Geoderma 306 (2017) 195–205

Contents lists available at ScienceDirect

Geoderma

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

Predicting soil properties in the Canadian boreal forest with limited data: Comparison of spatial and non-spatial statistical approaches

Julien Beguin[a,*], Geir-Arne Fuglstad[b], Nicolas Mansuy[a], David Paré[a]

[a] Natural Resources Canada, Canadian Forest Service, Laurentian Forestry Centre, Quebec, QC G1V 4C7, Canada
[b] Department of Mathematical Sciences, Norwegian University of Science and Technology, 7491 Trondheim, Norway

ARTICLE INFO

Editor: A.B. McBratney

Keywords:
Digital soil mapping
Boreal forest
Spatial autocorrelation
Bayesian analyses
Machine-learning
Random forests
Boosted regression trees
Kriging
Geostatistic
Cross-validation

ABSTRACT

Digital soil mapping (DSM) involves the use of georeferenced information and statistical models to map predictions and uncertainties related to soil properties. Many remote regions of the globe, such as boreal forest ecosystems, are characterized by low sampling efforts and limited availability of field soil data. Although DSM is an expanding topic in soil science, little guidance currently exists to select the appropriate combination of statistical methods and model formulation in the context of limited data availability. Using the Canadian managed forest as a case study, the main objective of this study was to investigate to which extent the choice of statistical method and model specification could improve the spatial prediction of soil properties with limited data. More specifically, we compared the cross-product performance of eight statistical approaches (linear, additive and geostatistical models, and four machine-learning techniques) and three model formulations ("*covariates only*": a suite of environmental covariates only; "*spatial only*": a function of geographic coordinates only; and "*covariates + spatial*": a combination of both covariates and spatial functions) to predict five key forest soil properties in the organic layer (thickness and C:N ratio) and in the top 15 cm of the mineral horizon (carbon concentration, percentage of sand, and bulk density). Our results show that 1) although strong differences in predictive performance occurred across all statistical approaches and model formulations, spatially explicit models consistently had higher R^2 and lower RMSE values than non-spatial models for all soil properties, except for the C:N ratio; 2) Bayesian geostatistical models were among the best methods, followed by ordinary kriging and machine-learning methods; and 3) comparative analyses made it possible to identify the more performant models and statistical methods to predict specific soil properties. We make modeling tools and code available (e.g., Bayesian geostastical models) that increase DSM capabilities and support existing efforts toward the production of improved digital soil products with limited data.

1. Introduction

Spatially explicit soil information is required to assess potential land use, predict vulnerabilities and implement biogeochemical models forecasting the impact of human activity and climate change on terrestrial ecosystems, as well as on the services they provide (Adhikari and Hartemink, 2016; Folberth et al., 2016). Considerable efforts have been made by the research community to harmonize and define common specifications of soil data sets from different origins (Arrouays et al., 2014). These efforts have led to the creation of large soil pedon databases that facilitate the mapping, monitoring and modeling of ecosystem processes at multiple spatial scales, making it possible to predict vegetation shifts (Kuhn et al., 2016) and changes in ecosystem productivity (Maire et al., 2015). Key outcomes of these advances culminated in the release of soil raster products at continental (Hengl et al., 2015) and global (Hengl et al., 2014) scales, together with quantitative estimates of uncertainty associated with predicted soil properties. The availability of soil quantitative estimates is a significant step toward integrating soil indicators into the assessment of ecosystem function and vulnerability (Folberth et al., 2016).

Digital soil mapping (DSM) involves the use of numerical methods to fit and validate statistical models on georeferenced soil information (dependent variables) using environmental covariates (independent variables) that represent soil-forming factors, and to map predictions and their uncertainty at a specified spatial resolution over a focal study area. Environmental covariates are obtained from various sources, including remote sensing products and digital elevation models (McBratney et al., 2003). When detailed expert-based soil maps are available, techniques of spatial disaggregation of polygon information are often used (Bui and Moran, 2001; Lamboni et al., 2016). However,

* Corresponding author.
E-mail address: julien.beguin@canada.ca (J. Beguin).

http://dx.doi.org/10.1016/j.geoderma.2017.06.016
Received 18 November 2016; Received in revised form 12 May 2017; Accepted 16 June 2017

over large regions, and more typically in forested regions, expert-based soil maps are often unavailable and the *scorpan* model approach (see McBratney et al., 2003), which matches environmental covariates with soil point data (pedons), is commonly used. A key challenge in DSM is that there is almost always a shortage of soil pedon data, which may lead to low model accuracy and/or misrepresentations of predicted soil attributes (Ahrens, 2008). How to make the maximum use of sparse data is thus a recurrent challenge in soil science. At the same time, this challenge offers a growing opportunity to develop new statistical approaches that improve soil predictive mapping.

Canada's forests, which cover over 390 million ha of land and represent 10% of the world's forest cover, are representative of this situation since only limited soil pedon database are available at the national level. Over the last decades, the pool of available numerical methods and statistical models combined with an increase in computing power and data availability have tremendously boosted DSM capabilities with limited data (McBratney et al., 2003; Grunwald, 2009; Brevik et al., 2016). Various new modeling tools are now freely available to predict and map soil types as well as continuous or discrete soil properties. The quality of these predictive soil maps, however, remains highly variable and depends of the interplay among four main key components: 1) the availability and quality of the data for both soil profiles and environmental covariates; 2) the inherent variation in nature complexity and heterogeneity of any focal soil property across spatial scales and soil depth; 3) the specification of statistical models (e.g., the choice of covariates, with linear vs non-linear effects, with simple vs interaction effect terms, with hierarchical structures or not, with the inclusion or not of a spatial component); and 4) the choice of statistical framework (e.g., Bayesian vs frequentist), statistical method and algorithm to fit these models (see Fig. 1), hereafter referred as 'statistical methods'.

Machine-learning techniques, in particular, have become very popular in predictive modeling (Hastie et al., 2009; Kuhn and Johnson, 2013), especially in DSM where numerous studies use random forests (Grimm et al., 2008), boosted regression trees (Grinand et al., 2008), *k*-nearest neighbors (Mansuy et al., 2014), Cubist (Rizzo et al., 2016), support vector machines (Were et al., 2015), and artificial neural networks (Behrens et al., 2005). In addition to the choice of statistical method, the choice of statistical models (model specification) includes the use of non-spatial vs spatially explicit models. When the geographical locations of sample plots are recorded, spatially explicit models are often used to account for spatial autocorrelation in the data or in model residuals (McBratney et al., 2003; Dormann et al., 2007; Hengl, 2009; Beale et al., 2010; Banerjee et al., 2014), which often improves the accuracy of predictions as well as the predictive performance of the models (Beguin et al., 2012).

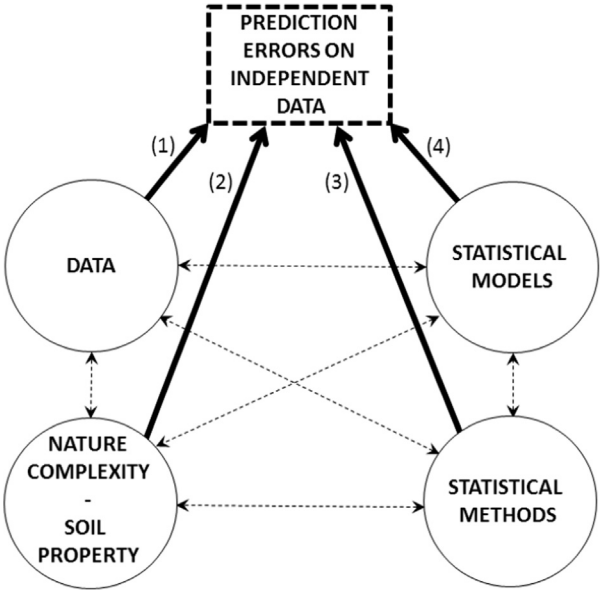

Fig. 1. Conceptual framework showing the interrelationships among the four main components that influence predictive errors in digital soil mapping when using statistical approaches. Getting the lowest prediction errors between observed and predicted soil properties on independent data is the main objective of digital soil mapping. Conceptually, the causes of prediction errors can be divided into four main components: (1) the quality and availability of the data (e.g., sample size, quality, spatial resolution and precision); (2) nature complexity or the level of heterogeneity in soil properties; (3) the choice of statistical framework (e.g., Bayesian vs frequentist), statistical method and algorithm, hereafter referred as 'statistical methods'; and (4) the choice of statistical model (e.g., spatial vs non-spatial, linear vs non-linear effects, simple vs interaction effect terms). Each of these components can act alone (bold arrows) or interact with other components (dashed arrows) to shape the accuracy of digital soil maps.

Although every spatial statistical method has its intrinsic way of modeling spatial correlation structure in the data (Li and Heap, 2014), the following are the most common in practice: 1) using additional covariates that are parametric or non-parametric functions of the sample geographic coordinates, such as in trend surface analyses with linear or additive models (Dormann et al., 2007), in spatial filtering regression (e.g., Moran Eigenvectors; Dray et al., 2006) or in autocovariate regression (Dormann et al., 2007); 2) using spatial covariance structure in the variance-covariance matrix with parametric function (e.g., variograms), such as in generalized least squares (GLS) models (Dormann et al., 2007) or in regression-kriging (Hengl et al., 2004); 3) using weighted matrices of interactions among neighboring sites, such as in conditional (CAR) and simultaneous (SAR) autoregressive models (Banerjee et al., 2014); and 4) using Bayesian hierarchical models where effects of the covariates, spatial effects and nugget effects are combined in an additive model (Banerjee et al., 2014). Bayesian methods may be computationally heavy, but there has been much recent development that makes them readily usable for data sets of realistic sizes of an order of 10,000 points and bigger (Sun et al., 2012; Lindgren and Rue, 2015).

While great efforts by the soil mapping community have led to standardized technical specifications regarding the spatial entity, the assessment of soil properties to be predicted, and the handling of uncertainties in DSM (Arrouays et al., 2014), less work has been done to compare the relative performance of a range of statistical methods and model specifications (e.g., spatial vs non-spatial) across multiple soil properties. Most DSM studies (but see Heung et al., 2016) use one or a few statistical approach(es) (Poggio et al., 2013; Nawar et al., 2015), typically with one type of model specification to analyze specific soil data sets, which often makes unclear the extent to which the combination of particular statistical approach and model formulation influences the outcome.

The main objective of this study was therefore to investigate to what extent the choice of statistical methods and model specification could improve the spatial prediction of forest soil properties with sparse soil data. More specifically, we compared the cross-product performance of eight statistical methods (linear, additive and geostatistical models, and four machine-learning techniques) and three different model specifications ("*covariates only*": model fitted with a suite of environmental covariates only; "*spatial only*": model fitted with only a spatial component derived from geographic coordinates of the plots; and "*covariates + spatial*": model fitted with both covariates and a spatial component) to predict five key forest soil properties (thickness and C:N ratio in the organic layer as well as carbon concentration, percentage of sand, and bulk density in the top 15 cm of the mineral horizon).

2. Material and methods

2.1. Study area

The study area covers 290 million ha of managed forests across Canada and extends from 52° to 138° West and from 42° to 60° North