Geoderma 355 (2019) 113912

Contents lists available at ScienceDirect

Geoderma

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

# Mapping LUCAS topsoil chemical properties at European scale using Gaussian process regression

Cristiano Ballabio[a,*], Emanuele Lugato[a], Oihane Fernández-Ugalde[a], Alberto Orgiazzi[a], Arwyn Jones[a], Pasquale Borrelli[b], Luca Montanarella[a], Panos Panagos[a]

[a] European Commission, Joint Research Centre (JRC), Ispra, Italy
[b] Environmental Geosciences, University of Basel, Switzerland

ARTICLE INFO

Handling Editor: Alex McBratney

*Keywords:*
pH
Cation Exchange Capacity
Nitrogen
Phosphorus
Potassium
C:N ratio
Mapping
Gaussian process regression

ABSTRACT

This paper presents the second part of the mapping of topsoil properties based on the Land Use and Cover Area frame Survey (LUCAS). The first part described the physical properties (Ballabio et al., 2016) while this second part includes the following chemical properties: pH, Cation Exchange Capacity (CEC), calcium carbonates ($CaCO_3$), C:N ratio, nitrogen (N), phosphorus (P) and potassium (K). The LUCAS survey collected harmonised data on changes in land cover and the state of land use for the European Union (EU). Among the 270,000 land use and cover observations selected for field visit, approximately 20,000 soil samples were collected in 24 EU Member States in 2009 together with more than 2000 samples from Bulgaria and Romania in 2012. The chemical properties maps for the European Union were produced using Gaussian process regression (GPR) models. GPR was selected for its capacity to assess model uncertainty and the possibility of adding prior knowledge in the form of covariance functions to the model.

The derived maps will establish baselines that will help monitor soil quality and provide guidance to agro-environmental research and policy developments in the European Union.

## 1. Introduction

Globally, soil and environmental challenges (climate change, pollution, water scarcity, biodiversity decline) are increasing dramatically (IPBES, 2019). Organizations such as the United Nations Convention to Combat Desertification (UNCCD), the Food Agriculture Organization (FAO) and the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) have recognized that soil is under threat globally (Montanarella, 2015). The development of solutions to combat land degradation requires data collection, expert knowledge, and scenario analysis through the modelling of soil properties and functions (Hartemink, 2015). During the past two decades, the soil science community has developed regional, continental and worldwide soil maps and databases, which have been used for soil resource assessment and risk evaluation (Arrouays et al., 2017).

During the past decade, the increased use of digital soil mapping (McBratney et al., 2003) approaches became a solution to increased requests for spatial soil data coming from research organizations, policy makers and the private sector. The extensive development of digital soil mapping models has been facilitated by the exponential increase in the availability of remote sensing data, computing power, and the development of Geographic Information Systems (Minasny and McBratney, 2016). Digital soil mapping has important advantages in the prediction of soil properties (e.g. evidence-based, open access data and software, transparent and repeatable methodology, etc.) compared to conventional soil mapping approaches (Minasny et al., 2018). Among others, Grundy et al. (2015) mapped 11 soil properties in Australia using a geostatistical approach, Padarian et al. (2017) contributed to the development of a Global Soil Map by modelling eight properties for Chile, Mansuy et al. (2014) generated national maps of soil properties for managed forests in Canada, Adhikari et al. (2014) developed soil organic carbon content maps for Denmark while Poggio and Gimona (2014) modelled soil organic carbon stocks in Scotland.

In the European Union (EU), Ballabio et al. (2016) have developed physical properties datasets (silt, clay, sand and coarse fragments) for the EU, together with maps of derived products (bulk density, available water capacity) using the Land Use and Cover Area frame Survey (LUCAS) topsoil database. In addition, de Brogniez et al., 2015 and Yigini and Panagos, 2016 have mapped the soil organic carbon content (expressed as percent and carbon stocks) using LUCAS topsoil database

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

https://doi.org/10.1016/j.geoderma.2019.113912
Received 4 March 2019; Received in revised form 26 July 2019; Accepted 12 August 2019
Available online 28 August 2019

and geostatistical models.

The objective of this paper is to produce chemical soil properties datasets using LUCAS topsoil database and advanced digital soil mapping methodologies. In detail, this paper proposes a soil mapping model for interpolating the 22,000 surveyed points in LUCAS for chemical soil properties at continental scale. Finally, we developed datasets for soil pH, Cation Exchange Capacity (CEC), calcium carbonates ($CaCO_3$), and total phosphorus, potassium and nitrogen, plus derived products based on soil organic carbon and nitrogen (e.g. C:N ratio) covering the 26 EU Member States (excluding Croatia and Cyprus).

It is not the objective of the paper to challenge any local or regional map of chemical properties implemented with higher density of surveyed points. Moreover, this paper will provide only a cursory explanation for the reasons for different spatial patterns on chemical properties as our focus is the digital soil mapping model development.

## 2. Mapping chemical properties at European scale

This section gives an overview of existing activities relevant to chemical properties using digital soil mapping. Soil chemical properties have a spatial dependence; their spatial patterns depend on soil forming processes (including climate, parent material and weathering), topography, climate, vegetation, time and anthropogenic influences (Yost et al., 1982).

**Soil pH** is dependent on the parent material, climate and soil organic carbon. Where precipitation levels are high, soil pH decreases over time through acidification due to leaching of base cations and corresponding build-up of hydrogen ions. In dry environments, where chemical weathering and leaching are less intense, soil pH may be neutral or alkaline, as a result of evaporation of alkaline groundwater. In general higher rainfall rates result in acid soils and the water balance seems to act globally as a main driving factor for soil pH (Slessarev et al., 2016), while topography and mineralogy may act as secondary drivers. At European scale, soil pH datasets have been produced by the Forum of European Geological Surveys (FOREGS) (Salminen et al., 2005) using 1588 soil samples across 28 countries. The project Geochemical Mapping of Agricultural and Grazing Land Soil in Europe (GEMAS) sampled 2200 points in agricultural land and 2118 points in permanent grasslands in 2008–2009. In 2012, a harmonised pH dataset was compiled by the JRC and disseminated through the European Soil Data Centre (ESDAC) (Panagos et al., 2012). The ESDAC pH dataset is based on 12,333 measurements from 11 different data sources which is a significant limitation, due to heterogeneity in the measurements, compared to the LUCAS soil sampling scheme. The recent publication of SoilGrids (Hengl et al., 2014) was an important advancement for soil pH data availability at global scale.

**Calcium carbonates** are derived from the weathering of lime-rich parent material (Lal, 2007). Soil carbonates, most commonly represented by calcium carbonate, have multiple functions in soils. Firstly, they help slowing soil acidification by acting as a buffer to increased levels of aluminium and hydrogen ions, thus also preventing the uptake of heavy metals by plants. Another important function is the stabilization and the improvement of both soil organic carbon content and soil structure. Sarmadian et al. (2010) used geostatistical methods to produce a topsoil calcium carbonates map. At continental scale, Wilford et al. (2015) predicted soil calcium carbonate concentrations in Australia using data from 1311 sites. According to Wilford et al. (2015), the key predictors of $CaCO_3$ in Australia include mean annual precipitation, mean annual radiation, soil types, mean annual temperature, and the MODIS vegetation coefficient of flatness.

**Cation Exchange Capacity** (CEC) is often considered as an indicator of soil quality and measures the ability of soil to hold and exchange cations (Saidi, 2012). CEC is related to stable aggregates and texture properties (Bronick and Lal, 2005). CEC is often estimated using pedotransfer rules or other algorithms such as artificial neural networks (Amini et al., 2005). Few studies have estimated CEC at country scale and, among them, Khaledian et al. (2017) derived correlations between CEC and physical attributes (clay, silt, sand) and chemical attributes (pH). Bishop and McBratney (2001) applied different geostatistical models to estimate the CEC in northern New South Wales, Australia.

**Table 1**
Naming and description of the covariates used in the models (the covariate code corresponds to the one used in the variable relevance plots of Fig. 2**).**

| Covariate code | Description |
|---|---|
| mir_PCAb*i* | Component *i* of PCA transformed of MODIS multitemporal Mean Infrared band for year 2009 |
| nir_PCAb*i* | Component *i* of PCA transformed of MODIS multitemporal Near Infrared band for year 2009 |
| red_PCAb*i* | Component *i* of PCA transformed of MODIS multitemporal Red band for year 2009 |
| blue_PCAb*i* | Component *i* of PCA transformed of MODIS multitemporal blue band for year 2009 |
| pheno_MODIS_LAEA.1 | Periodic component *i* of MODIS NDVI time series Fourier harmonic analysis |
| trend_MODIS_LAEA.1 | Trend component *i* of MODIS NDVI time series Fourier harmonic analysis |
| tmax*i*_500 | Average max temperature of month *i* from WorldClim |
| tmin*i*_500 | Average min temperature of month *i* from WorldClim |
| prec*i*_500 | Average precipitation of month *i* from WorldClim |
| bio*i*_500 | Bioclimatic index *i* from WorldClim |
| y | Latitude |
| x | Longitude |
| elevation | Elevation |
| valley height | Valley height index |
| gen_surface | Smoothed Elevation |
| ls | RUSLE topographic factor (Slope Length and Steepness factor) |
| aacn | Altitude above channel network |
| airflow_height | Effective Air Flow Heights (Böhner and Antonić, 2009) |
| downsl_dist_grad | Downslope Distance Gradient (Hjerdt et al., 2004) |
| corine.*i* | Class *i* of CORINE land cover |
| geo.*i* | Class *i* of ESDB parent material |

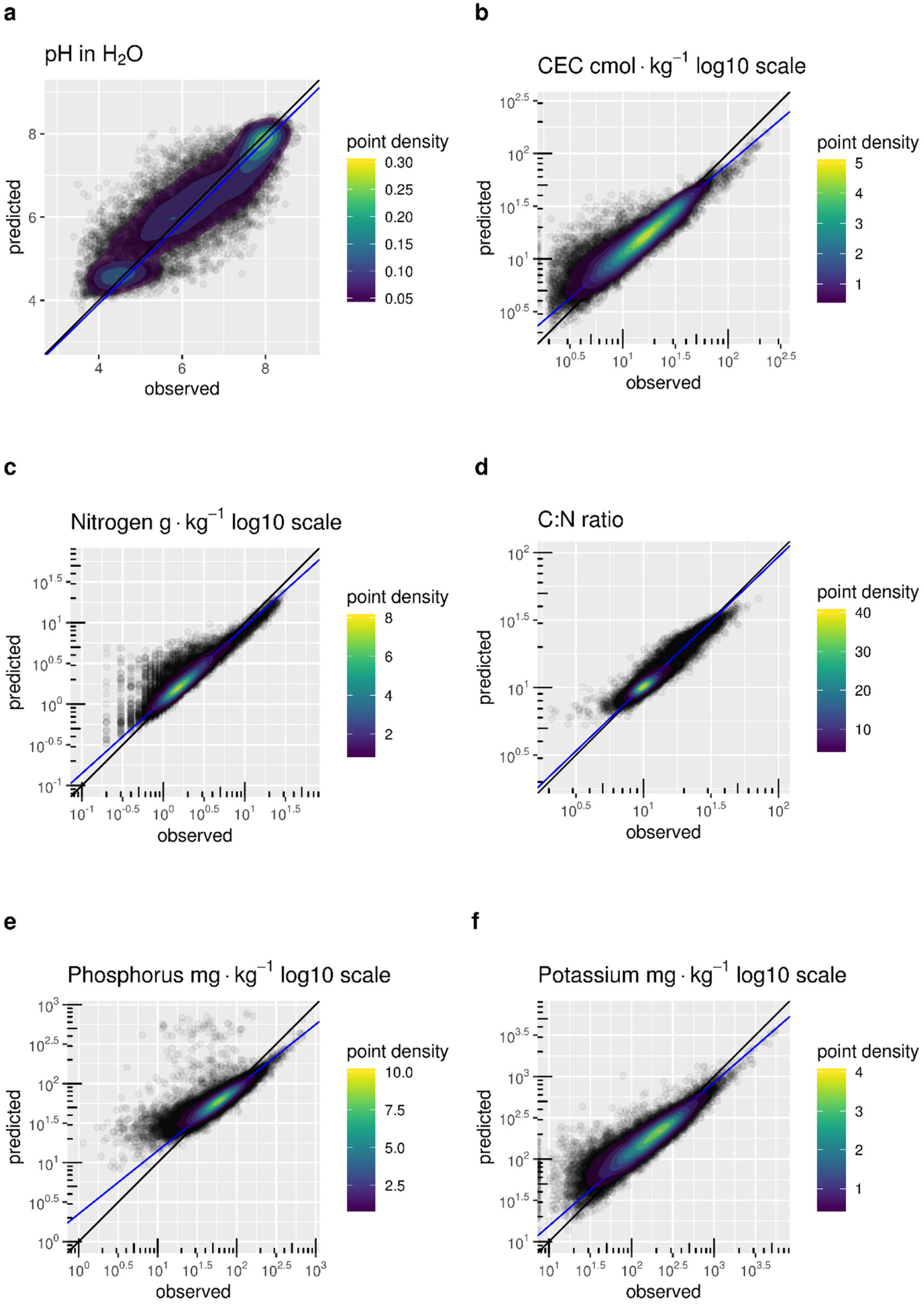
**Table 2**
GPR regression performance metrics for chemical properties. The mean and median values in LUCAS data are given as a reference.

| | Mean | Median | RMSE | MAE | RSE | $R^2$ |
|---|---|---|---|---|---|---|
| $CaCO_3$ g·kg$^{-1}$ | 52.78 | 1.00 | 78.29 | 40.84 | 0.76 | 0.61 |
| CEC cmol·kg$^{-1}$ | 16.08 | 12.70 | 11.02 | 6.64 | 2.55 | 0.35 |
| C:N ratio | 13.20 | 10.89 | 1.97 | 1.26 | 0.12 | 0.91 |
| Nitrogen g·kg$^{-1}$ | 12.10 | 1.80 | 2.40 | 1.2 | 2.59 | 0.60 |
| Phosphorous mg·kg$^{-1}$ | 37.61 | 29.10 | 17.52 | 11.70 | 0.82 | 0.74 |
| Ph in $H_2O$ | 6.30 | 6.30 | 0.78 | 0.62 | 0.57 | 0.65 |
| Ph in $CaCl_2$ | 5.70 | 5.80 | 0.68 | 0.53 | 0.36 | 0.76 |
| Potassium mg·kg$^{-1}$ | 199.17 | 142.20 | 121.89 | 70.98 | 0.53 | 0.75 |

**Nitrogen** (N) spatial distribution is not only affected by natural ecological processes, but also impacted by intensive human activities (K. Wang et al., 2013). This is an important challenge for accurate predictive mapping at regional scales. The C:N (organic carbon to nitrogen) ratio is an index of the organic matter turnover and nitrogen availability due to mineralization or immobilization of soil nitrogen. Using more than 4000 soil profiles around the world from the World Inventory of Soil Emission Potentials (WISE) database, Batjes (1996) estimated a wide range of mean C:N ratios for the 0–30 cm topsoil starting from 9.9 for arid Yennosols to 25.8 for Histosols.

**Phosphorus** (P) concentrations are also influenced by human activity. Fertilization can result in higher levels of P, especially in higher yields crops where high input of P fertilizers are reported (Tóth et al., 2014). Modern agriculture is much dependent on phosphorus fertilizers, and P supply is strategically critical at global level. Using a geostatistical model, Rossel and Bui (2016) mapped phosphorus stocks in Australia at approximately 90 m resolution.

**Potassium** (K) has different functions for plant life; it is a



*(caption on next page)*

**Fig. 1.** Predicted vs observed values for the topsoil properties discussed in Section 4.2. The blue line represent a linear fit for predicted-vs-observed data. The black line is the diagonal and the contours represent point densities. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

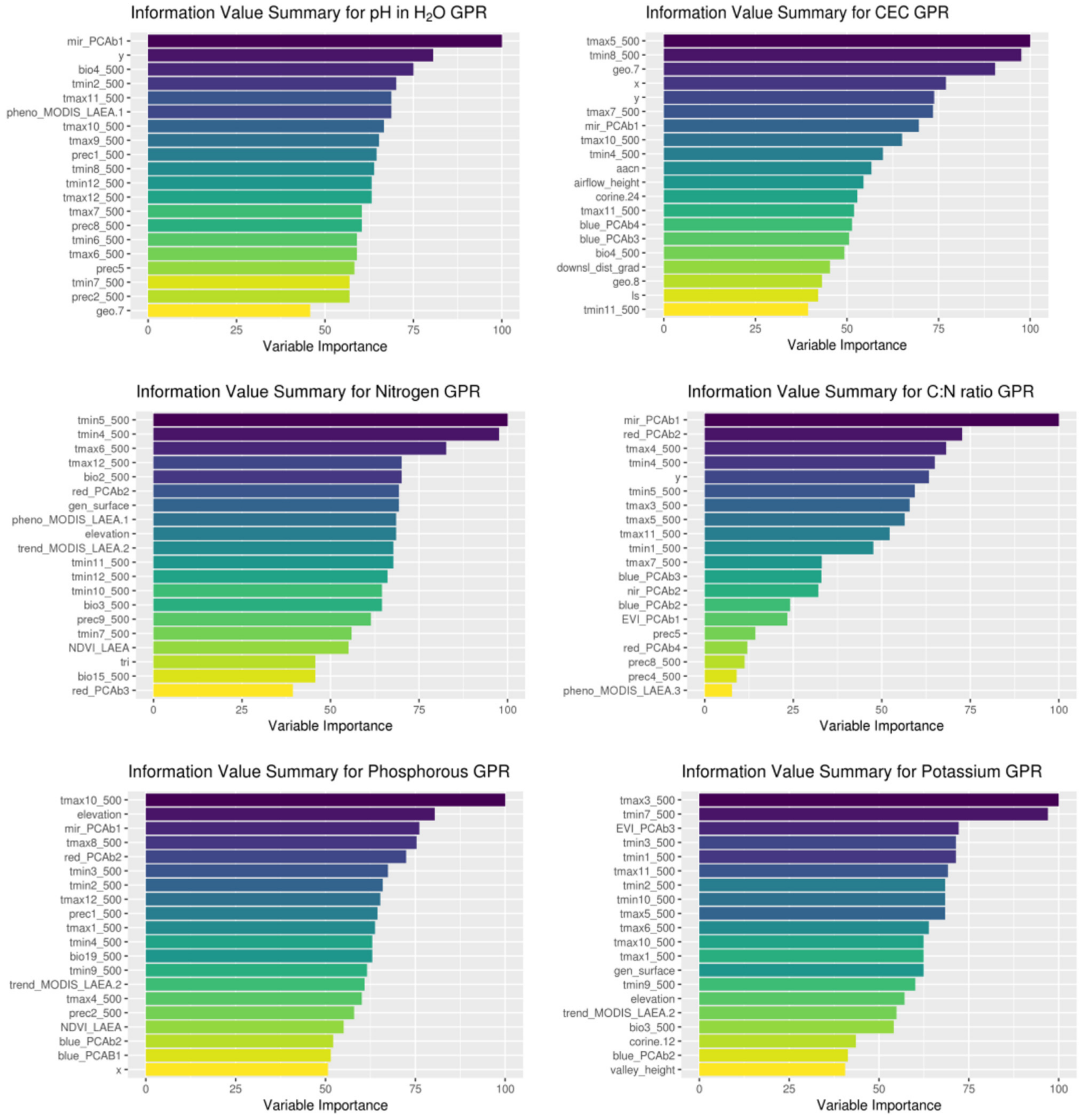


**Fig. 2.** Variable importance metrics for the topsoil properties discussed in Section 4.2.

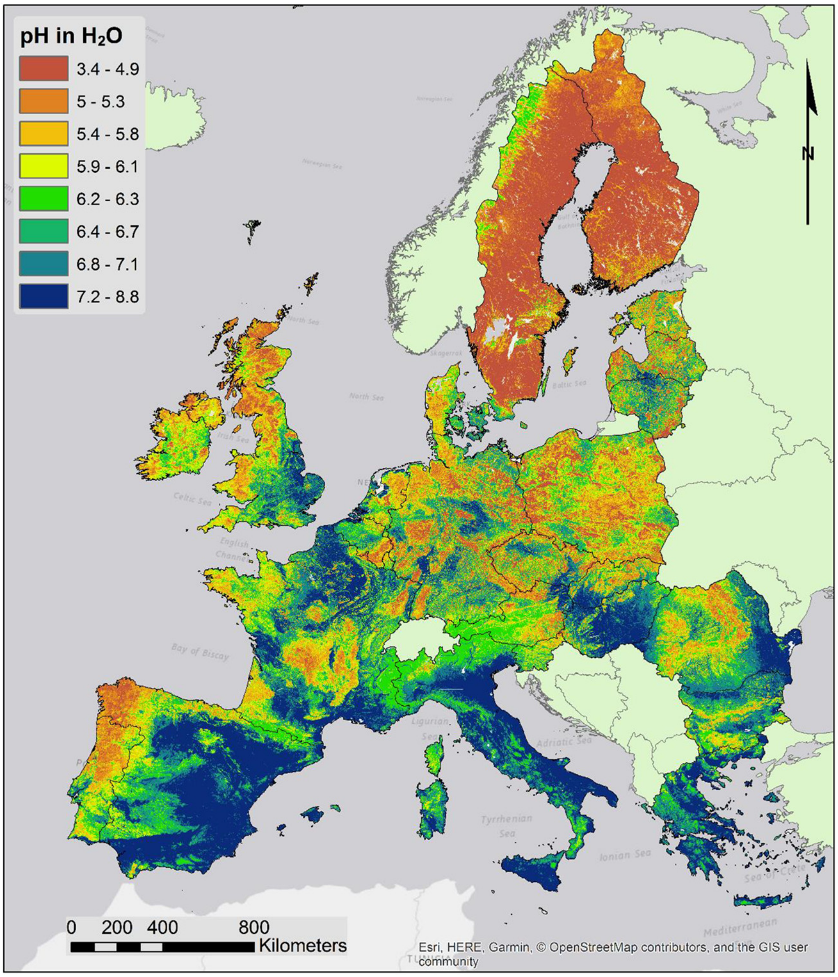
constituent of enzymes and acts as a regulator of drought tolerance and water use (M. Wang et al., 2013). In the soil, the principal sources of potassium are feldspars and micas, which release K during weathering (Hillel, 2008). Potassium depletion from soil is quite uncommon as cation exchange prevents leaching. Few studies have mapped potassium at continental scale; among them, Prado et al. (2008) mapped the potassium distribution in Brazil using a limited amount of data and extrapolated it to the whole country.

## 3. Material and methods

This section describes the main data input which is the LUCAS topsoil database, how it has been compiled based on the survey and the laboratory analysis.

### 3.1. Land Use/Land Cover Area Frame Survey (LUCAS) topsoil database

The Land Use/Land Cover Area Frame Survey (LUCAS) is a project to monitor land use and land cover changes across the EU. The LUCAS



Esri, HERE, Garmin, © OpenStreetMap contributors, and the GIS user community

**Fig. 3.** Map of topsoil pH in water.

survey is performed every three years, with the latest published LUCAS dataset dating back to 2015. It now covers all the 28 EU countries and includes field observations at more than 273,000 points. Soil samples are taken in about 10% of the surveyed locations every 6 years. The first LUCAS soil survey was done in 2009 collecting 19,969 topsoil samples (0–20 cm) from 25 out of 28 EU countries, excluding Romania, Bulgaria and Croatia (Orgiazzi et al., 2018). In the 2012 LUCAS survey, 2034 topsoil samples were collected from Bulgaria and Romania following the standard procedure of 2009. The overall sampling density of this pan-European soil survey is nearly one soil sample every 196 Km$^2$ (Panagos et al., 2013), which means one sample about every 14 km × 14 km. In this paper, we used the chemical properties based on 2009–2012 LUCAS topsoil as the analysis of the 2015 soil samples is still ongoing.

The LUCAS topsoil dataset is the most comprehensive and harmonised soil dataset at European scale, which allows pan-EU studies on the distribution of physical properties (clay, silt and sand) (Ballabio et al., 2016), soil erodibility (Panagos et al., 2014), soil organic carbon (de Brogniez et al., 2015) and the modelling of heavy metals diffuse pollution such as copper (Ballabio et al., 2018). The number of points selected is based on a stratification in order to cover all possible land uses (based on CORINE land cover classes) and country surface (Carre et al., 2013). Orgiazzi et al. (2018) described in detail the soil sampling procedure and the standards that the surveyors should follow. The soil samples were analysed for the percentage of coarse fragments, particle-size distribution (silt, clay, sand), pH, organic carbon, calcium carbonate, soluble phosphorous, total nitrogen, extractable potassium, Cation Exchange Capacity (CEC) and multispectral properties (Tóth et al., 2013a, 2013b). Due to problems in labelling, tagging, geo-referencing and mismanagement, 321 soil samples were excluded from LUCAS topsoil database, resulting in 21,682 total records.

### 3.2. Laboratory analysis of soil samples for chemical properties

The sample analysis was performed by a single laboratory, contributing to data comparability avoiding uncertainties due to analysis based on different methods or different calibrations in case of multiple laboratories. In a first phase, LUCAS topsoil samples were analysed for their physical and chemical properties following ISO standard procedures. In a later stage, an additional analysis for heavy metals was performed.

### 3.3. Auxiliary variables

To support the spatial predictions of soil properties, a series of datasets or covariates were selected according to their possible influence on soil chemical properties. The spatial resolution of the covariates was set to 250 m, as a compromise between the resolution of the Moderate-resolution Imaging Spectroradiometer (MODIS) data (500 m), the finer resolution of the DEM (25 m) and the coarser WorldClim climatic (1 km) datasets (Fick and Hijmans, 2017). Overall 100 numeric and 99 dummy covariates were considered in the first steps of the analysis. The dummy covariates were obtained from the coding of the categorical variables classes (CORINE and parent material type) into dichotomous variables.

After feature selection, the name and a description of the covariates

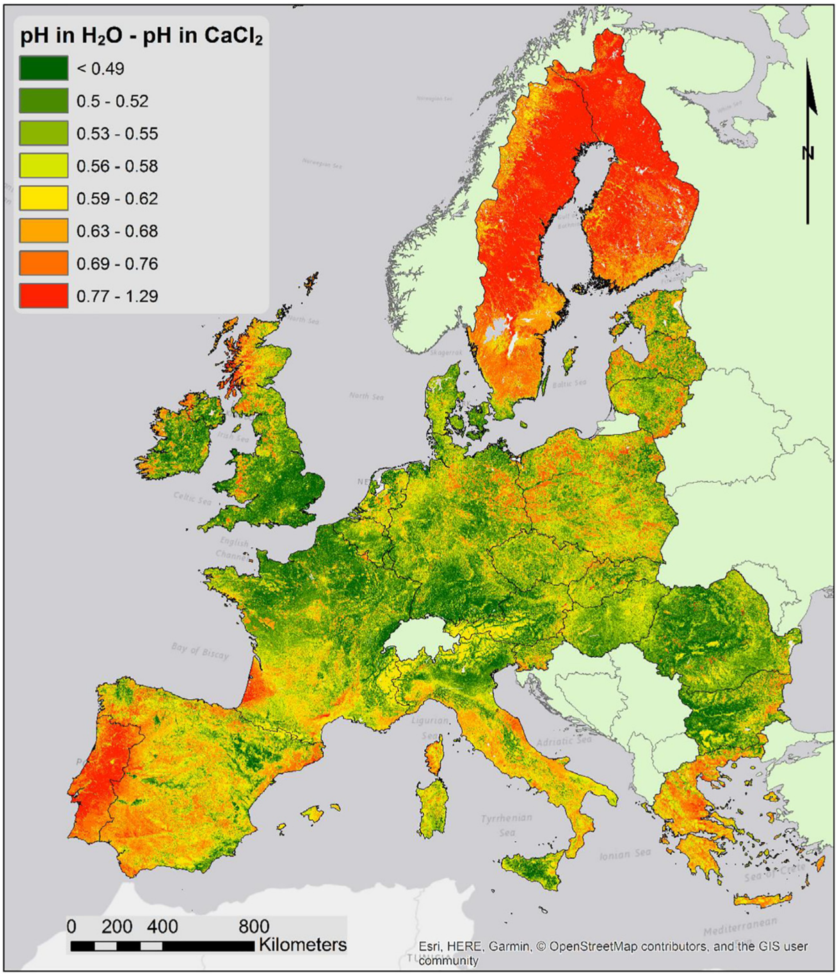


**Fig. 4.** Map of the difference between pH in water and pH in $CaCl_2$ 0.01 M solution.

retained in the final model is given in Table 1.

Cyprus was excluded from the analysis due to missing covariates.

### 3.3.1. MODIS and derived data

A series of MODIS image products for 2009 was collected; in particular, the MODIS Global vegetation indices (Didan, 2005). These products are characterised by a spatial resolution between 250 and 500 m and a temporal resolution of 16 days. The products include blue, red and near- and mid-infrared reflectance, centered at 469 nm, 645 nm, and 858 nm respectively. 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 $NDVI = (NIR - RED)/(NIR + RED)$, where NIR and RED stand for the spectral reflectance measurements acquired in the near-infrared and visible (red) 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 is defined as:

$$EVI = g \bullet \frac{NIR - RED}{NIR + c1 \bullet RED - c2 \bullet BLUE + L} \tag{1}$$

where NIR, RED, and BLUE are the respective surface reflectance in the corresponding spectral bands, $L$ is the canopy background adjustment, and $c1$ and $c2$ 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: $L = 1$, $c1 = 6$, $c2 = 7.5$, and $g$ (gain factor) = 2.5.

Phenological indices were derived from MODIS data using a first order harmonic model on the EVI and NDVI multi-temporal data. The harmonic uses a discrete Fourier processing that decomposes temporal curves in a linear trend plus amplitude, variance and phase metric terms. The harmonic model can be defined as

$$\widehat{Y}_t = \alpha_0 + \sum_{j=1}^{m} \alpha_j \, cos\left(\frac{j2\pi t}{l}\right) + \beta_j \, sin\left(\frac{j2\pi t}{l}\right) \tag{2}$$

where $\widehat{Y}_t$ is the vegetation index value, $t$ is the time value for a given pixel, $l$ is the cycle length (yearly) and $m$ is the order of the trigonometric polynomial and coincides with the number of harmonics of the expansion (set as one in this study), $\alpha_j$ and $\beta_j$ are the Fourier coefficients.

Harmonic analysis using Fourier series, has been used to model the temporal changes in the vegetation cover using satellite data for several decades (Menenti et al., 1993; Moody and Johnson, 2001; Olsson and Eklundh, 1994) and provides better spatial information on the different types of vegetation cover than using composite images alone.

Additionally, a Principal Component Analysis (PCA) transformation of the full MODIS 16 day images time series was performed for each band in order to extract relevant features. The PCA projects the time correlated input images into uncorrelated PCA components ordered according to their variance. Thus, the first few components account for most of the time related variation in each MODIS band.

### 3.3.2. Terrain parameters

The EU-DEM digital elevation model (Bashfield and Keim, 2011) was used to derive land features at a resolution of 25 m for all Europe.

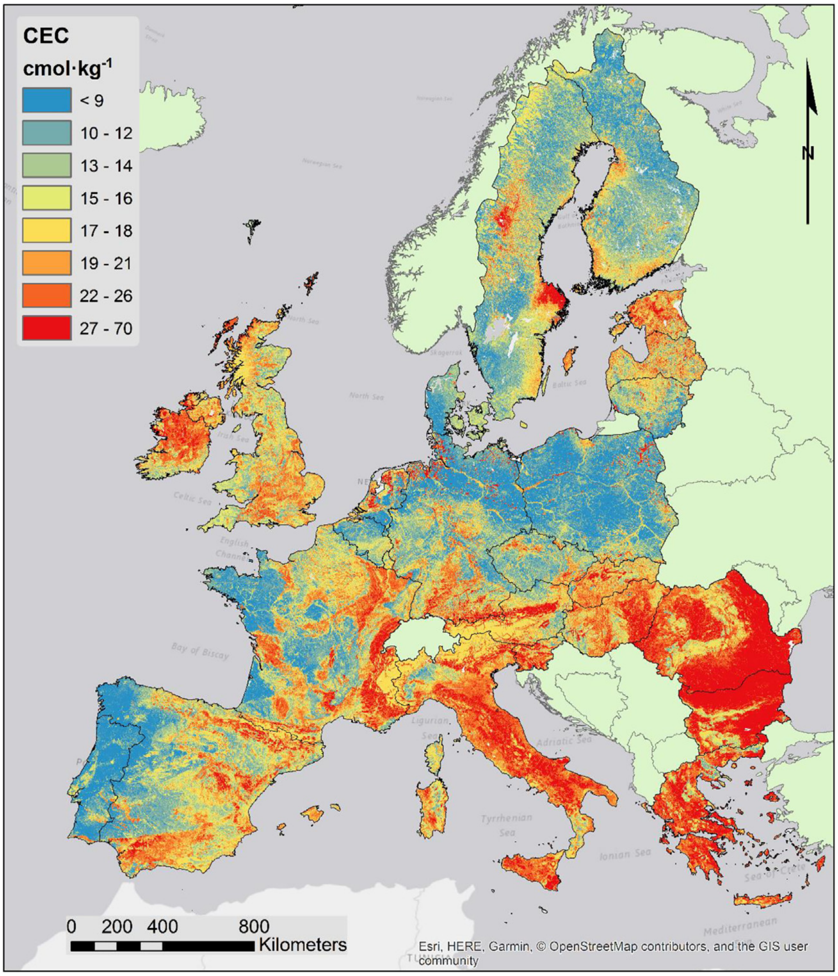


**Fig. 5.** Map of topsoil CEC.

Both the DEM and the derived surface parameters were then rescaled to 250 m. The derivation of land surface parameters was made using the SAGA GIS software. Among the various parameters derived and tested, the most relevant 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).

### 3.3.3. Land cover

The CORINE (CORdinate INformation on the Environment) is a raster format land cover database comprising 44 classes. CORINE is derived from Earth observation satellites using computer-aided photo-interpretation. The nominal scale of CORINE is 1:100,000 with a minimum mapping unit (MMU) of 25 ha and a change detection threshold of 5 ha. The CORINE dataset was used to represent the spatial distribution of land use and 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, 2014).

### 3.3.4. Climate data

Monthly temperature averages and extremes, and monthly average precipitation values 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 explicitly in the model in the form of monthly values of minimum and maximum temperature and monthly rainfall rates. Also the bioclimatic variables (Temperature and precipitation indexes) of WorldClim were included in the analysis. Given the high collinearity of climate data, a careful feature selection procedure was applied in the model training stage.

### 3.3.5. Legacy soil data and parent material geochemistry

In the first stage of this study, the European Soil Database (ESDB) (Panagos et al., 2012) was considered as a possible covariate to characterise soil properties. In this context, the ESDB was utilised as a multinomial variable by identifying and labelling soil types. However, the use of the ESDB soil data was found to provide little improvement to the model outcome and was then removed from the analysis. Nonetheless, the data within the ESDB was used to create a map of the parent material geochemistry that was included in the model.

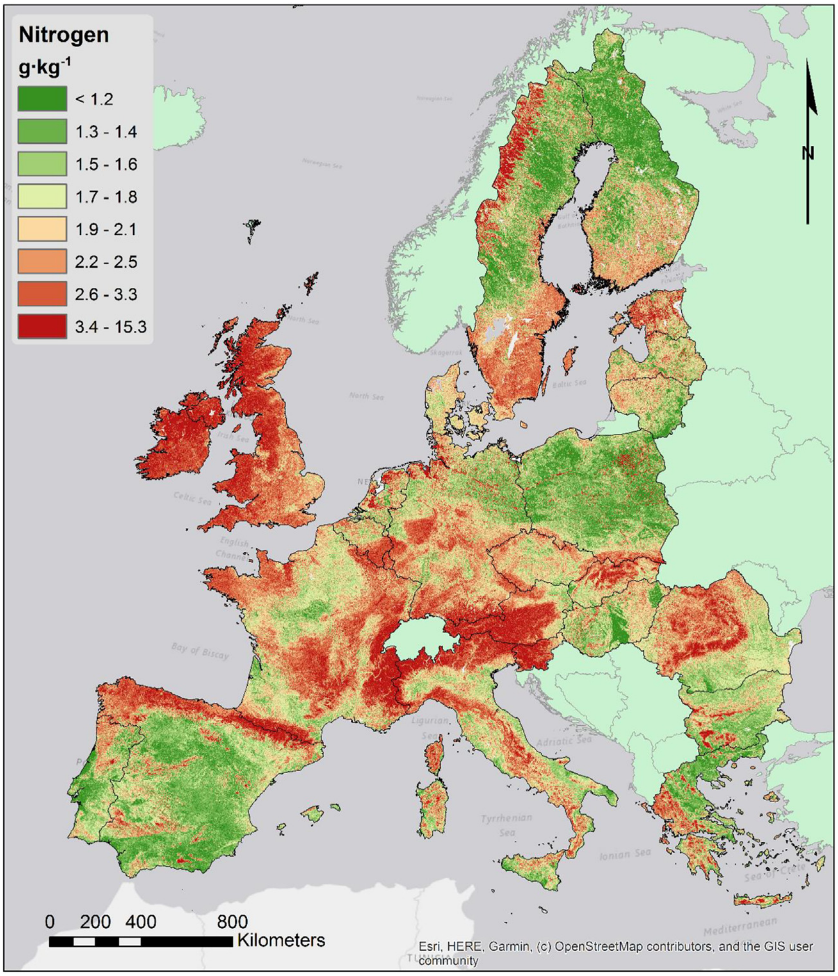
## 3.4. Gaussian process regression models for chemical properties

In order to assess the relation between environmental features and soil chemical properties distribution, Gaussian Process Regression (GPR) (Rasmussen and Williams, 2006) was utilised for inference and mapping.

GPR assumes that the output $y$ of a function $f$ with input x can be expressed as

$$y = f(\mathrm{x}) + \epsilon \tag{3}$$

where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$. This is analogous to linear regression. However in



Esri, HERE, Garmin, (c) OpenStreetMap contributors, and the GIS user community

**Fig. 6.** Map of topsoil nitrogen content.

GPR, not only the error term $\epsilon$, but also $f$ is treated as a random variable. In GPR, $f(x)$ is distributed as a Gaussian process

$$f(x) \sim \mathcal{GP}(\mu(x), k(x, x^*)) \tag{4}$$

where $f(x)$ is defined by its mean $\mu(x)$ and covariance $k(x, x^*)$.

The covariance function $k$ is also known as the kernel of the GPR and models the dependence of the function values between different values of x. In this respect, GPR is equivalent to kriging (Stein, 2012); however, while kriging is usually performed in geographical space, GPR is applied on an arbitrary number of covariates. The choice of the appropriate kernel is based on the structure, in terms of smoothness and peculiar patterns, of the data itself.

In this study, the Matérn kernel function (Stein, 2012) was used. The Matérn function is quite flexible as it can model data with different smoothness; the function can be written as

$$k(x, x^*) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}}{\ell}|x - x^*|\right)^{\nu} B_\nu\left(\frac{\sqrt{2\nu}}{\ell}|x - x^*|\right) \tag{5}$$

where $\nu$ and $\ell$ are positive adjustable parameters, $B_\nu$ is a modified Bessel function of the second kind of order $\nu$ and $\Gamma$ is the Gamma function. The $\ell$ acts as a scale parameter, while $\nu$ controls the process smoothness. In general, values of $\nu$ are kept within the range between 1/2, where the process becomes rough, and 7/2, where it becomes difficult to distinguish between finite values of $\nu$ and $\nu \to \infty$.

The kernel function is equivalent to a covariance function in kriging and its value can be considered as a measure of similarity between the two feature vectors.

GPR can be seen as a Bayesian Nonparametric approach to regression, where the function from the Gaussian processes takes values in a (possibly infinite) function space. Defining y as the vector of observed values of the dependent variable and X as the matrix of the corresponding covariates and defining $y^*$ as a set of points to be predicted with the corresponding matrix of covariates $X^*$, a random vector can be drawn from the join prior distribution of functions as

$$y^* \sim \mathcal{N}(0, K(X^*, X^*)) \tag{6}$$

where $K(X^*, X^*)$ is the covariance matrix between inputs at points to be predicted. However, any vector drawn from the prior will provide no knowledge about the observed data. In order to get the posterior distribution over functions, the joint prior distribution must be restricted to contain only those functions which agree with the observed data. So given that the joint distribution of y and $y^*$ is

$$\begin{bmatrix} y \\ y^* \end{bmatrix} \sim \mathcal{N}(0, [K(X, X) + \sigma_n^2 I K(X, X^*) K(X, X^*) K(X', X^*)]) \tag{7}$$

(where $K(X,X)$ is the covariance matrix between all observed points inputs (covariates), $K(X, X^*)$ is the covariance between observed points and points to be predicted and $\sigma_n{}^2 I$ is the identity matrix multiplied by the estimated (or presumed) variance of the observations) predictions for the new points $y^*$ corresponding to the covariates matrix $X^*$ can be derived as

$$y^* = K(X^*, X)[K(X, X) + \sigma_n^2 I]^{-1} y \tag{8}$$

and the variances for the elements of $y^*$ can be obtained from the diagonal of the covariance matrix $COV(y^*)$

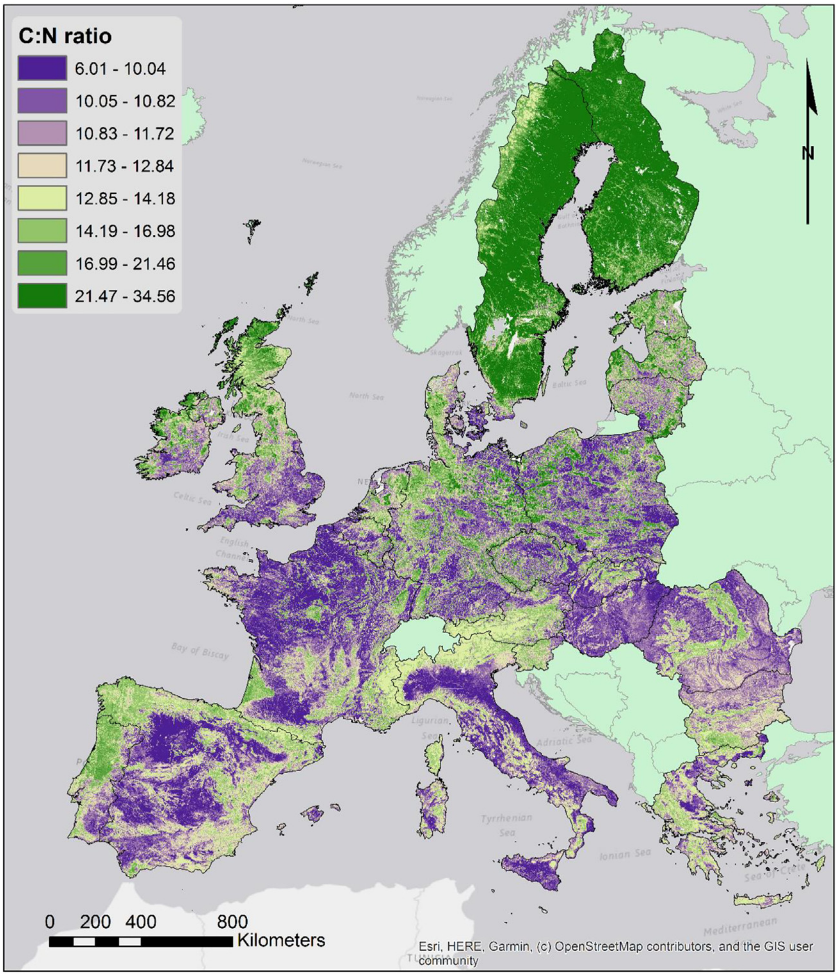


**Fig. 7.** Map of the topsoil C:N ratio.

$$\mathrm{COV}(y^*) = K(\mathrm{X}^*, \mathrm{X}^*) - K(\mathrm{X}^*, \mathrm{X})[K(\mathrm{X}, \mathrm{X}) + \sigma_n^2 \mathrm{I}]^{-1} K(\mathrm{X}, \mathrm{X}^*) \tag{9}$$

Moreover, as the posterior distribution of Eq. (7) can be rewritten as

$$\mu(\mathrm{x}) = \sum_{i=1}^{t} \mathrm{w}_i k(\mathrm{x}, \mathrm{x}_i) \tag{10}$$

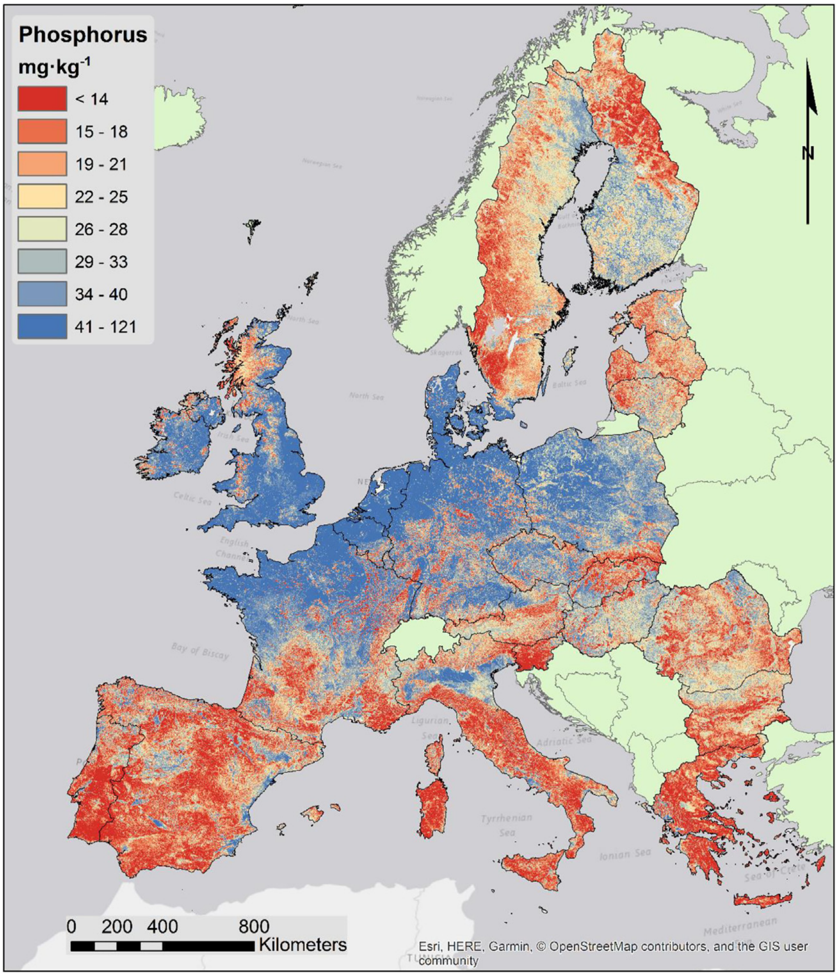
(where $\mathrm{x}_i$ is an observed value in X, and the weights come from the vector $\mathrm{w} = [K(\mathrm{X}, \mathrm{X}) + \sigma_n{}^2 \mathrm{I}]^{-1} \mathrm{y}$), GPR is effectively equivalent to a linear regression where inputs are projected into an higher dimensional space using basis functions (the kernel), predictions are then obtained by weighting the input values to the input values of the point to be predicted. In this manner GPR retains the conceptual simplicity of linear regression while having the capability of fitting arbitrarily complex relations of machine learning approaches.

An advantage of GPR over other machine learning approaches is that the process models both the expectation and the variance of the random variable, thus allowing mapping the prediction uncertainty. Moreover, the GPR allows the specification of the input data noise, so if prior knowledge about it is known, it can be used to avoid overfitting the data.

Another advantage of using GPR is that any linear combination of kernels is itself a valid kernel. This property can be used to model data that has different scale dependent patterns (Ballabio and Comolli, 2010) or are a composition of periodic and trend components. In this study, a single and a composition of two Matérn kernels were tested. Since the composition requires the tuning of more parameters, $\nu$ was kept constant across kernels in the composition. While the composition performed generally better than the single kernel, the gain was not significant enough as to justify the extra tuning time required. Therefore, we applied the GPR using a single kernel.

While GPR is a powerful technique, its main drawback is its computational complexity. Given that the computational burden scales as $\mathcal{O}(N^3)$ for model fitting, $\mathcal{O}(N)$ for model prediction and $\mathcal{O}(N^2)$ for variance prediction (where $N$ is the number of observations) the number of covariates is usually kept as small as possible. The computational scaling is especially problematic in spatial mapping where the number of predictions is easily in the order of millions of raster cells. To ease this issue it is possible to use the Nyström kernel matrix approximation (Drineas and Mahoney, 2005) to compress matrices to a more manageable size. Moreover, it is advisable to use massive parallel processing in order to split the raster data into more manageable subsets as the prediction task is easy to parallelize.

The GPR models parameters were tuned using a repeated k-fold cross-validation, with k = 10, in order to avoid overfitting. Moreover, model performance was also evaluated using the same procedure. Models selection was performed by Simulated Annealing (SA) (Kirkpatrick et al., 1983) in order to select the best set of covariates and thus reducing the chance of collinearity. SA was also utilised to estimate the relative information value of each covariate. As the selection is based on the reduction in the k-fold cross-validation error estimates, changes in the error due to a given predictor being added or removed from the model are tracked during the SA. This track of error changes is then aggregated as the variable information value.



Esri, HERE, Garmin, © OpenStreetMap contributors, and the GIS user community

**Fig. 8.** Map of topsoil phosphorous content.

## 4. Results

### 4.1. Model performance

In this study, GPR is quite efficient in predicting soil chemical properties with values of $R^2$ ranging from 0.91 to 0.35. In general, the properties more related with the vegetation cover (nitrogen content, pH) are the most successfully modelled, while properties such as CEC result in higher errors as measured by Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). This is likely due to the influence of other factors (e.g. soil clays mineralogy, soil age, etc.), that have little or no direct connection with the covariates included in this study. In particular, for soil age and mineral weathering, there are no available covariates that could be used at this scale. Moreover, properties such as CEC are inherently difficult to measure precisely, making their modelling subject to a higher noise to signal ratio.

GPR regression performance metrics for chemical properties (Table 2) shows the performances of the models in terms of RMSE, MAE, Relative Squared Error (RSE) and $R^2$; RMSE and MAE give error values in the measurement unit of the original variable, so their values can be compared with the mean and median values. RSE and $R^2$ are metrics that can be used to compare the performance of the different models. The values were computed from repeated k-fold cross-validation. For maps discussed in Section 4.2 the plots of predicted vs observed values are shown in Fig. 1, while variable information values are shown in Fig. 2. The plots of Fig. 1 seem to suggest the presence of some prediction bias for several of the properties. However, the prediction bias is present only for extreme values. In general, values within the 95th percentile are quite well predicted with residuals normally distributed and showing absence of bias.

### 4.2. Topsoil chemical properties maps

This section will discuss some of the properties mapped, namely:

1. The topsoil pH in water and the difference between topsoil pH in water and in $CaCl_2$ 0.01 M solution
2. The topsoil total nitrogen content and the topsoil C:N ratio
3. The topsoil total phosphorus content and the topsoil total potassium content

These properties were selected due to their distinctive spatial distribution and relevance in soil management.

#### 4.2.1. Topsoil pH

The map of topsoil pH (Fig. 3) shows a clear influence of the geochemical makeup of soil parent material. Areas where carbonate rocks are present show higher pH levels; this is particularly obvious in areas where soil erosion can enhance the influence of the parent material such as in the area surrounding the Mediterranean Sea. In particular most of Spain, southern France, Italy and Greece have neutral to alkaline soils. The effect of geology is also quite evident in Northern France and most of United Kingdom. Nevertheless, climate can also influence pH and sometime overcomes the effect of the parent material. This is quite evident in Ireland and north-eastern Spain where rainfall is intense enough to leach the topsoil, resulting in acid soils. Vegetation

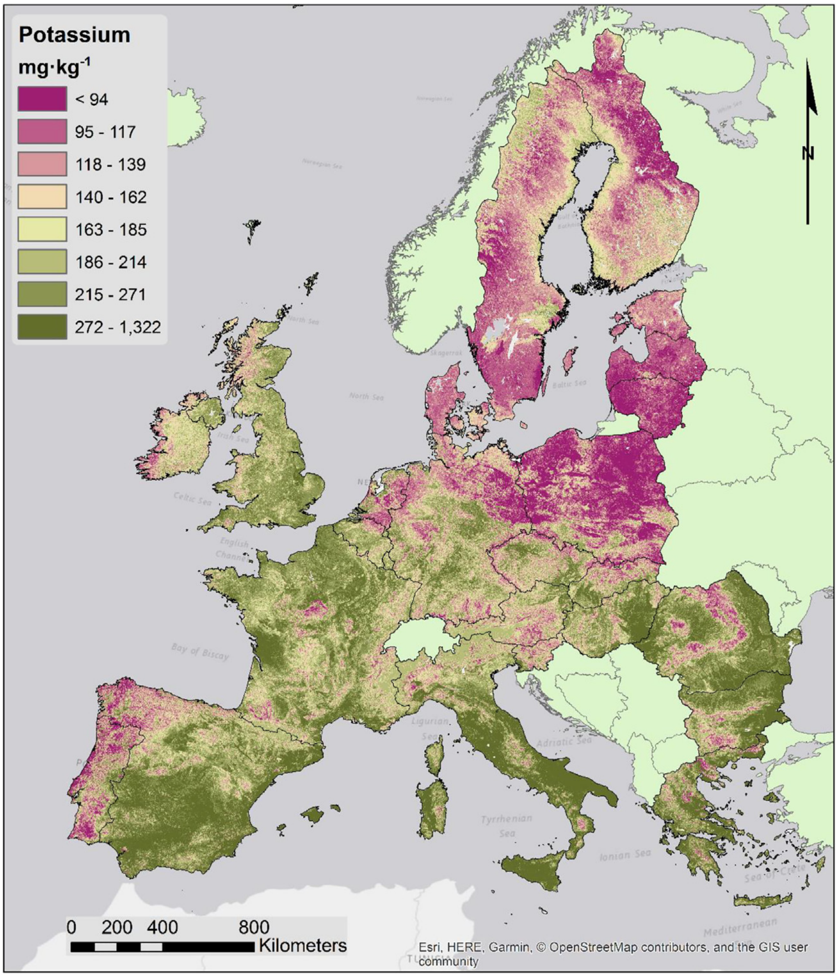


**Fig. 9.** Map of topsoil potassium content.

cover is another factor influencing soil pH; in general forest cover, especially of coniferous species, tend to lower topsoil pH as they return fewer base cations to the soil with the litter. On the contrary, steppe grassland can result in increased topsoil pH as in the plains of Hungary where Chernozem soils typically have a topsoil enriched in calcium ions from the underlying loess deposits. The effect of vegetation is even more striking at shorter scales where the difference of pH between patches of forest and cultivated soils can be quite abrupt.

The map of pH in $CaCl_2$ solution (not shown) has a similar outline as the map of pH in water. Despite similarities the map of pH in $CaCl_2$ is useful for determining the soil liming potential (Schofield and Taylor, 1955). Moreover, the difference between the two pH measures can give an idea of the exchangeable acidity of the topsoil (Fig. 4). Exchangeable acidity includes more or less ionized acid functions, weak organic acids and easily exchangeable cations. Commonly it is caused by clay hydrolysis, which results in some of $Al^{3+}$ cations passing into exchangeable positions; this occurs naturally in some processes, like podsolization, but can be exacerbated by anthropic activities causing rain acidification or by acidifying fertilizers. As shown by Fig. 4, greater differences in pH values are found in areas where podsolization is the prevalent pedogenetic process, such as Scandinava and the Atlantic coast of Iberia and the west coast of France.

### *4.2.2. Topsoil Cation Exchange Capacity*

The map of CEC (Fig. 5) is mostly influenced by the distribution of clay in the topsoil. A comparison between clay distribution (Ballabio et al., 2016) and CEC shows many similarities. Remarkably, CEC is also quite influenced by topography, where areas of sediment accumulation have general higher values of CEC. This is quite visible in The Netherlands, Northern Germany and Poland, where the areas surrounding rivers have higher than average CEC. The geochemistry of the parent material also influences CEC. Besides limestone, where the relation with higher clay content is evident, soils developed on calcareous and marl rocks also tend to accumulate clay as a consequence of the leaching of the carbonates leaving the more stable clays behind.

### *4.2.3. Topsoil nitrogen and C:N ratio*

The distribution of topsoil nitrogen (Fig. 6) is highly correlated with soil organic carbon, given that nitrogen is a main component of soil organic matter. While their ratio can vary, some carbon rich soils are also nitrogen rich, at least in terms of absolute quantities. Given this relation, it is quite clear that vegetation cover and climate are the main drivers in the distribution of nitrogen. As the map in Fig. 6 shows, forests and grasslands areas tend to have higher nitrogen content. Forests of Scandinavia, or those of the mountain areas are clearly outlined by the map. Climate also acts as a main driving force influencing nitrogen content along the Atlantic area; in particular, the United Kingdom and Ireland show higher N concentrations due to a fresh and humid climate which favours organic matter accumulation. Soil texture also plays a role in stabilizing organic matter and thus nitrogen. Areas with coarser soils, such as most of Poland, tend to have less nitrogen even if other conditions are favourable (e.g. vegetation, climate).

While the nitrogen concentration is relevant for assessing stocks and potential $N_2O$ emissions, the ratio between carbon and nitrogen can better represent the differences in the organic matter composition.

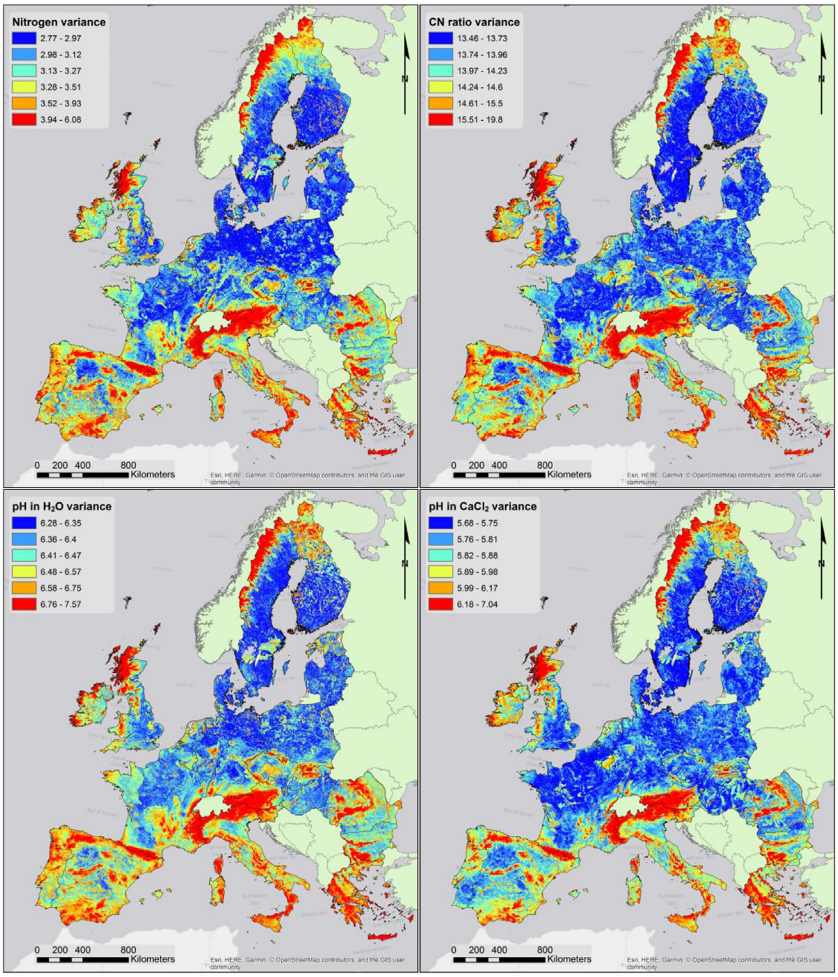


**Fig. 10.** GPR prediction variance for nitrogen, C:N ratio, pH in water and $CaCl_2$ 0.01 M solution.

Where higher rates correspond to more oligotrophic soils, typical of coniferous forests, or to peatland soils, lower rates are typical of more balanced nutrient-rich soils. Moreover, C:N ratio is a major determinant in the composition of the soil microbial community (Wan et al., 2015). However the C:N ratio is in turn influenced by the biota, with microbial dominated soils having lower C:N ratio than fungal dominated, in a typical feedback loop.

The map of the C:N ratio shown in Fig. 7 evidences the higher values in northern areas as well in areas of more intense rainfall. Vegetation distribution clearly influences the spatial distribution of the C:N ratio with higher values under coniferous trees and peatlands.

Surprisingly, the C:N ratio is better predicted (in terms of model error) by the GPR model than the other properties. This is likely due to the ratio being determined mostly by the vegetation type and not by other variables such as soil age and the geological makeup.

### 4.2.4. *Topsoil phosphorus and potassium*

The map of soil P (Fig. 8) shows a clear trend where land use appears to have a strong influence. In particular, most of the agricultural areas have higher levels of P. This is quite evident in areas like the River Po plain (Italy) where levels of P diverge from the national average. In general, areas with natural land cover and those with a prevalence of permanent crops correspond to lower levels of P.

The geological background seems to have a quite small influence, whereas climate is much more relevant. This is likely due to higher fertilization rates in relatively wetter climates. The P map produced in this study confirms models of P fertilization load (Potter et al., 2010).

Soil potassium distribution (Fig. 9) is mostly driven by parent material chemistry and climate. In particular, lower than average K concentrations are typical of the sandy soils of northeastern Europe, and of the relatively young soils of Scandinavia. Moreover, Portugal and northwestern Spain also exhibit lower levels of potassium likely due to leaching. In general, soils with higher clay content are better able to retain K, so the two variables show a similar spatial distribution (Ballabio et al., 2016).

## 4.3. *Prediction uncertainty*

The GPR predictive variance is a measure of uncertainty in the model prediction. Knowing the prediction uncertainty can be important when making predictions for areas characterised by different covariates values, compared to input data. This is analogous to kriging variance. However, while kriging variance is based on geographic distance, GPR variance is a function of the kernel distance in the covariates feature space. So while the patterns of kriging variance can be guessed by the final user from the spatial distribution of the observations, GPR variance cannot be easily assessed as it only partially depends on the samples spatial coverage.

Fig. 10 shows the prediction variances for pH, in water and $CaCl_2$ solution, nitrogen and C:N ratio. While only these variables are shown, maps of variance where produced for every chemical property. As expected, some of the areas with the highest variance values are associated with unsampled areas (i.e. mountain areas above 1000 m a.s.l.). Forest areas also tend to exhibit a relatively higher variance as do areas where the presence of organic soils is more likely (i.e. Scotland and Ireland). A similar behaviour is visible in the pH maps; it is worth noting that pH measured in water has a generally higher variance than pH in $CaCl_2$; this reflects the less stable measure of pH in water.

### 4.4. Data availability

The soil physical properties maps (Ballabio et al., 2016) were made available through ESDAC in September 2015. According to statistics derived for a review of LUCAS data (Orgiazzi et al., 2018), the physical properties are among the most requested datasets with almost 850 logged requests during a three year period (2015–2018). The proposed chemical properties datasets will be available in ESDAC with the publication of this study. The availability of data is an important obstacle for modellers as high spatial resolution datasets are not generally freely accessible.

## 5. Conclusions

This study provides a new set of maps of baseline topsoil chemical properties at 250 m resolution for twenty-six countries of the EU, covering an area of more than 4.5 million km$^2$. The modelling is based on Gaussian Process Regression technique that allows the estimation of prediction uncertainty. The best performing prediction was obtained for the C:N ratio ($R^2 = 0.91$), followed by phosphorus and potassium ($R^2 = 0.75$). The performance prediction of the rest of chemical properties in terms of $R^2$ is higher than 0.60 with the exception of CEC ($R^2 = 0.35$).

Topsoil pH is influenced by soil parent material, erosional effects, climate and vegetation. The calcium carbonate content is highly correlated with pH, having similar influencing factors. Soil nitrogen distribution is dependent on soil organic carbon, vegetation and climate and soil texture. The land use appears to be the main driver for phosphorus content in soils, as agricultural areas have higher concentrations due to fertilizer application. CEC is influenced by the clay distribution in soils, topography and parent material.

The main limitations of the study are the number of points and the quality of some input covariates.

With 22,000 sampled locations the LUCAS soil database is unique in Europe for the number of available observations, its spatial coverage and its temporal resolution. Moreover, LUCAS soil will be improved by the additional samples taken in the 2015 and 2018 campaigns. However, at present the limitation of unsampled areas in mountains higher than 1000 m makes the prediction highly uncertain in those areas. This limitation was removed in the 2015 survey.

The resolution of the available geological covariates is very coarse and their influence in the prediction of chemical properties was limited. On the contrary, the vegetation covariates were the most significant, which means that better vegetation products such as the ones derived from the EU-ESA Copernicus program can further improve the derived chemical property maps in the future.

While LUCAS point data are available upon request from the European Soil Data Centre (ESDAC), the interpolated maps of chemical properties offer a better overview of the distribution of soil chemical properties in the EU to the scientific community and to policy makers.

The chemical properties datasets, together with the physical properties, contribute to one of the main objectives of the GlobalSoilMap project (Arrouays et al., 2017), which is to combine worldwide predictions of soil properties towards a first product of GlobalSoilMap.

## Acknowledgements

The research was performed in the FP7 programme RECARE (Grant No 603498). The LUCAS Survey is coordinated by the European Statistical Office (EUROSTAT). Laboratory analysis were funded by DG AGRI, CLIMA and ENV.

## Conflict of interest

The authors confirm that there is no conflict of interest with networks, organizations and data centres referred to in this paper.

## References

Adhikari, K., Hartemink, A.E., Minasny, B., Kheir, R.B., Greve, M.B., Greve, M.H., 2014. Digital mapping of soil organic carbon contents and stocks in Denmark. PLoS One 9, e105519.

Amini, M., Abbaspour, K.C., Khademi, H., Fathianpour, N., Afyuni, M., Schulin, R., 2005. Neural network models to predict cation exchange capacity in arid regions of Iran. Eur. J. Soil Sci. 56, 551–559.

Arrouays, D., Leenaars, J.G., Richer-de-Forges, A.C., Adhikari, K., Ballabio, C., Greve, M., Grundy, M., Guerrero, E., Hempel, J., Hengl, T., 2017. Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. GeoResJ 14, 1–19.

Ballabio, C., Comolli, R., 2010. Mapping heavy metal content in soils with multi-kernel SVR and LiDAR derived data. In: Digital Soil Mapping. Springer, pp. 205–216.

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

Ballabio, C., Panagos, P., Lugato, E., Huang, J.-H., Orgiazzi, A., Jones, A., Fernández-Ugalde, O., Borrelli, P., Montanarella, L., 2018. Copper distribution in European topsoils: an assessment based on LUCAS soil survey. Sci. Total Environ. 636, 282–298.

Bashfield, A., Keim, A., 2011. Continent-wide DEM Creation for the European Union. Presented at the 34th International Symposium on Remote Sensing of Environment. The GEOSS Era: Towards Operational Environmental Monitoring. Sydney, Australia. pp. 10–15.

Batjes, N.H., 1996. Total carbon and nitrogen in the soils of the world. Eur. J. Soil Sci. 47, 151–163.

Bishop, T., McBratney, A., 2001. A comparison of prediction methods for the creation of field-extent soil property maps. Geoderma 103, 149–160.

Böhner, J., Antonić, O., 2009. Land-surface parameters specific to topo-climatology. Dev. Soil Sci. 33, 195–226.

Bronick, C.J., Lal, R., 2005. Soil structure and management: a review. Geoderma 124, 3–22.

Büttner, G., 2014. CORINE land cover and land cover change products. In: Land Use and Land Cover Mapping in Europe. Springer, pp. 55–74.

Carre, F., Rusco, E., Tóth, G., Jones, A., Gardi, C., Stolbovoy, V., 2013. 2. Soil sampling methodology. LUCAS Topsoil Surv. Methodol. Data Results 3.

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

Didan, K., 2005. MYD13Q1 MODIS/aqua vegetation indices 16-day L3 global 250m SIN grid V006 [data set] [WWW document]. URL. https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/myd13q1_v006, Accessed date: 26 February 2019.

Drineas, P., Mahoney, M.W., 2005. On the Nyström method for approximating a Gram matrix for improved kernel-based learning. J. Mach. Learn. Res. 6, 2153–2175.

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

Gallant, J.C., Dowling, T.I., 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. Water Resour. Res. 39.

Grundy, M., Rossel, R.V., Searle, R., Wilson, P., Chen, C., Gregory, L., 2015. Soil and landscape grid of Australia. Soil Res. 53, 835–844.

Hartemink, A.E., 2015. On global soil science and regional solutions. Geoderma Reg. 5, 1–3. https://doi.org/10.1016/j.geodrs.2015.02.001.

Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G., Walsh, M.G., 2014. SoilGrids1km—global soil information based on automated mapping. PLoS One 9, e105992.

Hillel, D., 2008. 11. Soil fertility and plant nutrition. In: Hillel, D. (Ed.), Soil in the Environment. Academic Press, San Diego, pp. 151–162. https://doi.org/10.1016/B978-0-12-348536-6.50016-2.

Hjerdt, K., McDonnell, J., Seibert, J., Rodhe, A., 2004. A new topographic index to quantify downslope controls on local drainage. Water Resour. Res. 40.

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. A Journal of the Royal Meteorological Society 25 (15), 1965–1978.

IPBES, 2019. Global Assessment Report on Biodiversity and Ecosystem Services of the Intergovernmental Science - Policy Platform on Biodiversity and Ecosystem Services. IPBES Secretariat, Bonn, Germany.

Khaledian, Y., Brevik, E.C., Pereira, P., Cerdà, A., Fattah, M.A., Tazikeh, H., 2017. Modeling soil cation exchange capacity in multiple countries. Catena 158, 194–200.

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., 1983. Optimization by simulated annealing. Science 220, 671–680.

Lal, R., 2007. Carbon management in agricultural soils. Mitig. Adapt. Strateg. Glob. Chang. 12, 303–322.

Mansuy, N., Thiffault, E., Paré, D., Bernier, P., Guindon, L., Villemaire, P., Poirier, V., Beaudoin, A., 2014. Digital mapping of soil properties in Canadian managed forests at 250 m of resolution using the k-nearest neighbor method. Geoderma 235, 59–73.

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

Menenti, M., Azzali, S., Verhoef, W., Van Swol, R., 1993. Mapping agroecological zones and time lag in vegetation growth by means of Fourier analysis of time series of NDVI images. Adv. Space Res. 13, 233–237.

Minasny, B., McBratney, A.B., 2016. Digital soil mapping: a brief history and some lessons. Geoderma 264, 301–311.

Minasny, B., Setiawan, B.I., Saptomo, S.K., McBratney, A.B., 2018. Open digital mapping as a cost-effective method for mapping peat thickness and assessing the carbon stock of tropical peatlands. Geoderma 313, 25–40.

Montanarella, L., 2015. Agricultural policy: govern our soils. Nat. News 528, 32.

Moody, A., Johnson, D.M., 2001. Land-surface phenologies from AVHRR using the discrete Fourier transform. Remote Sens. Environ. 75, 305–323.

Olsson, L., Eklundh, L., 1994. Fourier series for analysis of temporal sequences of satellite sensor imagery. Int. J. Remote Sens. 15, 3735–3741.

Orgiazzi, A., Ballabio, C., Panagos, P., Jones, A., Fernández-Ugalde, O., 2018. LUCAS Soil, the largest expandable soil dataset for Europe: a review. Eur. J. Soil Sci. 69, 140–153.

Padarian, J., Minasny, B., McBratney, A., 2017. Chile and the Chilean soil grid: a contribution to GlobalSoilMap. Geoderma Reg. 9, 17–28.

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.

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, 189–200.

Poggio, L., Gimona, A., 2014. National scale 3D modelling of soil organic carbon stocks with uncertainty propagation—an example from Scotland. Geoderma 232, 284–299.

Potter, P., Ramankutty, N., Bennett, E.M., Donner, S.D., 2010. Characterizing the spatial patterns of global fertilizer application and manure production. Earth Interact. 14, 1–22. https://doi.org/10.1175/2009EI288.1.

Prado, R., Benites, V., Machado, P., Polidoro, J., Dart, R., Naumov, A., 2008. Mapping potassium availability from limited soil profile data in Brazil. In: Digital Soil Mapping With Limited Data. Springer, pp. 91–101.

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian Processes for Machine Learning, Adaptive Computation and Machine Learning. MIT Press, Cambridge, Mass.

Rossel, R.A.V., Bui, E.N., 2016. A new detailed map of total phosphorus stocks in Australian soil. Sci. Total Environ. 542, 1040–1049.

Saidi, D., 2012. Importance and role of cation exchange capacity on the physicals properties of the Cheliff Saline Soils (Algeria). Procedia Eng. 33, 435–449.

Salminen, R., Batista, M., Bidovec, M., Demetriades, A., De Vivo, B., De Vos, W., Duris, M., Gilucis, A., Gregorauskiene, V., Halamić, J., 2005. Geochemical Atlas of Europe, Part 1, Background Information, Methodology and Maps. Geological Survey of Finland.

Sarmadian, F., Keshavarzi, A., Malekian, A., 2010. Continuous mapping of topsoil calcium carbonate using geostatistical techniques in a semi-arid region. Aust. J. Crop. Sci. 4, 603.

Schofield, R., Taylor, A.W., 1955. The measurement of soil pH 1. Soil Sci. Soc. Am. J. 19, 164–167.

Slessarev, E., Lin, Y., Bingham, N., Johnson, J., Dai, Y., Schimel, J., Chadwick, O., 2016. Water balance creates a threshold in soil pH at the global scale. Nature 540, 567.

Stein, M.L., 2012. Interpolation of Spatial Data: Some Theory for Kriging. Springer Science & Business Media.

Tóth, G., Jones, A., Montanarella, L., 2013a. LUCAS Topsoil Survey: Methodology, Data and Results. Publications Office.

Tóth, G., Jones, A., Montanarella, L., 2013b. The LUCAS topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union. Environ. Monit. Assess. 185, 7409–7425.

Tóth, G., Guicharnaud, R.-A., Tóth, B., Hermann, T., 2014. Phosphorus levels in croplands of the European Union with implications for P fertilizer use. Eur. J. Agron. 55, 42–52. https://doi.org/10.1016/j.eja.2013.12.008.

Wan, X., Huang, Z., He, Z., Yu, Z., Wang, M., Davis, M.R., Yang, Y., 2015. Soil C:N ratio is the major determinant of soil microbial community structure in subtropical coniferous and broadleaf forest plantations. Plant Soil 387, 103–116.

Wang, K., Zhang, C., Li, W., 2013. Predictive mapping of soil total nitrogen at a regional scale: a comparison between geographically weighted regression and cokriging. Appl. Geogr. 42, 73–85.

Wang, M., Zheng, Q., Shen, Q., Guo, S., 2013. The critical role of potassium in plant stress response. Int. J. Mol. Sci. 14, 7370–7390. https://doi.org/10.3390/ijms14047370.

Wilford, J., De Caritat, P., Bui, E., 2015. Modelling the abundance of soil calcium carbonate across Australia using geochemical survey data and environmental predictors. Geoderma 259, 81–92.

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

Yost, R., Uehara, G., Fox, R., 1982. Geostatistical analysis of soil chemical properties of large land areas. I. Semi-variograms 1. Soil Sci. Soc. Am. J. 46, 1028–1032.