A map of the topsoil organic carbon content of Europe generated by a generalized additive model

D. de Brogniez\textsuperscript{a,b}, C. Ballabio\textsuperscript{a}, A. Stevens\textsuperscript{b}, R. J. A. Jones\textsuperscript{c}, L. Montanarella\textsuperscript{a} & B. van Wesemael\textsuperscript{b}

\textsuperscript{a}Institute for Environment and Sustainability, Land Resource Management Unit, Joint Research Centre, Via E. Fermi 2749, 21027 Ispra (VA), Italy, \textsuperscript{b}Georges Lemaître Centre for Earth and Climate Research, Earth and Life Institute, Université Catholique de Louvain, 1348 Louvain-la-Neuve, Belgium, and \textsuperscript{c}National Soil Resource Institute (NSRI), School of Applied Sciences, Cranfield University, Bedford MK430AL, UK

Summary

There is an increasing demand for up-to-date soil organic carbon (OC) data for global environmental and climatic modelling. The aim of this study was to create a map of topsoil OC content at the European scale by applying digital soil mapping techniques to the first European harmonized geo-referenced topsoil (0–20 cm) database, which arises from the Land use/Cover Area frame statistical Survey (LUCAS). A map of the associated uncertainty was also produced to support careful use of the predicted OC contents. A generalized additive model (GAM) was fitted on 85% of the dataset ($R^2 = 0.29$), using OC content as dependent variable; a backward stepwise approach selected slope, land cover, temperature, net primary productivity, latitude and longitude as suitable covariates. The validation of the model (performed on 15% of the data-set) gave an overall $R^2$ of 0.27 and an $R^2$ of 0.21 for mineral soils and 0.06 for organic soils. Organic C content in most organic soils was under-predicted, probably because of the imposed unimodal distribution of our model, whose mean is tilted towards the prevalent mineral soils. This was also confirmed by the poor prediction in Scandinavia (where organic soils are more frequent), which gave an $R^2$ of 0.09, whilst the prediction performance ($R^2$) in non-Scandinavian countries was 0.28. The map of predicted OC content had the smallest values in Mediterranean countries and in croplands across Europe, whereas largest OC contents were predicted in wetlands, woodlands and mountainous areas. The map of the predictions’ standard error had large uncertainty in northern latitudes, wetlands, moors and heathlands, whereas small uncertainty was mostly found in croplands. The map produced gives the most updated general picture of topsoil OC content at the European Union scale.

Introduction

Soil organic matter (SOM) consists of partially decayed plant residues and microorganisms and the by-products of microbial growth and decomposition. Soil OM is generally agreed to contain 58% organic carbon (OC), and in most cases, it is effectively measured as organic carbon. Soil OM is a key component of soil as it influences its structure, aggregate stability, nutrient availability, water retention and resilience (Hallett et al., 2012; Robinson et al., 2012). Through these properties, soils contribute to ecosystem dynamics and provide ecosystem services vital to human activities, such as food production or the prevention of land degradation. As the greatest terrestrial carbon pool, soils also play a key role in climate change regulation processes (Batjes, 1996) and in recent years there has been a growing demand for up-to-date soil carbon information for global climatic and environmental models. Whilst soil surveys generally collect point data, the establishment of spatially continuous data layers, as maps, is necessary for global modelling (Poggio et al., 2013).

In Europe, several soil monitoring networks are in place and measurements of soil OC content are ubiquitous (Morvan et al., 2008; Panagos et al., 2013). A first attempt to calculate OC contents at the European level was made by using pedotransfer rules (Van Ranst et al., 1995), where the relationship between OC, soil type, land use and temperature was established by expert knowledge. Jones et al. (2005) revised the latter pedotransfer rules and proposed a novel approach by applying them on a stack of spatially continuous data layers. The resulting OC estimates were presented as the first topsoil (0–30 cm) organic carbon content
map of Europe, hereafter referred as OCTOP, although the authors stressed the lack of comprehensive geo-referenced, harmonized (in sampling and analysis methodologies) soil OC data to test the reliability of their map.

Numerous studies in Europe and throughout the world have predicted OC spatial distribution at field (Chen et al., 2008), sub-national (Rawlins et al., 2009) and national (Bui et al., 2009; Meersmans et al., 2011) scales. In addition, the GlobalSoilMap.net project is the first world-wide initiative that aims to predict and then map soil properties such as OC content at a fine resolution with state-of-the-art digital soil mapping (DSM) technologies (Sanchez et al., 2009). The application of the DSM techniques to geo-referenced soil data and the large amount of remotely sensed data available today has allowed soil scientists to provide the broader scientific community with spatially continuous quantitative estimates of OC content, and their associated uncertainty (Lagacherie & McBratney, 2007). However, Jandl et al. (2014) and Minasny et al. (2013) identified the challenges of monitoring and mapping the large spatial variability of soil carbon caused by both data scarcity and heterogeneity of measurement techniques, sampling times, depths and methodologies.

In response to the scarcity of harmonized up-to-date OC data, the first topsoil survey at the European Union (EU) level was implemented in 2009 when 20,000 soil samples were taken in 23 Member States. The topsoil samples were sent to a central laboratory for physico-chemical analyses, which included total soil carbon. The LUCAS (Land Use/Cover Area frame statistical Survey) topsoil survey (Montanarella et al., 2011) offers the opportunity to replace the OCTOP dataset (Jones et al., 2005), providing a spatial database for environmental and climatic modelling with a baseline for monitoring OC content in Europe’s soils (Sanchez et al., 2009). The aim of the present work was to analyse the LUCAS-topsoil carbon data and create a map of predicted topsoil OC content, with associated uncertainty, according to DSM principles.

Materials and methods

Soil data

The LUCAS survey. The LUCAS survey consists of the photo-interpretation and consequent land use/land cover classification of 1,000,000 geo-referenced points, located at the intersections of a 2 × 2 km² grid, covering the EU. In spring and summer of 2009, 200,000 points of the 1,000,000 were selected and visited in the field for validation. A topsoil survey was conducted simultaneously at approximately 10% of these sites in 23 countries, namely Austria, Belgium, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, the Netherlands, Poland, Portugal, Slovakia, Slovenia, Spain, Sweden and the United Kingdom. Malta and Cyprus were surveyed in spring 2011. Romania and Bulgaria were surveyed in 2012 (spring and summer) and Iceland in summer 2012 and 2013. This study used the samples from the 23 countries surveyed in 2009, as well as those from Malta and Cyprus. The topsoil sampling locations were selected, excluding areas above 1000 m altitude, by conditioned Latin hypercube sampling (Minasny & McBratney, 2006) using as covariates CORINE land cover 2000 (http://www.eea.europa.eu/publications/COR0-landcover) and NASA Shuttle Radar Topography Mission (SRTM; Farr et al., 2007) digital elevation data, including derived slope, curvature and aspect (Montanarella et al., 2011). The Latin hypercube method was chosen to select sampling locations representative of European land use and topographic features.

Top-soil sampling. A first subsample was taken at the exact sampling location to a depth of approximately 15–20 cm. Four other subsamples were taken at a distance of 2 m from the central hole, at the cardinal points. The surveyors removed vegetation residues, grass and litter, if any, in order to sample only the mineral topsoil. The five subsamples were bulked to create a 500-g sample and put in a labelled plastic bag (EC, 2009).

Laboratory analysis. The air-dried and sieved (< 2 mm) soil samples were analysed for 13 physico-chemical properties: here we provide details for just carbon. Total carbon content was determined by dry-combustion in a CN analyser (VarioMax, Elementar Gmbh, Hanau, Germany). The carbonates present in the sample were then determined volumetrically by addition of hydrochloric acid and measuring the volume of CO₂ emitted (Sherrod et al., 2002). Soil organic carbon content was determined by subtracting the carbonate C content from the total carbon content.

Environmental covariates

The digital soil mapping (DSM) approach adopted fitted a statistical regression model between the soil property to predict and the independent variables at the same location. The soil property values are then predicted at unsampled locations by applying the fitted model to the covariates, collected as spatially continuous data layers. The environmental covariates used to predict OC content in this study are listed below.

Land surface parameters

NASA-SRTM digital elevation model (30 arc second resolution, equivalent to 100 m at European latitudes) and slope (derived) data were used (Farr et al., 2007).

Land cover

CORINE (CORdinate INformation on the Environment) land cover is an inventory of 44 classes coordinated by the European Environment Agency and presented as a raster dataset at 100 and 250 m resolution. The 2006 version of CORINE was used, except for Greece, for which data were not available for that year (EEA - European Environment Agency, 2012). The Greek data were extracted from CORINE2000 and simply merged with the 2006 data because the nomenclature remained unchanged. We reclassified the CORINE data into 16 classes (see Table S1).
Climatic data

Accumulated annual temperature (AAT) and the ratio of annual precipitation (P) and annual potential evapotranspiration (P0) totals were used as covariates. The former was generated from data collected for all Europe and interpolated, using an inverse-spline function and the adiabatic lapse rate (6°C decline in temperature for every 1000 m rise in altitude) at 1 km resolution for the PESERA erosion project (Kirkby et al., 2008). The ratio of annual precipitation and annual potential evapotranspiration, referred to as PP0, was generated by combining two datasets. The annual accumulated precipitation was calculated by summing the monthly total precipitation available from the WorldClim dataset for the period 1950–2000 at a spatial resolution of 1 km (Hijmans et al., 2005). Just as for temperature, P0 data were obtained from the Climaticdata map, and was generated by combining two datasets. The annual accumulated precipitation was calculated by summing the monthly total precipitation available from the WorldClim dataset for the period 1950–2000 at a spatial resolution of 1 km (Hijmans et al., 2005). Just as for temperature, P0 data were obtained from the Climaticdata map, and was generated by combining two datasets. The annual accumulated precipitation was calculated by summing the monthly total precipitation available from the WorldClim dataset for the period 1950–2000 at a spatial resolution of 1 km (Hijmans et al., 2005). Just as for temperature, P0 data were obtained from the Climaticdata map, and was generated by combining two datasets. The annual accumulated precipitation was calculated by summing the monthly total precipitation available from the WorldClim dataset for the period 1950–2000 at a spatial resolution of 1 km (Hijmans et al., 2005).

Net primary productivity

The Moderate Resolution Imaging Spectroradiometer (MODIS) annual net primary productivity (NPP) product (Running et al., 2004), available at 1 km spatial resolution, was averaged between 2000 and 2009 and used to produce a unique raster layer.

Spatial resolution of the covariates

A spatial resolution of 500 m was chosen as a compromise between the different original resolutions of the covariates. Continuous variables were rescaled by bilinear interpolation, which determined the output value of a cell based on a weighted distance average of the four nearest input cell centres. Land cover (discrete variable) was rescaled by assigning the value of the nearest neighbour cell to the rescaled, output cell. Slope was derived from the DEM at the original resolution (about 100 m) and then rescaled to 500 m; by doing this, we avoided the generation of artefacts by downscaling. Climatic data and NPP had a resolution of 1 km and were thus down-scaled. Bilinear resampling was considered to be feasible because the climatic gradients are necessarily smooth as these maps capture the effect of meteorological processes at meso (2–2000 km) and synoptic (>2000 km) scales. This choice was also supported by the fact that the variograms of climatic variables show gentle increment and ranges around 700 km.

SOC prediction model

The initial dataset was split into a calibration (85%) and a validation (15%) set by Latin hypercube sampling (Minsny & McBratney, 2006). The stratification was conditioned by the following variables: elevation, slope, net primary productivity, temperature, PP0, latitude, longitude, measured OC content and CORINE land cover. Knowing that land cover has a large impact on OC content, we developed the model on samples for which observed land cover (from the LUCAS survey) and CORINE land-cover inventory were in agreement to avoid using wrong land-cover classes to calibrate the model. However, using observed land cover (LUCAS) instead of mapped/predicted (CORINE) land cover has potentially the consequence of under-estimating the prediction error variance (Kempen et al., 2010). To check this, we fitted a model on the entire dataset and found no differences in cross-validation results.

A generalized additive model (GAM) was fitted on the calibration set: GAMs are a generalization of linear regression models in which the coefficients can be expanded as smooth functions of covariates (Hastie & Tibshirani, 1986). They are semi-parametric and can account for non-linear relationships between dependent variables and covariates (Equation (1)):

$$E(Y|X_1, X_2, \ldots, X_p) = \alpha + f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p),$$

where $X_1, X_2, \ldots, X_p$ represent the predictors, $Y$ is the response variable and $f_j$’s are the smooth functions.

As for generalized linear models, GAMs specify a distribution for the response variable $Y$ and use a link function $g$ relating the conditional mean $\mu(Y)$ of the response variable to an additive function of the predictors as follows:

$$g[\mu(Y)] = \alpha + f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p).$$

To prevent an ‘over-fit’, thin plate regression splines were fitted by maximum penalized likelihood (Wood, 2006a). A backward stepwise approach was then followed to select the best set of covariates and to determine the relative influence of each of the covariates on the overall prediction capabilities of the model (Poggio et al., 2013). The Akaike Information Criterion (AIC) and the deviances explained were calculated and compared for each of the models created (Akaike, 1974).

The selected model was then applied to the points of the validation set. Predicted and measured OC content were compared and both root mean square errors (RMSE) and normalized root mean square error (RMSE divided by the observed data range; NRMSE) were calculated. The coefficient of determination was calculated for the validation procedure.

Mapping

We refitted a model with the same covariates on all available samples by using the set of covariates identified by the methodology presented above and applied it to the stack of EU-wide spatially continuous covariates. Maps of predicted OC contents and model standard error were prepared, with urban areas, large water bodies and areas above 1000 m altitude masked out (because no topsoil samples were taken above that altitude and mapping these areas would have been pure unconstrained extrapolation). The standard error, which shows the theoretical range of deviation in the prediction made by the model, was calculated for every pixel of the created map, and was based on the sampling from the posterior covariance matrix of the fitted parameters. In order to validate the standard error map, the $z$-score for the observations in the validation set was calculated. This was performed by simulating coefficient vectors from the posterior distribution of the GAM coefficients, by drawing samples...
from a multivariate normal distribution. The samples drawn from
the posterior distribution of the coefficients were then used to gen-
erate samples from the posterior distribution for the observations
in the validation set upon which the standard error was calculated.
Finally the z-score was calculated as the ratio between the observed
error and the standard error. The distribution of the z-score should
be normal, with \( \mu \) close to 0 and \( \sigma^2 \) close to 1.

Results and discussion

Data exploratory analysis

Prior to regression analysis, the measured OC values, the environ-
mental covariates and the relationship between them were stud-
ied. OC content ranged from 1.0 g to 586.4 g kg\(^{-1}\) (mean value,
48.5 g kg\(^{-1}\)) in the calibration set and from 1.0 to 586.8 g kg\(^{-1}\)
(mean = 53.7 g kg\(^{-1}\)) in the validation set (Figure 1). The his-
tograms reveal a positively-skewed bimodal distribution with the
two modes observed around 20 and 500 g kg\(^{-1}\). The first local max-
imum highlights the predominance of mineral soils in Europe and
the second indicates the presence of organic soils. Organic soils
are classified as Histosols by the World Reference Base for Soil
Resources (IUSS Working Group WRB, 2007), which are charac-
terized by the presence of a histic or a folic horizon. The former has
a lower limit of OC content of between 12 and 18%, depending on
the clay content of the mineral fraction. The latter contains more
than 20% OC content. The threshold of 20% (200 g kg\(^{-1}\)) OC con-
tent will be taken in this work as the reference limit for the presence
of organic soils. The histograms show that at least 75% (up to the
third quartile) of the samples are mineral (Figure 1).

A correlation matrix between the dependent and independent vari-
able of our model was computed firstly to study the linear relation-
ship between OC and the covariates and secondly to prevent the
presence of collinear covariates in the regression model (Table 1).
Covariates that correlated most with OC content were latitude
(0.4) and temperature (−0.34). As for the independent variables,
the correlation matrix reveals large negative correlations between
temperature and latitude (−0.89) as well as between temperature
and \( \text{PP}_{0} \) (−0.52). Latitude and longitude also have relatively large
correlation values with several covariates.

Scatterplots of the measured OC content within the different
feature spaces of covariates were plotted for mineral and organic
soils (Figure 3). The former appear ubiquitously in the feature
spaces whereas the latter only occur along a very limited range
of covariate values. Moreover, the variation of mineral soils’ OC
content with the different covariates seemed to follow a trend

Looking at the OC content for each land cover class, we observed
that, unsurprisingly, wetlands have the largest OC content in
Europe, with a mean value of 365.1 g kg\(^{-1}\) (median = 452.5 g kg\(^{-1}\),
Figure 2). According to the LUCAS survey nomenclature, wet-
lands encompass marshes and peat bogs, which explains the
large OC content. Woodlands had the second largest OC content
(mean = 99.7 g kg\(^{-1}\), median = 41.9 g kg\(^{-1}\)), followed by shrub
lands and grasslands, which have similar values (mean = 59.1
and 42.0 g kg\(^{-1}\), median = 30.1 and 27.3 g kg\(^{-1}\), respectively).
Croplands have the smallest average OC content (mean =
18.7 g kg\(^{-1}\), median = 14.5 g kg\(^{-1}\)), although the middle 50% of
data samples in uncropped land have smaller values than
croplands (mean = 23.7 g kg\(^{-1}\), median = 10.5 g kg\(^{-1}\)).

Figure 1 Histogram and summary descriptive statistics of measured OC content in calibration (a) and validation (b) sets.

© 2014 The Authors. European Journal of Soil Science published by John Wiley & Sons Ltd on behalf of British Society of Soil Science
European Journal of Soil Science, 66, 121–134
Soil organic carbon content at pan-European level

Figure 2 Boxplot of OC content (logarithmic scale) for each land-cover class, as defined in the LUCAS survey nomenclature. The dot in the boxes is the mean, the horizontal line in the middle of the boxes is the median, while their lower and upper limits correspond to the first and third quartiles (the 25th and the 75th percentiles). The number of soil samples taken in each class is indicated along the X axis (n).

Table 1 Correlation matrix of measured OC content and covariates used in the regression model

|             | Elevation | Slope       | NPP  | Temperature | PP0   | y               | x   | OC   |
|-------------|-----------|-------------|------|-------------|-------|-----------------|-----|------|
| Elevation   | 1.00      | 0.39        | −0.06| 0.13        | −0.20 | −0.42           | −0.32| −0.09|
| Slope       | −         | 1.00        | 0.27 | 0.10        | 0.11  | −0.25           | −0.07| −0.02|
| NPP         | −         | −           | 1.00 | 0.08        | 0.32  | −0.15           | −0.16| −0.01|
| Temperature | −         | −           | −    | 1.00        | −0.52 | −0.89           | −0.43| −0.34|
| PP0         | −         | −           | −    | −           | 1.00  | 0.44            | 0.08 | 0.22 |
| y           | −         | −           | −    | −           | −     | 1.00            | 0.50 | 0.40 |
| x           | −         | −           | −    | −           | −     | −               | 1.00 | 0.16 |
| OC          | −         | −           | −    | −           | −     | −               | 1.00 | −    |

whereas organic soils showed no specific response to the range of covariate values.

SOC model

Calibration of the model. A Gamma distribution was chosen for the GAM as OC content was continuous and strictly positive. Identity and log-link function were tested. The analysis of the residuals’ Quantile-Quantile (Q-Q) plot showed better results with the log-link function which was therefore chosen for the model fitting (results not shown). The more-or-less straight line observed on the Q-Q plot of the calibration residuals suggests an approximate Gamma distribution of the regression residuals (Figure 4). The curved pattern with slope increasing from left to right suggests that the distribution is skewed to the right. The plot highlights the
Figure 3 Scatterplot between measured OC content and covariates in mineral (a) and organic (b) soils. The colour gradient indicates the density of points, with darker colours representing high density. X-axes units are elevation (m), slope (%), NPP (g C m\(^{-2}\) year\(^{-1}\)), temperature (°C year\(^{-1}\)), PP\(_0\) (−), x (degrees) and y (degrees). The correlation coefficient (r) is shown.

© 2014 The Authors. European Journal of Soil Science published by John Wiley & Sons Ltd on behalf of British Society of Soil Science

European Journal of Soil Science, 66, 121–134
potential weakness of the chosen distribution towards large OC values.

The backward stepwise selection of the covariates showed that elevation and PP0 were influencing the prediction model the least (Table 2). Slope, temperature, NPP, land cover, longitude and latitude were selected as covariates for the final regression model. The interaction between latitude and longitude was included in the model using a non-parametric function in two features using smooth interaction as scale-invariant tensor products (Wood, 2006b). The interaction accounts for the fact that the two variables belong to the same physical feature space. Elevation is often interpretable as a proxy for climatic variables in regression models predicting soil properties. The presence of temperature data in our model probably justifies the removal of elevation by the stepwise selection procedure. The correlation of −0.52 observed between temperature and PP0 explains the loss of the latter variable during the selection (Table 1). The variation in AIC scores calculated after the removal of each covariate from the model revealed that land cover is by far the most influential covariate in the model (Table 2). Several authors have also recognized the important influence of land use and land cover on a soil’s organic carbon pool (Jones et al., 2005; Meersmans et al., 2011). The combination of longitude and latitude was the second most influencing variable, followed by net primary productivity. The ‘fitting performance’ ($R^2$) was 0.29, which is similar to values found by Meersmans et al. (2011).

As well as the model performance, the smooth function presented in Figure 5 illustrates the modelled relationship between annual accumulated temperature (AAT) and OC content. Overall, OC content decreases as AAT increases, with a rapid decline observed between 1000 and 2000°C year$^{-1}$ and then a slower decline for AAT greater than 2000°C year$^{-1}$. The decrease in decomposition rate with increasing temperature has been explained by Arrhenius (1889). Below 1000°C year$^{-1}$, however, we observed an opposite trend as OC content increases with temperature. Areas where these low temperatures were recorded include the north of Scotland, Finland and Sweden, where the smallest values of NPP were also observed. The increase in OC content as AAT increases in these cold regions is related to the increase in vegetation cover that contributes to OC input. The small number of samples taken in these temperature conditions (shown on the bottom bar) and the large confidence interval observed encouraged us to regard this last trend with caution.

**Table 2** Summary table of the GAM models tested by backward step-wise removal of the covariates. The difference in Akaike Information Criterion (AIC) scores and deviance explained were calculated

| xy | Corine | Elevation | Slope | Temperature | PP0 | NPP | ΔAIC | Δdev explained |
|----|--------|-----------|-------|-------------|-----|-----|------|----------------|
| x  | x      | x         | x     | x           | x   | x   | 0    | 0              |
| −  | x      | x         | x     | x           | x   | x   | 178  | 0.7            |
| x  | −      | x         | x     | x           | x   | x   | 1976 | 6.6            |
| x  | x      | −         | x     | x           | x   | x   | 16   | 0.1            |
| x  | x      | x         | −     | x           | x   | x   | 49   | 0.2            |
| x  | x      | x         | x     | −           | x   | x   | 63   | 0.2            |
| x  | x      | x         | x     | x           | −   | x   | 19   | 0.1            |
| x  | x      | x         | x     | x           | x   | −   | 126  | 0.4            |

**Figure 5** Smooth function of temperature as produced by the GAM model. The solid line is the predicted value of OC content as a function of temperature. The upper and lower dashed lines are the 95% confidence interval. The bar at the bottom of the graph indicates the presence or absence of a sample.

**Model validation.** The validation of the model gave an $R^2$ value of 0.27, which is comparable to coefficients obtained in similar studies (Meersmans et al., 2011; Kumar et al., 2012). We believe that the relatively poor model fitting that we obtained was because of the complexity of the processes of OC accretion and mineralization operating at different spatial scales and the great variety of landscapes and soil types found across Europe (Arrouays et al., 2012). Also, the covariates, having a resolution of 500 m, may not be able to explain the exact conditions encountered at the validation profile. For instance, CORINE gives the dominant land cover, but mixed pixels are common and hence the observed land cover for the validation profile might not correspond to the CORINE class. More generally, OC variations occurring at scales smaller than the grid size of...
the map cannot be captured by the model and will not be depicted in the map. Predicted values plotted against observed OC contents had poor results for Scandinavia ($R^2 = 0.09$), but the remaining countries had better results ($R^2 = 0.28$). In those northern latitudes, the model performance was poor for the prediction of both mineral and organic soils (Figure 6). Olsson et al. (2009) and Heikkinen et al. (2013) showed that, in Sweden and Finland, OC contents in mineral soils tend to decrease as latitude increases, contrary to what is observed in other parts of Europe. Although the results of temperature smoothing indicate an increase in OC content with temperature in those regions (Figure 5), the north to south increase in OC content in mineral soils of Scandinavia is not well predicted (Figure 6b). The presence of organic soils at those latitudes is probably partly responsible for that, by putting a bias on the model towards large OC contents. As for the difficulty in predicting OC content larger than 200 g kg$^{-1}$, we observe that most organic soils in Europe are in fact under-predicted (Figure 6a,b). This can be explained by the unimodal distribution used by the model, whose mean value is tilted towards the vast majority of residuals coming from mineral soils, therefore leaving residuals from the organic part of the population in the tail of the distribution. Rawlins et al. (2009) encountered similar problems in their study in Northern Ireland and concluded that mineral and peat soils should be modelled separately. This could be achieved using for instance the methods developed in Poggio et al. (2013).

Overall, the model had an RMSE of 79 g kg$^{-1}$ (NRMSE = 85%). The model had an RMSE of 42 g kg$^{-1}$ and an $R^2$ of 0.21 for the prediction of mineral soils, as well as an RMSE of 287 g kg$^{-1}$ and an $R^2$ of 0.06 for organic soils. These results indicate that the use of a unimodal distribution results in a better prediction of the mineral soils for which the OC values are in the centre of this distribution.

When considering the land cover classes separately, predictions for croplands on mineral soils are the most accurate (Table 3). Croplands represent the second main land cover type in Europe (after woodlands) by occupying nearly a quarter of the area and they are distributed evenly across the EU (Eurostat, European Commission, 2011). Within the organic soils, the model performs better in woodland, which is the most common land cover type in Europe, with a 39% share of the total surface area. Most of the organic soil samples in the calibration set were taken in woodlands, which may explain the better performance of the model for that land cover type. Conversely, in mineral soils, the model performed the worst at predicting OC content in woodlands. A reason for this might be the great heterogeneity observed in forest soils, with measured OC contents ranging from 1 to 586 g kg$^{-1}$. Even slight differences in sampling depth and removal of the ectorganic horizons can induce large variation in OC content because of the strong vertical gradient in OC contents of forest soils.

**OC predictions and standard error maps**

The largest OC contents were observed in Ireland, the United Kingdom, Sweden, Finland, Estonia and Latvia, mostly in wetlands (peat lands), woodlands and in mountainous areas (Figure 7). These findings are in line with current knowledge on spatial distribution of topsoil OC content in Europe (Jones et al., 2005).

---

Figure 6 Predicted values plotted against observed OC contents in the validation set. Other countries (a) and Scandinavia (b).
Soil organic carbon content at pan-European level

Table 3 Model accuracy (RMSE, root mean square error, g kg\(^{-1}\)) and normalized root mean squared error (NRMSE, %) calculated for mineral and organic soils (OC content > 200 g kg\(^{-1}\)) and for each LUCAS land-cover class.

| Soils   | Land cover  | RMSE / g kg\(^{-1}\) | NRMSE / % | n  |
|---------|-------------|-----------------------|------------|----|
| Mineral | Bare land   | 31                    | 161        | 54 |
| Mineral | Cropland    | 12                    | 96         | 965|
| Mineral | Grassland   | 37                    | 126        | 653|
| Mineral | Shrub       | 44                    | 119        | 68 |
| Mineral | Wetlands    | 245                   | –          | 1  |
| Mineral | Woodland    | 74                    | 192        | 443|
| Organic | Cropland    | 412                   | –          | 1  |
| Organic | Grassland   | 290                   | 320        | 20 |
| Organic | Shrub       | 289                   | 319        | 7  |
| Organic | Wetlands    | 183                   | 295        | 4  |
| Organic | Woodland    | 290                   | 282        | 108|

The normalized root mean square error is calculated as the ratio of the root mean square error and the range of observed values.

The smallest OC concentrations were predicted to be mostly in the Mediterranean countries and also parts of France, Germany, Poland, the Czech Republic, Slovakia and Hungary, where the land cover is defined as croplands or permanent crops. Intensive management practices, as implemented on much agricultural land in Europe, increase the mineralization of soil organic matter and therefore reduce OC contents (Lal, 2002). The visual comparison of the spatial distribution of organic soils proposed by our model and by Montanarella et al. (2006) indicates good agreement in Ireland, the United Kingdom, Scandinavia and Baltic countries. However, our model seems to predict fewer organic soils in the Netherlands and in Poland. The reason for the observed discrepancies might be the differences in the methods used to produce both maps. The distribution of organic soils in Europe (Montanarella et al., 2006) was inferred by using a polygon-based approach, following the identification of different key soil types within the European Soil Map and Data-base (King et al., 1994; Heineke et al., 1998).

Large standard errors are observed in northern latitudes but also in inland wetlands or moors and heathlands (Figure 8). Few samples

![Figure 7 Map of predicted topsoil organic carbon content (g C kg\(^{-1}\)).](image-url)
were taken in the highlands of Scotland, in Wales, in south-western Ireland or in northern Sweden and Finland, where OC variation tends, moreover, to be very large (Figure 9). In all these areas, OC predictions have large standard errors. Mountain ranges such as the Alps (Italy, France and Austria), the Carpathians (the Czech Republic, Slovakia and Poland), the Apennines (Italy), the Central Massif and the Vosges (France) and the Pindus (Greece) had large standard errors in their areas below 1000 m altitude (areas above 1000 m altitude were masked out in Figure 8). The model, whose weakness at larger OC contents we understand, failed to predict the generally large OC content measured in these mountain ranges. Areas where a large standard error is estimated should be considered with caution. In contrast, areas where a small standard error is calculated (mostly corresponding to the croplands of Europe) give predictions of OC content that more accurately approximate the real values. These areas are therefore the most reliable predictions in Figure 7. The standard error map gave a z-score distribution with \( \mu = 0.043 \) and \( \sigma^2 = 2.01 \). These values are, however, biased because of the skewness of the distribution of residuals and when the right tail values are excluded \( \mu = 0.013 \) and \( \sigma^2 = 0.91 \).

Comparison with existing SOC map

Here, we compare the values predicted on the validation set of our model with those predicted by Jones et al. (2005) for the same pixel. Because the validation procedure showed a tendency of the model to under-predict severely the OC content of Scandinavian soils, we chose to separate the comparison of both models accordingly. Figure 10 shows that the prediction models generally agree on the prediction of OC content of mineral soils in ‘other countries’. Although the two models were substantially built in different ways (pedotransfer rule or regression model), the logic behind their prediction of OC content is quite similar. Jones et al. (2005) and we recognized the large influence of land cover and temperature on OC dynamics in soils and fed the respective models with such information, which led to similar outcomes. In addition, whilst our input data were actual measured OC values, OCTOP predicts OC content from measurements from representative soil profiles synthesized by expert soil scientists from across Europe (King et al., 1994; Van Ranst et al., 1995). Areas of disagreement are mostly where our model predicts the OC content for mineral soils whilst the OCTOP...
model predicts OC content greater than 200 g kg\(^{-1}\); in Scandinavia, there is not any clear pattern of agreement or disagreement.

The two models can also be compared on mean OC estimates by land-cover class (Figure 11). As the validation set used for the GAM did not encompass enough points in every land-cover class to allow for such comparison, we extracted predictions from the map. Globally, we can see that the average predictions compare well, except for the inland wetlands class for which the GAM estimates a larger OC content. The observed match is probably through the use of a common layer in the prediction models, namely land cover. Also, the error bars are always smaller in the case of the GAM. Because in OCTOP soil variation is expressed by discrete soil classes, a mismatch between the large soil units and land cover classes at larger scales can occur. This creates a more heterogeneous range of predicted values per land cover class, resulting in wider standard deviations. The latter supports our initial assumption that OCTOP could be improved by using more detailed soil spatial information layers or by modelling soil properties at a finer resolution using DSM techniques.

**Conclusion**

Our study has proposed the first topsoil OC content map of Europe which is based on direct harmonized measurements stored in a soil database, produced using DSM techniques. The model shows a fairly good accuracy for most of the EU \((R^2 = 0.28)\), except for Scandinavia \((R^2 = 0.09)\) where organic topsoils predominate. This was also indicated by the validation procedure, which revealed that the unimodal distribution imposed by the fitted model was not suited to accurately predicting the bimodal distribution of OC content of all soils in Europe. Hence modelling mineral and organic soils separately might give better results. However, if the final output is to be spatial, the lack of an accurate spatial data layer predicting the localization of organic soils in Europe will remain a limitation for producing a continuous map using both models. The comparison of our map with OCTOP underlined the influence of land cover on OC content in soils and it showed that the use of (discrete) soil classes instead of continuous fields to express soil OC variation is a major source of uncertainty.
Figure 10 Scatterplot of OC content predictions proposed by Jones et al. (2005) and by our model in Scandinavia (b) and other countries (a).

Figure 11 Comparison of mean OC content predictions (g kg\(^{-1}\)) per land-cover class, for OCTOP (Jones et al., 2005) and for the GAM model (this study). Dots are mean OC values and the error bars represent two standard deviations (2\(\sigma\)).
The map produced gives the most up-to-date general picture of topsoil OC content at the European Union scale and is not intended to be a substitute for national-scale or local maps that are based on more detailed spatial information. Moreover, it is important that the uncertainty associated with the predicted values from this study is understood by the end-users and should encourage careful use and interpretation of the spatial values. The maps produced in this study will be freely available for download from the European Soil Data Centre website (http://eusoils.jrc.ec.europa.eu/).

Supporting Information

The following supporting information is available in the online version of this article:

Table S1. Reclassification of the Corine Land Cover (CLC) 44 classes into 16 new classes.

Acknowledgements

The authors wish to acknowledge the SOIL teams at the European Commission’s Directories General Eurostat, Environment and Joint Research Centre (JRC) for their encouragement and continuous effort to maintain and improve the LUCAS dataset. The support of Dr Alan Belward, Head of the Land Resource Management Unit, and of Dr Maria Betti, Director of the Institute for Environment and Sustainability, Joint Research Centre (JRC), is also acknowledged.

References

Akaike, H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19, 716–723.

Arrhenius, S. 1889. Uber die reaktionsgeschwindigkeit bei der inversion von rohrzucker durch sauren. Zeitschrift für Physikalische Chemie, 4, 226–248.

Arrouays, D., Marchant, B.P., Saby, N.P.A., Meersmans, J., Orton, T.G., Arrhenius, S. 2009. Soil organic carbon content at pan-European level.

EC – European Commission, Eurostat. 2009. LUCAS 2009 Technical reference document C1. Instruction for surveyors. URL http://epp.eurostat.ec.europa.eu/portal/page/portal/lucas/documents/LUCAS2009_C1-Instructions_Revised20130925.pdf [accessed on 16 September 2014]

EEA – European Environment Agency 2012. EIONET : Corine Land Cover 2006 [WWW document]. URL http://sia.eionet.europa.eu/CLC2006 [accessed on 23 January 2013].

Eurostat, European Commission 2011. Land Cover and Land Use Statistics at Regional Level [WWW document]. URL http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Land_cover_and_land_use_statistics_at_regional_level#Further_Eurostat_information [accessed on 14 November 2013].

Farr, T.G., Rosen, P.A., Caro, E., Crippen, R., Duren, R., Hensley, S. et al. 2007. The shuttle radar topography mission. Reviews of Geophysics, 45, 1–33.

Hallett, P.D., Loades, K.W. & Krümmelbein, J. 2012. Soil physical degradation: threats and opportunities to food security. In: Soils and Food Security (eds R.E. Hester & R.M. Harrison), pp. 198–226. Royal Society of Chemistry, Cambridge.

Hastie, T. & Tibshirani, R. 1986. Generalized additive models. Statistical Science, 1, 297–318.

Heikkilä, J., Ketoja, E., Nuutinen, V. & Regina, K. 2013. Declining trend of carbon in Finnish cropland soils in 1974–2009. Global Change Biology, 19, 1456–1469.

Heinze, H.J., Eckelmann, W., Thomasson, A.J., Jones, R.J.A., Montanarella, L. & Buckley, B. 1998. Land Information Systems: Developments for Planning the Sustainable Use of Land Resources. Office for Official Publications of the European Communities, Luxembourg.

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. International Journal of Climatology, 25, 1965–1978.

IUSS Working Group WRB 2007. World Reference Base for Soil Resources 2006, First Update 2007. FAO, Rome.

Jandl, R., Rodeghiero, M., Martinez, C., Cotrufo, M.F., Bampa, F., Van Wesemael, B. et al. 2014. Current status, uncertainty and future needs in soil organic carbon monitoring. Science of the Total Environment, 468–469, 376–383.

Jones, R.J.A., Hiederer, R., Rusco, E. & Montanarella, L. 2005. Estimating organic carbon in the soils of Europe for policy support. European Journal of Soil Science, 56, 655–671.

Kempen, B., Heuvelink, G.B.M., Brus, D.J. & Stooyvorgel, J.J. 2010. Pedometric mapping of soil organic matter using a soil map with quantified uncertainty. European Journal of Soil Science, 61, 333–347.

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.

Kirkby, M.J., Irvine, B.J., Jones, R.J.A., Govers, G. and PESEERA team. 2008. The PESEERA coarse scale erosion model for Europe. I. – Model rationale and implementation. European Journal of Soil Science, 59, 1293–1306.

Kumar, S., Lal, R. & Liu, D. 2012. A geographically weighted regression kriging approach for mapping soil organic carbon stock. Geoderma, 189–190, 627–634.

Lagacherie, P. & McBratney, A.B. 2007. Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping. In: Digital Soil Mapping an Introductory Perspective (eds P. Lagacherie, A.B. McBratney & M. Vliz), pp. 3–22. Elsevier, Oxford.

Lal, R. 2002. Introduction. In: Agriculture Practices and Policies for Carbon Sequestration in Soil (eds J.M. Kimble, R. Lal & R.F. Follett), pp. 3–5. CRC Press, Boca Raton, FL.

Meersmans, J., van Wesemael, B., Goidts, E., Van Molle, M., De Baets, S. & De Ridder, F. 2011. Spatial analysis of soil organic carbon evolution in Belgian croplands and grasslands, 1960–2006. Global Change Biology, 17, 466–479.

Minsny, B. & McBratney, A.B. 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences, 32, 1378–1388.
Minasny, B., McBratney, A.B., Malone, B.P. & Wheeler, I. 2013. Digital Mapping of Soil Carbon. Advances in Agronomy, Volume 118. Academic Press, Newark, NJ.

Montanarella, L., Jones, R.J.A. & Hiederer, R. 2006. The distribution of peatland in Europe. Mires & Peat, 1, 1–10.

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

Morvan, X., Saby, N.P.A., Arrouays, D., Le Bas, C., Jones, R.J.A., Verheijen, F.G.A. et al. 2008. Soil monitoring in Europe: a review of existing systems and requirements for harmonisation. Science of the Total Environment, 391, 1–12.

Olsson, M.T., Erlandsson, M., Lundin, L., Nilsson, T., Nilsson, Å. & Stendahl, J. 2009. Organic carbon stocks in Swedish Podzol soils in relation to soil hydrology and other site characteristics. Silva Fennica, 43, 209–222.

Panagos, P., Hiederer, R., Van Liedekerke, M. & Bampa, F. 2013. Estimating soil organic carbon in Europe based on data collected through an European network. Ecological Indicators, 24, 439–450.

Poggio, L., Gimona, A. & Brewer, M.J. 2013. Regional scale mapping of soil properties and their uncertainty with a large number of satellite-derived covariates. Geoderma, 209–210, 1–14.

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. European Journal of Soil Science, 60, 44–54.

Robinson, D.A., Emmett, B.A., Reynolds, B., Rowe, E.C., Spurgeon, D., Keith, A.M. et al. 2012. Soil natural capital and ecosystem service delivery in a world of global soil change. In: Soils and Food Security (eds R.E. Hester & R.M. Harrison), pp. 41–68. Royal Society of Chemistry, Cambridge.

Running, S.W., Nemani, R.R., Heinsch, F.A., Zhao, M., Reeves, M. & Hashimoto, H. 2004. A continuous satellite-derived measure of global terrestrial production. BioScience, 54, 547–560.

Sanchez, P.A., Ahamed, S., Carré, F., Hartemink, A.E., Hempel, J., Huisings, J. et al. 2009. Digital soil map of the world. Science, 325, 680–681.

Sherrod, L.A., Dunn, G., Peterson, G.A. & Kolberg, R.L. 2002. Inorganic carbon analysis by modified pressure-calcimeter method. Soil Science Society of America Journal, 66, 299–305.

Van Ranst, E., Thomasson, A.J., Daroussin, J., Hollis, J.M., Jones, R.J.A., Jamagne, M. et al. 1995. Elaboration of an extended knowledge database to interpret the 1:1,000,000 EU Soil Map for environmental purposes. In: European Land Information Systems for Agro-environmental Monitoring, European Soil Bureau Research Report, pp. 71–84. Office for Official Publications of the European Communities, Luxembourg.

Wood, S.N. 2006a. Generalized Additive Models: An Introduction with R. Chapman & Hall/CRC, Boca Raton, FL.

Wood, S.N. 2006b. Low rank scale invariant tensor product smooths for generalized additive mixed models. Biometrics, 64, 1025–1036.