Geoderma 232–234 (2014) 69–80

Contents lists available at ScienceDirect

Geoderma

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

Scope to predict soil properties at within-field scale from small samples using proximally sensed γ-ray spectrometer and EM induction data

J. Huang [a], R.M. Lark [b], D.A. Robinson [c], I. Lebron [c], A.M. Keith [d], B. Rawlins [b], A. Tye [b], O. Kuras [b], M. Raines [b], J. Triantafilis [a,*]

[a] School of Biological, Earth and Environmental Science, The University of New South Wales, Kensington, NSW 2052, Australia
[b] British Geological Survey, Keyworth, Nottingham NG12 5GG, UK
[c] NERC, Centre for Ecology and Hydrology, Environment Centre Wales, Deiniol Road, Bangor, Gwynedd LL57 2UW, UK
[d] NERC, Centre for Ecology and Hydrology, Bailrigg, Lancaster LA1 4AP, UK

ARTICLE INFO

Article history:
Received 9 August 2013
Received in revised form 15 January 2014
Accepted 28 April 2014
Available online 23 May 2014

Keywords:
EC_a
Induction
EM
Fuzzy k-means
Gamma-ray spectrometry
Soil variability
Characterization and distribution
Digital soil mapping

ABSTRACT

Spatial predictions of soil properties are needed for various purposes. However, the costs associated with soil sampling and laboratory analysis are substantial. One way to improve efficiencies is to combine measurement of soil properties with collection of cheaper-to-measure ancillary data. There are two possible approaches. The first is the formation of classes from ancillary data. A second is the use of a simple predictive linear model of the target soil property on the ancillary variables. Here, results are presented and compared where proximally sensed gamma-ray (γ-ray) spectrometry and electromagnetic induction (EMI) data are used to predict the variation in topsoil properties (e.g. clay content and pH). In the first instance, the proximal data is numerically clustered using a fuzzy k-means (FKM) clustering algorithm, to identify contiguous classes. The resultant digital soil maps (i.e. $k = 2$–10 classes) are consistent with a soil series map generated using traditional soil profile description, classification and mapping methods at a highly variable site near the township of Shelford, Nottinghamshire UK. In terms of prediction, the calculated expected value of mean squared prediction error (i.e. $\sigma^2_{p,C}$) indicated that values of $k = 7$ and 8 were ideal for predicting clay and pH. Secondly, a linear mixed model (LMM) is fitted in which the proximal data are fixed effects but the residuals are treated as a combination of a spatially correlated random effect and an independent and identically distributed error. In terms of prediction, the expected value of the mean squared prediction error from a regression ($\sigma^2_{p,R}$) suggested that the regression models were able to predict clay content, better than FKM clustering. The reverse was true with respect to pH, however. We conclude that both methods have merit. In the case of the clustering the approach is able to account for soil properties which have non-linearity's with the ancillary data (i.e. pH), whereas the LMM approach is best when there is a strong linear relationship (i.e. clay).

1. Introduction

Spatial predictions of soil properties are needed for various purposes including agriculture and engineering as well as scientific disciplines such as soil science, ecology and hydrology (Goovaerts, 1997). For example, maps of clay content can be used to ascertain land-use potential, whilst maps of soil pH can indicate lime requirement to counteract soil acidity, or potential nutrient availability. However, the costs associated with soil sampling and laboratory analysis are substantial, and spatial prediction requires considerable sample effort given the observation by Webster and Oliver (1992) that approximately 100 sample points are required to estimate a spatial statistical model. One way to improve soil sampling efficiency is to combine direct measurement of soil properties with collection of cheaper-to-measure ancillary data. Ancillary data can be used to improve precision with which properties are predicted from relatively few direct observations. Hence the growing interest in proximal geophysical sensing methods (Robinson et al., 2008) which have been applied to a range of problems including soil salinity assessment (Lesch et al., 2005), prediction of depth to clay (Jung et al., 2006), soil moisture determination (Robinson et al., 2012), determination of soil cation exchange capacity (Triantafilis et al., 2009a) and deep drainage estimation (Woodforth et al., 2012).

In this paper we consider two possible approaches. The first is to use ancillary data to form a set of land classes by a numerical

Abbreviations: EMI, Electromagnetic induction; EC_a, bulk soil electrical conductivity.

* Corresponding author.
E-mail address: j.triantafilis@unsw.edu.au (J. Triantafilis).

http://dx.doi.org/10.1016/j.geoderma.2014.04.031

clustering algorithm. The mean value of the soil property in each class, estimated from samples within each class, can then be used for prediction. This approach could be useful because it makes no assumptions about the nature of the relationship between the soil property and the ancillary variables and because precise estimates of class means can be obtained from bulk samples formed by aggregating individual sample cores within the class thereby reducing analytical costs. One practical question for the implementation of this approach is how many classes should be defined. This is usually addressed by considering the distribution of the ancillary variables used to form the classes, looking for evidence of compact structures in feature space (e.g. Triantafilis et al., 2009b). The rationale of this approach is that the classes so-identified reflect natural clusters in the feature space rather than an arbitrary partition, and so should reflect underlying sources of variation in the soil. Another approach (not used in this context to date) is prediction-based. As we consider more and smaller classes the within-class variance of the soil properties we wish to predict will, in general diminish, but the prediction error does not necessarily because the class mean is estimated with less precision as a fixed sample effort is divided between more classes (Huang et al., 2014).

A second and more commonly-used approach is linear predictive modeling, essentially a multiple regression of the target soil property on the ancillary variables. Ideally this is done using data obtained from a probability sample so residuals can be treated as independent. The model is then used to form a prediction of the target property at a site where only ancillary data is known. Often data are not collected according to a probability design, in which case a linear mixed model (LMM) fitted in which covariates are fixed effects but the residuals are treated as a combination of a spatially correlated random effect and an independent and identically distributed error (Lark et al., 2006). The prediction of the soil property at an unsampled site is then a combination of a regression-type prediction from proximally-sensed covariates and a kriging-type prediction of residuals from the fixed effects model at sampled sites (e.g. Gooley et al., 2014).

In this paper we consider both approaches, showing how the question 'how many classes?' can be addressed in terms of the uncertainty of resulting predictions, and compared with the linear mixed model. We illustrate this with a case study in which γ-ray spectrometry and the apparent electrical conductivity using an electromagnetic (EM) induction instrument were measured as ancillary data across two fields located east of the village of Shelford near Nottingham in the UK. We formed classes from the ancillary data using fuzzy k-means (FKM) analysis. We then analyze data on soil properties along with the classes formed from the ancillary data and the ancillary data themselves. We show how the precision of class means as predictors of soil properties (for fixed total sample effort) varies with the number of classes and compare this criterion for the number of classes with measures based on the distribution of the ancillary data. We also compare these measures of precision with comparable ones for direct prediction from the ancillary data by a linear model.

2. Materials and methods

2.1. Study area

The study fields (Fig. 1) are located east of the village of Shelford, which lies approximately 4 km east of Nottingham in

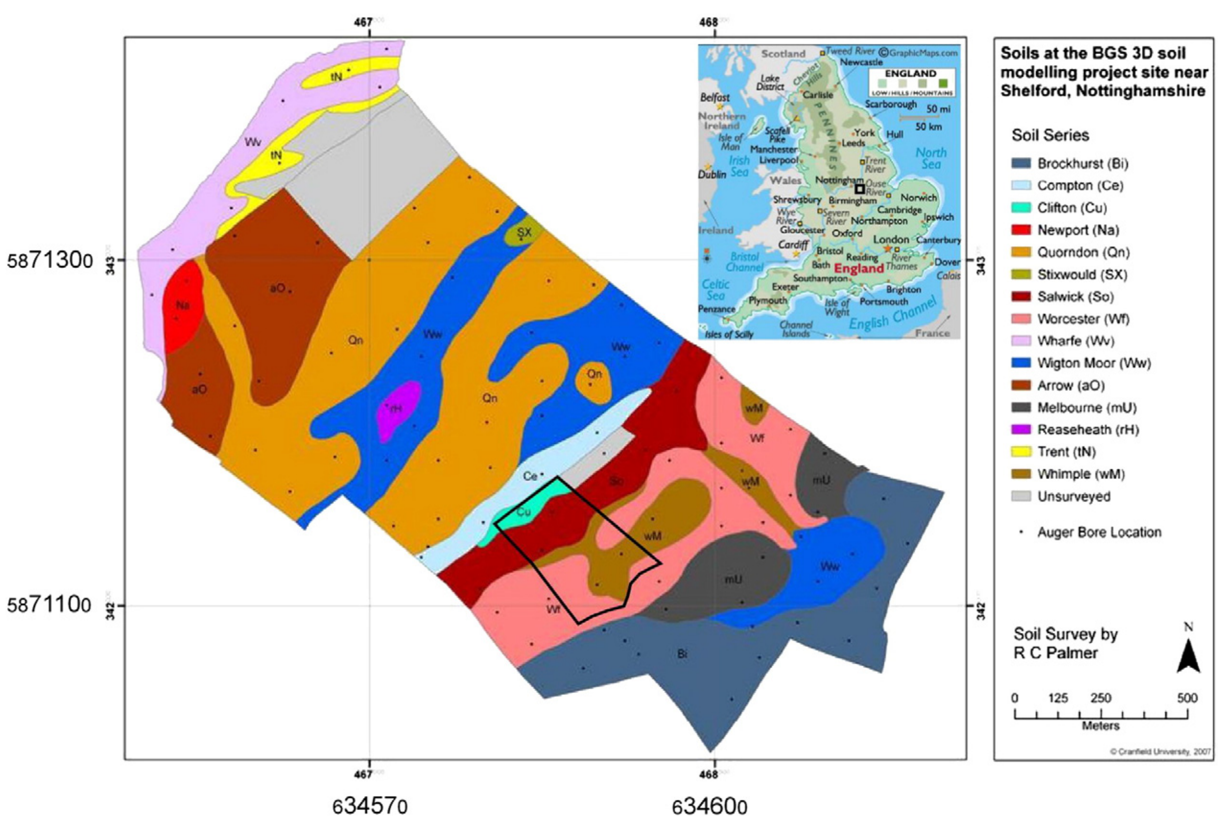

Fig. 1. Location of study area east of Nottingham and River Trent and the soil series map.