Geoderma Regional 7 (2016) 201–215

Contents lists available at ScienceDirect

Geoderma Regional

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

Geostatistical prediction and simulation of European soil property maps

Gerard B.M. Heuvelink [a,*], Johannes Kros [b], Gert Jan Reinds [b], Wim De Vries [b]

[a] Soil Geography and Landscape group, Wageningen University, PO Box 47, 6700 AA Wageningen, The Netherlands
[b] Alterra, Wageningen University and Research Centre, PO Box 47, 6700 AA Wageningen, The Netherlands

ARTICLE INFO

Article history:
Received 16 November 2015
Received in revised form 8 March 2016
Accepted 11 April 2016
Available online 13 April 2016

Keywords:
Geostatistics
Cokriging
Europe
Soil properties
Uncertainty
Mapping
Cambisols
Fluvisols

ABSTRACT

A geostatistical model was developed and applied to predict six soil properties and soil horizon thickness for mineral A, B and C soil horizons at the European scale and quantify the associated prediction uncertainties. The soil properties are pH, organic carbon content, organic nitrogen content, clay and sand contents and bulk density. The geostatistical model takes a regression cokriging approach, in which correlations between soil properties and across soil horizons are taken into account. Non-stationarities in the means and variances are represented by mapping units of the generalised European soil and land cover maps. The model was calibrated using the combined WISE, SPADE 1 and EFSDB databases, which jointly contain approximately 3600 soil profiles, irregularly distributed over Europe. The resulting model showed for most soil properties strong dependencies on soil type and land cover, moderate correlations between soil property residuals, strong correlations across horizons, and moderate spatial correlation of regression residuals. Kriging predictions and simulations were made on a 5 km by 5 km grid. Uncertainties in the resulting maps are large, particularly in under-sampled parts of Europe and in strata with large spatial variation. We conclude that geostatistical prediction and simulation are useful techniques to quantify uncertainties in soil property maps at the European scale, but that many more observations are required to fully exploit the relationship with explanatory variables and improve mapping accuracy. One important advantage of the techniques used is that they yield a full probabilistic model, as required by Monte Carlo uncertainty propagation analyses of spatially distributed dynamic models that use soil properties as uncertain input. In particular, the results of this study have been used to analyse how uncertainty in soil properties propagate through terrestrial greenhouse gas emission models.

1. Introduction

The soil is an important controlling factor of many environmental processes, such as forest- and crop growth, nitrate leaching and greenhouse gas emission. As a result, dynamic models that aim to mimic these processes require soil information as input. End users of environmental process models, however, are increasingly dissatisfied with the predicted model output only, but also require quantification and communication of the associated uncertainties (e.g. Brown et al., 2005; Refsgaard et al., 2007, Van Der Sluijs, 2012; Bastin et al., 2013). Thus, it is necessary to analyse how errors and imperfections in model inputs propagate through the model. This can be done with Monte Carlo simulation, which repeatedly runs the model with random draws from the probability distribution of the uncertain inputs and computes summary statistics of the sample of model outputs (Hammersley and Handscomb, 1979; Heuvelink, 1998). Monte Carlo uncertainty propagation analysis is nowadays widely used in soil science studies (e.g. Kros et al., 1999; Balakrishnan et al., 2005, Lehrter and Cebrian, 2010; Nol et al., 2010; Freni et al., 2011; Van den Berg et al., 2012; Fitton et al., 2014).

Monte Carlo uncertainty propagation analysis of dynamic soil process models requires sampling from joint probability distributions of uncertain soil inputs. This is the most difficult part of the analysis, because it is rare that full probability distributions are available. It is not enough to have just a single measure of the uncertainty, such as the root mean squared error or variance. Instead, a full probability model is required that includes spatial correlations (i.e., variograms) of the uncertainties and cross-correlations between uncertainties in different soil properties and between the same soil property at different depths (Heuvelink et al., 2007; Heuvelink, 2014). Cross-correlations are important because these can dramatically influence the uncertainty in model output (Heuvelink, 1998). Spatial correlation is important when models include spatial interactions or when spatial aggregates of model outputs are presented (e.g. the average greenhouse gas emission over regions or countries). Arguably the only viable way to arrive at a full probabilistic description of spatially distributed soil properties is through a geostatistical approach (Heuvelink, 2014). This first requires the definition and calibration of a statistical model that describes the spatial structure of the soil properties, their mutual relationships and dependency on deterministic explanatory variables ('covariates') such as topography, geology, climate and land cover. Next the statistical model is used to condition the soil properties to point observations

* Corresponding author.
E-mail address: gerard.heuvelink@wur.nl (G.B.M. Heuvelink).

http://dx.doi.org/10.1016/j.geodrs.2016.04.002
2352-0094/

and explanatory variables through (regression) kriging (Hengl et al., 2004). For Monte Carlo uncertainty propagation, kriging is replaced with spatial stochastic simulation (Goovaerts, 1999). While kriging makes optimal predictions, spatial stochastic simulation uses a pseudo-random number generator to sample from the conditional probability distribution.

The aims of this paper are: i) to define a multivariate geostatistical model for seven soil properties (pH, organic carbon (C) content (g/kg), organic nitrogen (N) content (g/kg), clay content (%), sand content (%), bulk density (g/cm^3) and horizon thickness (cm); for simplicity we deem horizon thickness also a 'soil property') for three major soil horizons (A, B and C) for Europe (here used as a synonym for EU25 + 5); ii) to condition the model to publicly available point observations and explanatory variables and derive maps of predictions and prediction error standard deviations using regression kriging; and iii) to generate conditional simulations of the soil properties for use in a Monte Carlo uncertainty propagation analysis. Such analysis can be done with the INTEGRATOR model (De Vries et al., 2010, 2011; Kros et al., 2010), which has previously been used to predict uncertainties in the emission of nitrogen to air (e.g. ammonia, nitrous oxide) and water (e.g. nitrate, ammonium) from both agricultural and natural terrestrial systems (Kros et al., 2012).

2. Materials and methods

2.1. Soil profile data and explanatory variables

The member states of the EU25 + 5 all have their own national and regional databases with soil profile information and associated field and laboratory observations, which in total amounts to millions of soil records. For instance, only for The Netherlands there are over 330,000 unique soil profiles with field characterizations, about 8000 of which have laboratory observations of basic soil properties. However, most of these databases are not publicly available or easily downloadable, and neither are these harmonised to a common standard. Part of the data is harmonised and centrally stored in the European Soil Database 1:1,000,000 (ESDBv2; http://eusoils.jrc.ec.europa.eu/ESDB_Archive/ESDB) and of these the SPADE 1 dataset (Hiederer et al., 2006) has geo-referenced observations and is free to use. The size of the dataset is limited, however, to only 382 profiles of measured soil properties. In addition, the global ISRIC-WISE database (Batjes, 2009) offers another 1012 profiles in Europe. Third, there is a European Forest Soil Database (EFSDB, Reinds, 1994), which contains 2192 profile observations, but these are only located in forest soils. Recently, the LUCAS soil database (Toth et al., 2013) was also made available under a specific license. However, this dataset only contains information about the topsoil and was therefore not used in this research. Fig. 1 gives the profile locations of the three datasets used. Most profiles had three mineral soil horizons (A, B and C) but those that had more were simplified by omitting the E horizon. In total, approximately 3500 profiles were assembled, but not all profiles had observations of all seven soil properties for all three horizons. Table 1 gives the total number of available observations for each combination of soil property and horizon.

Given that the total number of soil profiles is limited it is worthwhile to include additional information about the spatial variation of soil properties as contained in spatially exhaustive data layers, such as the 1:1,000,000 soil map of the European Communities (CEC, 1985), further denoted as EC soil map, the CORINE EU land cover map (EEA, 2009) and the environmental zone map of Europe (Metzger et al., 2005; Jongman et al., 2006). These maps are freely available in digital form and provide relevant and partially complementary information about the soil properties and can thus improve mapping accuracy. Since all three maps have a fairly extensive legend, generalisation is required to be able to use these as explanatory variables in a geostatistical model. The EC soil map was generalised by reclassifying it into four categorical maps that focus on main differences in base status (four classes), organic matter (three classes), soil wetness (three classes) and soil texture (six classes). Table 2 provides the correlation between the resulting soil groups and those of the EC soil map. These four soil characteristics were chosen such that they could easily be derived from soil type and texture class alone. The number of classes per map was kept deliberately small to support parameter estimation in the geostatistical modelling phase. The land cover map was generalised into three main classes (grassland, arable land and nature, while omitting other land use types).

Maps of the explanatory variables are given in Fig. 2.

2.2. Geostatistical modelling

Let Z_i ($i = 1,..,21$) refer to one of seven soil properties at one of three horizons. Each Z_i is defined as:

$$Z_i(s) = \sum_{k=1}^{p_i} (\mu_{ik} + \sigma_{ik} \cdot \varepsilon_i(s)) \cdot f_{ik}(s) \quad (1)$$

where s is geographic location and f_{ik} are binary maps that at any location are either 0 or 1, and where for any location we have $\sum_{k=1}^{p_i} f_{ik}(s) = 1$. These binary maps were derived from the maps given in Fig. 2, by first overlaying all maps and next merging the newly formed classes, such that for the i-th soil property p_i classes remain. Merging is required to reduce the total number of classes and keep sufficient observations within each class to be able to estimate μ_{ik} and σ_{ik}, as explained below. Parameter μ_{ik} represents the mean of soil property Z_i in the k-th class while σ_{ik} represents its standard deviation. The stochastic residual ε_i is assumed isotropic, second-order stationary and normally distributed with zero mean and unit variance. Its spatial correlation function is denoted by $\rho_i(h)$, where h is Euclidean spatial distance. Also, the stochastic residuals of two soil properties Z_i and Z_j (i.e., two different soil properties or the same soil property at different horizons) may be spatially cross-correlated, as characterised by the spatial cross-correlation function $\rho_{ij}(h)$. For soil properties that have skew distributions, the geostatistical model presented in Eq. (1) is formulated for the (natural) log-transformed soil property.

Calibration of the geostatistical model requires estimation of the parameters μ_{ik} and σ_{ik}, and correlation functions $\rho_i(h)$ and $\rho_{ij}(h)$. The first parameters are estimated simply by taking the arithmetic mean and standard deviation of all observations within each class. This implies that each class must have sufficient observations to yield reliable estimates. Overlay of the six maps of Fig. 2 potentially yields $4 \times 3 \times 3 \times 6 \times 3 \times 13 = 8424$ classes, although in reality only 439 occur. However, these were still too many, since for most soil properties only about 2000 to 3000 observations are available per horizon, while for bulk density the total number of observations per horizon is only about 560 to 820 (see Table 1). Overlay classes that were judged less distinctive were therefore merged, by using expert judgement and comparison of histograms of observations between classes. Table 3 shows that the merge may be different for each soil property, since the influence of explanatory variables differs between soil properties. It should be noted that the forced reduction to a manageable number of classes indicates that much of the information contained in the explanatory variables may be lost. For instance, the environmental zone map remains unused for all soil properties, not because it is non-informative, but because other explanatory variables are judged more important and too few per-class observations would remain if it were included.

The resulting class maps that are used to define the f_{ik} of Eq. (1) are given in Fig. 3. The number of classes varies between six (for clay, sand and bulk density) to thirteen (for organic N content) and are independent of the soil horizon for a given soil property. The distinguished classes generally include land cover, while texture class is included as a proxy for clay and sand content, and organic matter class as a proxy for organic C and organic N content. As an example, Table 4 shows the distinguished nine classes for pH derived from land cover and base