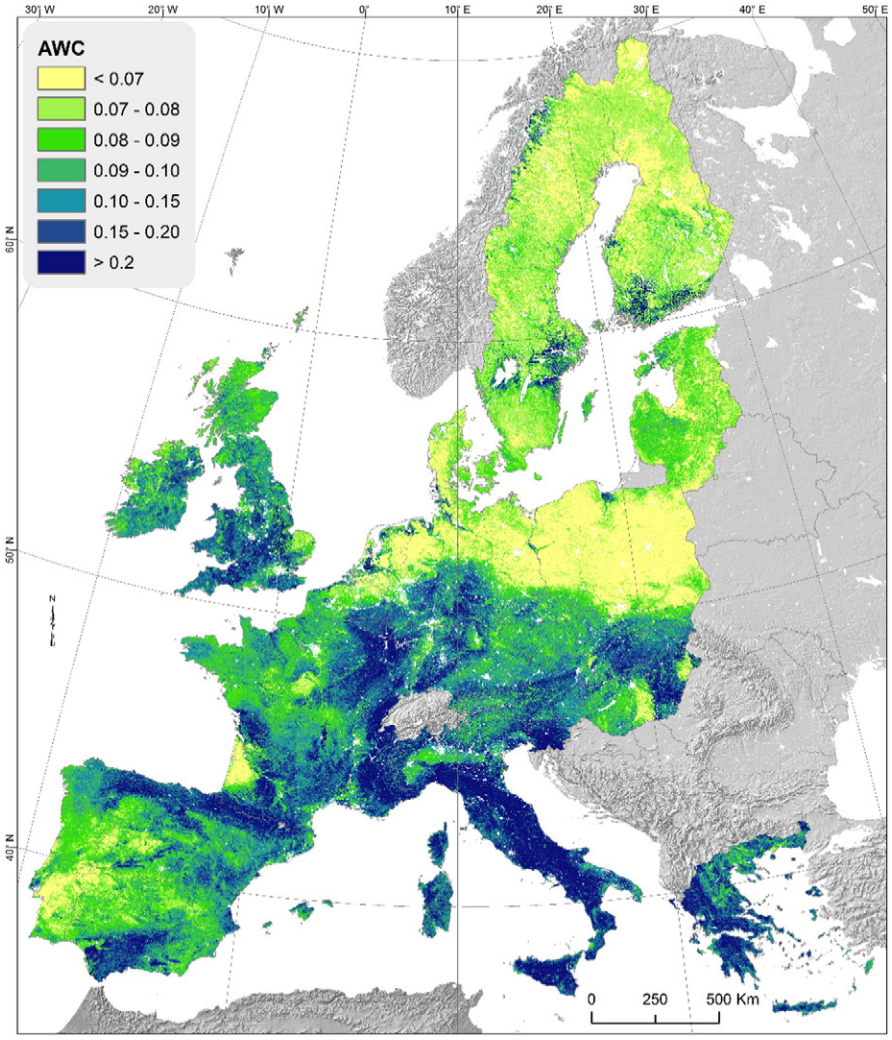

Fig. 11. Map of Available Water Capacity for the topsoil fine earth fraction.

for Europe. The testing of the MARS model, the predictors (Spectral data, Terrain, Land Cover and climatic data) used and the LUCAS topsoil dataset were tested as a way towards mapping soil physical properties at the high resolution required by the *GlobalSoilMap* specifications.

In the near future, the chemical properties (pH, ECEC, etc.) will be predicted at the same spatial resolution. Eventually, physical and chemical properties will make available seven out of the twelve soil attributes for the European Union, first at a coarser 500 m resolution, then at the *GlobalSoilMap* specified resolution.

The study also proposed the possible use of the physical properties datasets and their derived products. The data availability is a cornerstone for modellers who have no access to high spatial resolution data. With the publication of this study, modellers and in general scientists will be able to download the physical properties dataset from the European Soil Data Centre. Besides the application for soil science modelling in general, the particle size distribution datasets and their derived products can be used in different areas such as water management, hydrology, agricultural management and design of crop rotation scenarios.

Acknowledgements

The authors wish to acknowledge the SOIL team at the European Commission's Joint Research Centre (JRC) for the development, maintenance and distribution of the LUCAS dataset.

References

Aitchison, J., 1982. The statistical analysis of compositional data. J. R. Stat. Soc. Ser. B Methodol. 139–177.
Aitchison, J., 1986. The statistical analysis of compositional data. Chapman & Hall, London.
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., Mendonca-Santos, M.d.L., Minasny, B., Montanarella, L., Odeh, I.O.A., Sanchez, P.A., Thompson, J.A., Zhang, G.-L., 2014. Chapter Three — GlobalSoilMap: Toward a Fine-Resolution Global Grid of Soil Properties. In: Sparks, Donald L. (Ed.), Advances in Agronomy. Academic Press, pp. 93–134.
Asch, K., 2003. The 1:5 Million International Geological Map of Europe and Adjacent Areas: Development and Implementation of a GIS-enabled Concept; SA 3. In: Hannover (Ed.)Geologisches Jahrbuch BGR, Stuttgart (Schweitzerbart (Stuttgart), 190 pp., 45 fig., 46 tab.).
Barron, A.R., Xiao, X., 1991. Discussion: Multivariate adaptive regression splines. Ann. Stat. 67–82.
Büttner, G., Maucha, G., Kosztra, B., 2011. European validation of Land Cover changes in CLC2006 project. EARSeL Symposium, Prague.
Carré, F., Jacobson, M., 2009. Numerical classification of soil profile data using distance metrics. Geoderma 148 (3), 336–345.
Cools, N., Delanote, V., Scheldeman, X., Quataert, P., De Vos, B., Roskams, P., 2004. Quality assurance and quality control in forest soil analyses: a comparison between European soil laboratories. Accredit. Qual. Assur. 9, 688–694. http://dx.doi.org/10.1007/s00769-004-0856-4.
Dane, J.H., Topp, G.C., Campbell, G.S., Horton, R., Jury, W.A., Nielsen, D.R., van Es, H.M., Wierenga, P.J., Topp, G.C., 2002. Part 4, Physical Methods. Methods Soil Anal.
Davidian, M., Carroll, R.J., 1987. Variance Function Estimation. J. Am. Stat. Assoc. 82, 1079–1091. http://dx.doi.org/10.1080/01621459.1987.10478543.
de Brogniez, D., Ballabio, C., van Wesemael, B., Jones, R.J., Stevens, A., Montanarella, L., 2014. Topsoil Organic Carbon Map of Europe. Soil Carbon 393.
De Gruijter, J., Brus, D.J., Bierkens, M.F., Knotters, M., 2006. Sampling for natural resource monitoring. Springer.
Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Stat. 1–67.

Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resour. Res. 39, 1347. http://dx.doi.org/10.1029/2002WR001426.

Haase, D., Fink, J., Haase, G., Ruske, R., Pécsi, M., Richter, H., Altermann, M., Jäger, K.-D., 2007. Loess in Europe—its spatial distribution based on a European Loess Map, scale 1:2,500,000. Quat. Sci. Rev. 26, 1301–1312. http://dx.doi.org/10.1016/j.quascirev.2007.02.003.

Hempel, J.W., McBratney, A.B., Arrouays, D., McKenzie, N.J., Hartemink, A.E., 2014. GlobalSoilMap project history. Glob. Basis Glob. Spat. Soil Inf. Syst. 3.

Hengl, T., 2006. Finding the right pixel size. Comput. Geosci. 32, 1283–1298. http://dx.doi.org/10.1016/j.cageo.2005.11.008.

Hiederer, R., 2013. Mapping Soil Properties for Europe — Spatial Representation of Soil Database Attributes. (No. EUR26082EN), Scientific and Technical Research series. European Commission.

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

Horn, R., Domżżał, H., Słowińska-Jurkiewicz, A., Van Ouwerkerk, C., 1995. Soil compaction processes and their effects on the structure of arable soils and the environment. Soil Tillage Res. 35, 23–36.

Jones, R.J., Spoor, G., Thomasson, A., 2003. Vulnerability of subsoils in Europe to compaction: a preliminary analysis. Exp. Impact Prev. Subsoil Compact. Eur. Union 73, pp. 131–143. http://dx.doi.org/10.1016/S0167-1987(03)00106-5.

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. http://dx.doi.org/10.1111/j.1365-2389.2005.00728.x.

Jones, A., Panagos, P., Barcelo, S., Bouraoui, F., Bosco, C., Dewitte, O., Gardi, C., Hervás, J., Hiederer, R., Jeffery, S., Montanarella, L., Penizek, V., Tóth, G., Van Den Eeckhaut, M., Van Liedekerke, M., Verheijen, F., Yigini, Y., 2012. The State of Soil in Europe - A contribution of the JRC to the European Environment Agency's Environment State and Outlook Report – SOER 2010. No. EUR 25186 EN in EUR - Scientific and Technical Research series. Publications Office of the European Union.

Kibblewhite, M.G., Miko, L., Montanarella, L., 2012. Legal frameworks for soil protection: current development and technical information requirements. Terr. Syst. 4, 573–577. http://dx.doi.org/10.1016/j.cosust.2012.08.001.

King, D., Daroussin, J., Tavernier, R., 1994. Development of a soil geographic database from the Soil Map of the European Communities. Catena 21, 37–56. http://dx.doi.org/10.1016/0341-8162(94)90030-2.

Lal, R., 2004. Soil Carbon Sequestration Impacts on Global Climate Change and Food Security. Science 304, 1623–1627.

Lark, R.M., Bishop, T.F.A., 2007. Cokriging particle size fractions of the soil. Eur. J. Soil Sci. 58, 763–774. http://dx.doi.org/10.1111/j.1365-2389.2006.00866.x.

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

McLauchlan, K.K., 2006. Effects of soil texture on soil carbon and nitrogen dynamics after cessation of agriculture. Geoderma 136 (1), 289–299.

Montanarella, L., Tóth, G., Jones, A., 2011. Soil components in the 2009 LUCAS survey– in the European Union. In: Quality, Land, Information, Land Use (Eds.), G. Tóth & T. Németh). Office for Official Publications of the European Communities, Luxembourg, pp. 209–220.

Nachtergaele, F., Batjes, N., 2012. Harmonized world soil database. FAO.

Odgers, N.P., Sun, W., McBratney, A.B., Minasny, B., Clifford, D., 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma 214–215, 91–100. http://dx.doi.org/10.1016/j.geoderma.2013.09.024.

Palm, C., Sanchez, P., Ahamed, S., Awiti, A., 2007. Soils: A Contemporary Perspective. Annu. Rev. Environ. Resour. 32, 99–129. http://dx.doi.org/10.1146/annurev.energy.31.020105.100307.

Panagos, P., 2006. The European soil database 5. GEO Connex, pp. 32–33.

Panagos, P., Van Liedekerke, M., Jones, A., Montanarella, L., 2012. European Soil Data Centre: Response to European policy support and public data requirements. Land Use Policy 29, 329–338.

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

Panagos, P., Meusburger, K., Ballabio, C., Borrelli, P., Alewell, C., 2014. Soil erodibility in Europe: A high-resolution dataset based on LUCAS. Sci. Total Environ. 479–480, 189–200. http://dx.doi.org/10.1016/j.scitotenv.2014.02.010.

Plant, J.A., Whittaker, A., Demetriades, A., De Vivo, B., Lexa, J., n.d. The geological and tectonic framework of Europe, in: Geochemical Atlas of Europe 2003.

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. de L., Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vågen, T.-G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., Zhang, G.-L., 2009. Digital Soil Map of the World. Science 325, 680–681.

Saxton, K.E., Rawls, W., Romberger, J.S., Papendick, R.I., 1986. Estimating generalized soil-water characteristics from texture. Soil Science Society of America Journal 50 (4), 1031–1036.

Svendsen, J.I., Alexanderson, H., Astakhov, V.I., Demidov, I., Dowdeswell, J.A., Funder, S., Gataullin, V., Henriksen, M., Hjort, C., Houmark-Nielsen, M., Hubberten, H.W., Ingólfsson, Ó., Jakobsson, M., Kjær, K.H., Larsen, E., Lokrantz, H., Lunkka, J.P., Lysså, A., Mangerud, J., Matiouchkov, A., Murray, A., Möller, P., Niessen, F., Nikolskaya, O., Polyak, L., Saarnisto, M., Siegert, C., Siegert, M.J., Spielhagen, R.F., Stein, R., 2004. Late Quaternary ice sheet history of northern Eurasia. Quat. Environ. Eurasian North QUEEN 23, 1229–1271. http://dx.doi.org/10.1016/j.quascirev.2003.12.008.

Semmel, A., Terhorst, B., 2010. The concept of the Pleistocene periglacial cover beds in central Europe: a review. Quaternary International 222 (1), 120–128.

Tóth, G., Jones, A., Montanarella, L., 2013. The LUCAS topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union. Environmental monitoring and assessment 185 (9), 7409–7425.

Van Genuchten, M.T., 1980. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Sci. Soc. Am. J. 44, 892–898.

Wösten, J.H.M., Pachepsky, Y.A., Rawls, W.J., 2001. Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics. J. Hydrol. 251, 123–150. http://dx.doi.org/10.1016/S0022-1694(01)00464-4.