Geoderma 277 (2016) 69–82

Contents lists available at ScienceDirect

Geoderma

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

Bayesian spatial modelling of soil properties and their uncertainty: The example of soil organic matter in Scotland using R-INLA

Laura Poggio [a,*], Alessandro Gimona [a], Luigi Spezia [b], Mark J. Brewer [b]

[a] The James Hutton Institute, Craigiebuckler, AB15 8QH, Aberdeen, Scotland, (UK)
[b] Biomathematics and Statistics Scotland, Craigiebuckler, AB15 8QH, Aberdeen, Scotland, (UK)

ARTICLE INFO

Article history:
Received 22 October 2015
Received in revised form 26 April 2016
Accepted 29 April 2016
Available online 26 May 2016

Keywords:
Bayesian inference
remote sensing
soil properties mapping
uncertainty propagation

ABSTRACT

As any model for digital soil mapping suffers from different types of errors, including interpolation errors, it is important to quantify the uncertainty associated with the maps produced. The most common approach is some form of regression kriging or variation involving geostatistical simulation. Another way of assessing the spatial uncertainty lies in the Bayesian approach where the uncertainty is described by the posterior density. Typically Markov Chain Monte Carlo is used to compute the posterior density; however, this process is computationally intensive. The aim of this paper is to present an example of Bayesian uncertainty evaluation using (Bayesian) latent Gaussian models fitted using INLA (Integrated Nested Laplace Approximation) and with the SPDE (Stochastic Partial Differential Equation) approach for modelling the spatial correlation. For illustration, soil organic matter content in the Grampian region of Scotland (UK, about 12,100 km^2) was modelled for topsoil (2D) and whole-profile data (3D). Results were assessed using in-sample and out-of-sample measures and compared for distribution similarity, variogram and spatial structure reproduction, computational load and uncertainty ranges. The results were also compared with outputs from an extension of scorpan-kriging. The Bayesian framework using INLA offers a viable alternative to existing methods for digital soil mapping, with comparable validation results, important computational gains, good assessment of uncertainty and potential for integrated modelling uncertainty propagation.

1. Introduction

Soils play an important role in the environment and in ecosystem functions. Soil are highly variable in space and uncertainty is inherently associated with soil modelling and mapping at different scales. Numerous studies in recent years have modelled various soil properties with the support of environmental covariates. Most of these studies focused on mapping the horizontal variability with separate models for each of the soil layers considered, most often the topsoil. More recently, approaches were described and applied to take into account the vertical variability and the relationships across soil horizons (Malone et al., 2009).

Any model for digital soil mapping suffers from different types of errors, including interpolation errors, and it is therefore important to quantify the uncertainty associated with the maps produced. Various studies have emphasised the importance of modelling and assessing the uncertainty of the results of each mapping exercise (e.g. Minasny et al., 2013; Martin et al., 2011). The uncertainty can have important effects on further modelling especially when combined with the uncertainty inherited from other processes, such as a climate model. The lack of uncertainty assessment does not provide information on the reliability of predictions, limiting the potential for decision making (McBratney, 1992; Ogle et al., 2010).

Most of the methods used to date in the assessment of spatial uncertainty for digital soil mapping are within the frequentist framework or the conventional geostatistical approach as defined by Diggle et al. (1998). The most common method is some form of regression kriging or study of variation involving geostatistical simulation within the frequentist framework with uncertainty calculated from a (large) number of realisations (e.g. McBratney et al., 2003; Hengl et al., 2004; Grimm et al., 2008; Poggio and Gimona, 2014). Frequentist methods cannot easily reveal and model the uncertainty of all model parameters (Banerjee et al., 2014), but some studies have focussed on the estimation of uncertainty within a frequentist geostatistical approach (see e.g. Marchant and Lark, 2004, 2007; Zhu and Stein, 2005). An alternative to the frequentist approach is Bayesian modelling, where uncertainty is described explicitly by the posterior density (Banerjee et al., 2014; Cameletti and Blangiardo, 2015). The Bayesian approach produces credibility intervals which explicitly indicate the probability that the parameters lie within a specific range. In contrast, assuming repeatability of the experiment under the same spatial and temporal conditions, confidence intervals from the frequentist approach indicate the percentage of occasions that the interval contains the "true", unknown parameter

* Corresponding author.
E-mail address: laura.poggio@hutton.ac.uk (L. Poggio).

http://dx.doi.org/10.1016/j.geoderma.2016.04.026
0016-7061/

value (Banerjee et al., 2014; Cameletti and Blangiardo, 2015). Bayesian approaches have tended to rely on Markov Chain Monte Carlo (MCMC) simulations to make inference. MCMC algorithms have been widely applied to a number of environmental applications (e.g. Kazianka and Pilz, 2012; Hararuk and Luo, 2014; Patil et al., 2011; Zhang, 2002; Verdin et al., 2015). However, fewer applications exist in the domain of soil science, apart from work on Bayesian kriging (e.g. Aelion et al., 2009) and Bayesian hierarchical modelling (e.g. Minasny et al., 2011; Xiong et al., 2015; Sun et al., 2013). The Bayesian approach has been applied to hydrological properties (Marthews et al., 2014), such as soil moisture (Gao et al., 2014) or water retention curves (Yang et al., 2015).

MCMC is flexible and able to deal with virtually any type of data and models (e.g. Gaussian univariate Bayesian spatial regression models), but involves computationally- and time-intensive simulations (Banerjee et al., 2008); this can be a limiting factor in Bayesian spatial applications with respect to the size of data sets which can be analysed, although the availability of parallel computing can facilitate computation for larger data sets (Minasny et al., 2011; Schmidberger et al., 2009). Recently a computationally efficient alternative to MCMC was developed for so-called latent Gaussian models—the Integrated Nested Laplace Approximation approach (INLA, Rue et al., 2009). The set of latent Gaussian models includes many forms of regression, including (generalised) linear mixed spatial models, but is relatively limited in terms of general hierarchical modelling. A recent review (Falk et al., 2015) provides a comparison of algorithms and computational implementations applied to remote sensing images. INLA can be combined with the Stochastic Partial Differential Equation (SPDE) approach (Lindgren et al., 2011; Lindgren and Rue, 2015) for efficient modelling of spatial point data and geostatistical applications (Bivand et al., 2015). INLA with SPDE (henceforth "INLA + SPDE") has been applied to a variety of environmental problems, such as those in: Illian et al. (2013); Carson and Flemming (2014); Serra et al. (2014). Recent studies used this approach to model changes in fire regimes in a large area of the Amazon forest (Gutierrez-Velez et al., 2014). An initial study applying INLA + SPDE to soil properties was presented in Poggio et al. (2014, 2016).

The main aim of this paper is to present a Bayesian framework using latent Gaussian models fitted with INLA and using the SPDE method for modelling the spatial correlation (INLA + SPDE) including both lateral (2D) and vertical (3D) variability of soil properties to introduce it to the Digital Soil Mapping (DSM) toolkit. To illustrate the approach, we mapped the soil organic matter content in Scottish soils at a regional scale. The process and results of INLA + SPDE were compared with results from an extension of the scorpan-kriging approach, i.e. hybrid geostatistical Generalized Additive Models (GAM, Wood, 2006), combining GAM with Gaussian simulations (henceforth GAM + GS, Poggio and Gimona, 2014). The comparison will show that the proposed method has comparable results with respect to validation, is more computationally efficient and is more flexible in dealing with spatial uncertainty and its propagation.

2. Data and test areas

2.1. Test areas

In Scotland there is a clear distinction between organic and mineral soils, which often results in a bimodal distribution of soil properties, especially organic matter content (Chapman et al., 2009; Poggio et al., 2013).

Within Scotland, a test area was chosen to represent a wide range of soils, both mineral and organic. The Grampian region of Scotland (about 12,100 km^2) covers the whole of NE Scotland (Fig. 1) with a variety of landscapes and soils. It includes large river catchments and the Cairngorm mountains, with some of the highest peaks in Scotland. In this region there is a good spread of both mineral and organic soils, with different patch sizes and fragmentation soil types.

2.2. Response variable

The Scottish Soils Database contains information and data on soils from locations throughout Scotland. It contains the National Soil Inventory of Scotland (NSIS) profile samples collected on a regular 10 km grid of sampled locations (Lilly et al., 2010) and physical and chemical data from a large number of soil profiles taken to characterise the soil mapping units.

Fig. 2 shows the distribution for this response variable in the different test cases and the number of available soil profiles in each. For 2D applications, only topsoil values were used, while in 3D applications values for horizons to a depth of one metre were used. In order to provide validation of the models the data available were split into two sets, training and validation, in a ratio of 3:1, where the validation set was sampled randomly. In this study, the soil property considered was the percentage of soil organic matter (SOM). As the original data were percentages, they were mapped onto the real line by the logit function. This transformation also provided (approximately) Normally distributed observations.

2.3. Covariates

The covariates included are freely and globally available and were selected to describe, directly or indirectly, the most important scorpan factors, namely topography, vegetation, climate and geographical position.

2.3.1. Morphology

The Digital Elevation Model (DEM) used as a covariate in the fitted models was SRTM (Shuttle Radar Topography Mission), further processed to fill in no-data voids (Jarvis et al., 2006; Rodriguez et al., 2006). SRTM has a spatial resolution of 90m with global coverage. The measures used were elevation and slope as the steepest slope angle, calculated using the D8 method (O'Callaghan and Mark, 1984).

In order to match the resolution of the other covariates the medians in each grid cell of 1×1 km were used.

2.3.2. Remote sensing

A set of indices was derived from the Terra Moderate Resolution Imaging Spectro-radiometer (MODIS) 8 and 16 day composite products. The data were acquired from the NASA ftp website (ftp://e4ftl01u.ecs.nasa.gov/MOLT/) for 12 years between 2000 and 2011. The individual images were restored to fill cloud gaps (Poggio et al., 2012). The indices were selected for their capability to differentiate spectral responses from different bare soils, vegetation cover and mixed situations:

1) Enhanced Vegetation Index (EVI; Huete et al., 2002),
2) the Normalised Difference Water Index (NDWI; Gao, 1996)

$$NDWI = \frac{NIR - SWIR}{NIR + SWIR} \quad (1)$$

NDWI was calculated with NIR (Near InfraRed) and SWIR (Short Wave InfraRed) band: SWIR = 2130 (Gu et al., 2008).

Medians over the 12 years were used as covariates.

2.4. Soil information used for prior definition

In order to provide prior information on where peat, highly organic and mineral soils are likely to be present, an additional data set was used. This information was derived from two sources: 1. reclassification of the available traditional soil map, and 2. environmental clustering of the data available.