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], J. Triantafilis[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]

[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.

Correspondence: E-mail: John Triantafilis [j.triantafilis@unsw.edu.au]

Key Words: EC_a, induction, EM, fuzzy k-means, gamma-ray spectrometry, soil variability, characterization and distribution, digital soil mapping

Abbreviations: EMI Electromagnetic induction, EC_a bulk soil electrical conductivity

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 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 asa fixed sample effort is divided between more classes.

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., 2013).

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 analyse 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.

Figures 1 and 2 (Near here)

2. Materials and methods

2.1 Study area

The study fields (Figure 1) are located east of the village of Shelford, which lies approximately 4 km east of Nottingham in the valley of the River Trent. Geologically, the area consists of recent Holocene alluvial deposits and Pleistocene river terraces formed ~24,000 years BP. These lie unconformably over a series of Triassic mudstones and sandstones (Mercia Mudstone) to form a shallow floodplain aquifer. The annual average precipitation is 615 mm and evaporation 517 mm. The eastern field is used for pasture, whereas the field to the west for no-till arable farming (rape/wheat).

Recently a detailed (1:10,000) free soil survey was undertaken (Palmer, 2007). Figure 1 shows the soil series map. These are all simple map units that correspond to fifteen series of the classification of the Soil Survey of England and Wales (Avery, 1980). Four are in Trent alluvium (i.e. Wharfe, Trent, Compton and Stixwould); five on Trent River terraces (i.e. Newport, Arrow, Reaseheath, Quorndon and Wigton Moor series); and, six on the Triassic Mercia Mudstone (e.g. Worcester, Whimple, Brockhurst, Melbourne, Salwick and Clifton).

In brief, the northern end of the arable field is located on Mercia Mudstone where Clifton (Cu) and Salwick (So) are found. At the southern end Worcester (Wf) is common; however, slightly lighter-textured soil occurs in the middle and in a thin strip along the northern part where the Wimple (wM) map unit is located on the same parent material (i.e. mudstone). The pasture field located a short distance (~1.3 km) to the west is not in the mapped area, but from the topographic relations we might

reasonably expect the southwest side is associated with the deep permeable loams of the Arrow (aO) series of the Trent River terrace, whilst to the northeast the permeable alluvial medium loams of the Wharfe (Wv) series, associated with the Trent alluvium, dominate.

2.2 Ancillary instruments and data collection

Two sources of ancillary data were acquired with proximal sensors. To characterize the topsoil we collected γ-ray spectrometry data. An Exploranium GR-320 portable γ-ray spectrometer mounted in a backpack was used in conjunction with a handheld GPX-21 detector containing a 76 × 76-mm NaI (Tl) scintillation crystal, held at an approximate height of 1 m. At this height the instrument detects γ-radiation from an area within approximately a 10-metre radius (Atomic Energy Commission, USA, 1972). The detector was energy-stabilised with a small 133Ba source. The measured environmental levels of radiation are given in counts per second (cps) for the total counts (TC), percent for potassium (K %), and parts per million for uranium (U ppm), and thorium (Th ppm). In the study field these were recorded using integration over a 5 s time window as the operator walked slowly over the ground. Data were collected using a TDS Ranger palmtop running Pocket GIS software. Positioning data (SBAS enabled) were obtained from an internal Compact Flash card GPS receiver in the Ranger palmtop. γ-ray spectrometer transects are shown in Figure 2b.

In order to infer subsurface and subsoil variation we also collected EM data with a DUALEM-1S, because it incorporates single horizontal co-planar (HCP) and perpendicular (PRP) receiver arrays

that operate at a low frequency (9 kHz). The transmitter is located at one end with the distance to the centre of the HCP receiver being 1 m. The depth of EC_a measurement is approximately 0-1.5 m (1mHcon). The distance from the transmitter to the PRP receiver is 1.1 m which enables depth of EC_a of 0-0.5 (1mPcon) (DUALEM-421S Manual, 2008). The DUALEM-1S was connected to an Archer field computer (Juniper Systems Inc. Logan, UT, USA) and Bluetooth Sirf-III Royaltek BT-GPS receiver (Royal Tek, Kuei Shan, Tao Yuan, Taiwan). Measurements were integrated using the HGIS software package (Starpal inc., 2531 wapiti road, Fort Collins, CO 80525, USA).

The DUALEM-1S surveys were conducted across the two fields after harvest on Sept 2-3, Oct 14 and Nov 9-10, 2011. The instrument was held 0.2 m above the ground, with the instrument aligned in parallel with the direction of travel. Measurements were made by traversing the fields across the prevailing slope and following a predetermined route. In the arable field the transect spacing of ~15m and ~12 m in the pasture field. DUALEM-1S transects are shown in Figure 2b. EC_a outliers associated with the remnants of a buried metallic pipe running north-south in the western field were removed.

2.3 Soil sampling and laboratory analysis

The soil of both fields was sampled on a square grid with an interval of 25 m (see Figure 2c). There were 68 sample sites in the pasture and 137 in the arable field. Soil sampling was conducted on Nov 9 2011 (pasture) and Sept 03 2011 (arable field). Soil samples were collected for the depth interval

0–0.15 m with a gouge auger. Each of the 205 samples was stored in an individual water-tight plastic bag, taken back to the laboratory, weighed and then stored in a cool room.

The samples were dried, homogenized and sieved to 2mm. The particle size distribution was determined from a subsample by first determining the various particle size fractions using a laser diffractometer. The results were reported in terms of percent sand, silt and clay as defined by the Soil Survey of England and Wales (Hodgson, 1976). Soil pH was measured using a 1 part soil to 2.5 parts water dilution. We report results for clay content and pH in terms of differences among the classes obtained from the FKM analysis, to aid pedological interpretation and consider these soil properties for evaluating the use of the proximal soil sensor data for prediction of soil mapping units and LMM.

2.4 Fuzzy k-means (FKM) analysis

There are an increasing number of papers where ancillary data such as EM measurements of apparent electrical conductivity are analysed to form classes with a numerical clustering algorithm. This includes the use of k-means (FKM) to cluster EM with either NDVI (Dang et al., 2011) or topographic wetness index (Priori et al., 2013) or more commonly FKM analysis of EM with Quickbird imagery (Guo et al., 2013) or γ-ray spectrometry data (Van Meirvenne et al., 2013). Here we use the FKM method. It is described in detail by McBratney et al. (1992).

In brief, the similarity between an individual i and a cluster c is measured to determine how much they are alike in multi-variate space (Bezdek, 1981). The best outcome minimizes the objective function $J(\mathbf{M},\mathbf{C})$:

$$J(\mathbf{M},\mathbf{C}) = a \sum_{i=1}^{n} \sum_{c=1}^{k} m_{ic}^{\phi} d_{ic}^{2}(x_i, c_c) \quad (1)$$

where $\mathbf{M} = m_{ic}$ is a $n \times k$ matrix of membership values (n denoting the number of objects), $\mathbf{C} = (c_{cv})$ is a $k \times p$ matrix of class centers (p denotes the number of variables), c_{cv} is the value of the center of class c for variable v, $x_i = (x_{i1},\ldots,x_{ip})^{\mathrm{T}}$ is the vector representing individual i, $c_c = (c_{c1},\ldots,c_{cp})^{\mathrm{T}}$ is the vector representing the center of class c, and $d_{ic}^{2}(x_i, c_c)$ is the square distance between x_i and c_c according to a distance measure (d_{ic}^{2}). We chose Euclidean given our local knowledge of the geology of the area (Bezdek, 1981).

The fuzziness exponent (ϕ) determines degree of fuzziness. When $\phi = 1$ this is equivalent to the hard partition. As ϕ increases, memberships tend to become uniform. The fuzziness performance index (FPI) and the normalized classification entropy (NCE) were then used to identify values for ϕ and k. This is because the FPI is a measure of continuity between classes:

$$FPI = 1 - \frac{kF - 1}{k - 1} \quad (2)$$

where F is the partition coefficient;

$$F = \frac{1}{n} \sum_{i=1}^{n} \sum_{c=1}^{k} (m_{ic})^2 \quad (3)$$

An FPI value of 1 suggests a very fuzzy classification, whilst a value approaching 0 indicates a hard one. The NCE is a measure of disorganization in data partitioning:

$$NCE = \frac{H}{\log k} \tag{4}$$

where H is the entropy function;

$$H = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{k} m_{ic}\log(m_{ic}) \tag{5}$$

Values approaching 0 indicate that the classes are well structured, whilst values approaching 1 suggest the classes are disorganized. Triantafilis et al. (2013) suggest that values around 0.5 provide a balance between continuity and structure. A quantitative can be discerned for ϕ and k using the derivative of $J(\mathbf{M},\mathbf{C})$ with respect to ϕ (Bezdek, 1981):

$$\frac{dJ(\mathbf{M},\mathbf{C})}{d\phi} = \sum_{i=1}^{n}\sum_{c=1}^{k} m_{ic}^{\phi}\ \log\left(m_{ic}\right) d_{ic}^{2} \tag{6}$$

To determine the number of k, the outcome of $J(\mathbf{M},\mathbf{C})$ partitioning of the ancillary data into k = 2 to 10 classes using increments in ϕ of 0.2 and between ϕ = 1.2 to 2.4 is considered. A suitable value of ϕ for a given value of k is determined when the derivative of $-J(\mathbf{M},\mathbf{C})$ with respect to ϕ is largest (McBratney and Moore, 1985).

2.5 Linear mixed model (LMM)

When soil data were collected according to a systematic sampling scheme, it would not be appropriate to fit a linear model by ordinary least squares, since the residuals cannot be treated as independent random variables. Rather we propose a linear mixed model for the data of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta} + \boldsymbol{\varepsilon}, \quad (7)$$

where $\mathbf{y}$ is a $n\times1$ vector of values of the target soil variable, $\mathbf{X}$ is a $n\times p$ design matrix, $\boldsymbol{\beta}$ is a $p\times1$ vector of fixed effects coefficients, $\boldsymbol{\eta}$ is a $n\times1$ vector the elements of which are a realization of a spatially correlated random variable and $\boldsymbol{\varepsilon}$, is a $n\times1$ vector the elements of which are a realization of an independent and identically distributed random variable. The elements of the design matrix are the predictor variables and the fixed effects coefficients correspond to these. For example, if the predictors are p classes then element $\{i,j\}$ of $\mathbf{X}$ is 1 if the I th observation corresponds to the jth class and zero otherwise. There is therefore exactly one element equal to 1 in each row of the design matrix. In this case the elements of $\boldsymbol{\beta}$ are the estimated mean values of the target soil variable in the respective classes.

The correlated random variable $\boldsymbol{\eta}$ is assumed to be normal and has mean zero and variance parameters which are familiar from the geostatistical literature. These are an overall variance, σ^2_η, and a distance parameter for a selected variogram function (e.g. the range of a spherical variogram). The error variable $\boldsymbol{\varepsilon}$ also has zero mean and a variance σ^2_ε. In an alternative form of this model the elements in the design matrix may be, in the first column, a column of ones and in the next p-1 columns the values of the p-1 proximally sensed auxiliary variable, in which case the vector $\boldsymbol{\beta}$ contains an intercept

and p-1 regression coefficients. We fitted models of the form in Equation (7) for target soil variables and with the fixed effects either the class of maximum membership for the FKM clustering of the ancillary variables with k = 2-10 or a subset of the ancillary variables in a regression type model. The fitting was done using the LME procedure from the NLME library for the R platform (Pinheiro et al., 2013; R Development Core Team, 2010).

Under this procedure, variance parameters for the random effects are first estimated by residual maximum likelihood (REML) and the fixed effects coefficients are then estimated by weighted least squares. The null hypothesis that all class means are equal (where the fixed effects are classes) or that the regression coefficients are zero (where the fixed effects are continuous variables), is tested by the Wald statistic. These methods are described by Lark et al. (2006). After model fitting summary statistics and histograms of the residuals were examined to check that these appeared consistent with an assumption of normality.

The subset of ancillary variables for the model with continuous fixed effects was selected by first fitting a full model with all of the γ-ray (K, U, Th and TC) and DUALEM (1mHcon and 1mPcon) data as predictors. This was then compared to a series of reduced models by dropping each predictor in turn, and the full and reduced models were compared by computing their log-likelihood ratio and testing this against chi-squared with one degree of freedom (Verbeke and Mohlenbergs, 2000). Any predictor where the reduced model formed by dropping it was not significantly worse than the full model was rejected. This procedure was repeated until no further predictors were rejected. This

procedure requires maximum likelihood rather than REML estimation since residual likelihoods cannot be compared between models with different fixed effects. Once a predictor set was selected the model was then re-estimated by REML.

2.6 Computation of the prediction error variance for class means

The objective of this paper is to compare two approaches to the prediction of soil properties from the ancilliary variables, given that a relatively small set of direct measurements of the soil property is available. To do this we compute the expected value of the mean squared prediction error for the alternative methods:

$$\sigma^2_p = E[\{y - y^*\}^2] \quad (8)$$

where y denotes the value of the target variable at some unsampled location and y* denotes the predicted value.

When the predictor is the mean of a class (here obtained by cluster analysis of the ancillary data) then the mean-squared prediction error in class *i* is

$$\sigma^2_{p,i} = \sigma^2_i (1 + 1/n_i) \quad (9)$$

where σ^2_i is the variance of the target property within class *i* and the mean of class *i* was estimated from n_i independently and randomly selected observations within the class (Brus and Lark, 2013). In this study we use a pooled within-class variance (σ^2_w). If π_i denotes the relative area of the *i*th class out

of k and N is the total number of observations then the expected value of the mean squared prediction error for classes is:

$$\sigma^2_{p,C} = \sum_{i=1,\ ..k} \sigma^2_w \pi_i \ (1 + 1/N\pi_t) = \sigma^2_w(1+k/N). \quad (10)$$

In general, as k increases we expect the classes to become internally more uniform with respect to soil properties, so σ^2_w should decrease. However, it is apparent that the term in brackets on the right-hand side of Equation (10) will increase with increasing k, and that this increase will be greater the smaller is N. In summary, $\sigma^2_{p,C}$ will only decrease with increasing k if the reduction in the within-class variance is large enough to compensate for the fact that the fixed total sample size is spread more thinly over more classes which contributes to the uncertainty with which the class means are estimated.

In this study we computed $\sigma^2_{p,C}$ for each target soil property for k = 2-10 classes formed by the FKM algorithm. To do this we require a value of σ^2_w. This was obtained from the LMM, Equation (7), fitted to the observed soil data for the corresponding classification. The sum of the variances of the random effects in the model for k classes as the random effects, $\sigma^2_{\eta,k} + \sigma^2_{\varepsilon,k}$, was treated as the expected value of the variance for the random variable. This approach is used elsewhere to compute values for the variances of design-based sample estimates from the results of model-based analyses — Cochran (1977); Lark (2011) provides an example in soil science. The expected value of the mean squared prediction error for our classification into k classes for some sample size N is computed here as

$$^*\sigma^2_{p,C}(N \mid k) = (\sigma^2_{\eta,k} + \sigma^2_{\varepsilon,k})(1+k/N). \quad (11)$$

2.7 Computation of the prediction error variance for regression models

For the case of prediction direct from a selected subset of proximally sensed variables by a multiple regression-type predictor the mean-squared prediction error for a particular prediction at an unsampled site is:

$$\sigma^2_{p,R} = \sigma^2 (1+ \mathbf{x}^T\{\mathbf{X}^T\mathbf{X}\}^{-1}\mathbf{x}) \quad (12)$$

where $\mathbf{X}$ is the design matrix for the data set used to predict the model and $\mathbf{x}$ is a vector in which the ith element is the difference between the value of the ith predictor for the particular prediction and the overall mean of the ith predictor (Dudewicz and Mishra, 1988). The term σ^2 is the residual variance of the fitted regression. To compute the expected value of $\sigma^2_{p,R}$ for some simple random sample of size N we evaluated Equation (12) for values of the selected predictor variables at each of the M=205 observation sites and computed the average.

We obtained the expression $\mathbf{X}^T\mathbf{X}$ by computing it from the design matrix, $\mathbf{X}$, for our original M observations and rescaling it for a sample of size N and, as before, we used the sum of the REML estimates of the variances of the random effects in the corresponding LMM (Equation 7), $\sigma^2_{\eta,R} + \sigma^2_{\varepsilon,R}$, as the expected value of σ^2. The expected value of the mean squared prediction error from a regression estimated for some sample size N is therefore computed here as

$$^*\sigma^2_{p,R}(N) = (\sigma^2_{\eta,R} + \sigma^2_{\varepsilon,R})\{1+ 1/M \sum_{i=1,\ldots M} \mathbf{x}_i^T\{(N/M)\mathbf{X}^T\mathbf{X}\}^{-1}\mathbf{x}_i\} \quad (13)$$

where $\mathbf{x}_i$ is the vector of predictor values for the ith of our original M = 205 sampled locations and $\mathbf{X}$ is the design matrix for that same set.

2.8 Data analysis

The γ-ray spectrometry (i.e. K, U, Th and TC) and DUALEM-1S (i.e. 1mHcon and 1mPcon) data were first interpolated onto a common 10-m grid. This was done by ordinary kriging (OK) within a neighborhood of 20-30 and a local variogram. The Vesper program (Minasny et al., 1999) was used. Numerical clustering of the OK ancillary data was conducted by FKM analysis using FuzME 3.0 (Minasny and McBratney, 2002).

Figure 3 (Near here)

3. Results & discussion

3.1 Spatial distribution of proximally sensed data

Figure 3a) shows the spatial variation of TC. The smallest to intermediate-small counts (<40 cps) are found in the centre of the pasture field. Conversely, intermediate-large (50-60 cps) and large (> 60 cps) TC define the southern parts of the arable field. Similar patterns are evident in K (Figure 3b) and Th (not shown). It is worth noting the intermediate radioelement values (e.g. K % = 1.5-2 %) at the southern end of the arable field and in addition the large K readings (>2.5 %) to the south. Figure 3c) shows the spatial variation of the 1mHcon. For the most part the spatial patterns are similar. This is particularly the case in the pasture field where EC_a is smallest (< 6 mS/m). A difference between the γ-

ray and EC_a data is observed in the western half of the southern part of the arable field. Here 1mHcon is intermediate-large (18-24 mS/m) compared to the eastern half, which is large (> 24 mS/m).

Figures 4 and 5 and Tables 1 (Near here)

3.2 FKM analysis

Figure 4a) and b) shows the FPI and NCE, respectively for clustering of the sensor data for different values of k and ϕ. It is evident that the FPI is at a local minimum when k = 2, 3 and 4 when ϕ = 1.4 and when k = 3 for values of ϕ from 1.6 to 2.4. At several values of ϕ local minima are evident for k = 5 (ϕ = 1.6), k = 7 (ϕ = 1.8) and k = 8 (ϕ = 2.0). As with the FPI, the NCE is at a minimum for each value of ϕ considered and when k = 3. Local minima are also evident for k as indicated above. Given the equivocal nature of the results, that is the FPI and NCE do not provide a clear indication of what an appropriate number of k might be in the ancillary data, we looked to the plot of ϕ versus $-dJ(\mathbf{M},\mathbf{C})/ d\phi$. Figure 4c shows that where the derivative of $-J(\mathbf{M},\mathbf{C})$ is a maximum, McBratney and Moore (1985) indicate this is where ϕ is optimal. In most cases (i.e. k = 5-7) this occurs when ϕ = 2.0. Given these results we selected an exponent of ϕ equal to 2.0 and we compare the results of k = 2-10 classes with the soil series previously recognized by Palmer (2007).

3.3 Spatial distribution of the FKM classes

Figure 5a) shows the result for k = 4. Of the four classes, three are identified in the pasture field. The largest contiguous area is demarcated by 4A which is defined by small radioelement values (e.g. K

= 0.86 %) and EC_a (e.g. 1mPcon = 7.69 mS/m). Class 4B defines the eastern margin, whilst 4C is found within 4B as inclusions. In the arable field, 4B defines the northern third, having slightly larger radioelements (e.g. K = 1.90 %) and EC_a (e.g. 1mPcon = 14.78 mS/m) and matches the area ascribed by the So series. At the southern end, the western half is defined by 4C which was mapped as the wM series. Of all the classes, 4D has the largest radioelement (e.g. K = 2.36 %) and EC_a (e.g. 1mPcon = 42.11 mS/m). Moreover the areal extent of 4D coincides with the location of the Wf series.

Figure 5b) shows the spatial distribution for k = 5. The only real difference is that 4A broadly corresponds to 5A and another class (i.e. 5E), and 5B defines the northern end of the arable field. Figure 6c) shows similar patterns for k = 7. Here 5D broadly corresponds to 7D and 7F, whereby the latter has slightly smaller radioelement values (e.g. K = 2.31 %) and EC_a (e.g. 1mPcon = 40 mS/m). These results are consistent with Tye et al. (2011), who generated an automated resistivity profile map. In doing so they inferred that the distinct linear variation in the area denoted by 7D is due to outcrops of siltstone beds of the Gunthorpe Member. It is also worth noting that class 5B similarly corresponds to 7B and 7G. The former represents the So series, whilst the latter represents the previously unrecognized Cu series (see Figure 5c). When k = 8 classes are considered 7E corresponds to 8E and 8H, otherwise the remaining classes are equivalent (Figure not shown).

Figures 6 and 7 (Near here)

3.4 Mean squared prediction error of a map

Figure 6 shows the results of the calculated expected value of mean squared prediction error (i.e. $\sigma^2_{p,C}$) of the estimated class means for different total sample sizes for each of the k = 2-10 maps. Figure 6a) shows the result for clay. It is evident that with increasing k, $\sigma^2_{p,C}$ decreases. This is particularly the case between k = 2 and 6 classes. A minimum is reached when k = 7 and 8. Beyond this $\sigma^2_{p,C}$ increases because, while the classes may become increasingly internally uniform, the total sample effort is now divided over too many classes to provide adequate estimates of class means.

This approach to the selection of the number of classes to use is novel. It differs from previously-used approaches which focus purely on the internal uniformity of the classes because it considers the pragmatic question of the precision of predictions based on the classification. In particular note that the value of the criterion depends on the total sample size that we assume we have available for calibration. The strength of evidence for a minimum value of the $\sigma^2_{p,C}$ at some value of k is also sensitive to the size of the calibration sample.

Figure 6b) shows the equivalent plot but for pH. Here $\sigma^2_{p,C}$ decreases to a local minimum at k = 4, but increases again until a global minimum is attained at k = 7 and 8. This is consistent with the FPI and MPE metrics of the FKM algorithm; which indicated that k = 7 or 8 might be a suitable number of classes in the ancillary data.

3.5 REML analysis of FKM classes

Here we interpret the mean values of soil properties for classes obtained by FKM analysis of the sensor data with k = 7 because their spatial distribution most closely reflects the soil map (Palmer, 2007). Figure 7 shows the mean values of the soil properties and their standard error for each of the classes obtained from the LMM estimated by REML. In all cases the Wald statistic allows us to reject the null hypothesis of no difference among the class means.

The particle size fraction appears most revealing about the partitioning of the k = 7 classes. Figures 7a) shows that 7A has the smallest clay (22.3 %). Conversely, class 7D had the largest clay (45.7 %) followed closely by 7F. The four remaining classes (i.e. 7B, 7C, 7E and 7G) have similar mean clay (28.5-34.3 %). Figure 7b) shows that soil pH is largest for 7D (7.1) and the closely related 7F (6.9). It is worth noting that 7B (5.6) and 7G (6.9) have markedly different pH. This is interesting, given the FKM analysis of the proximally sensed data did not differentiate these classes until k = 7. Similarly, and whilst 7A (6.1) and 7E (6.2) are distinguishable early on in the classification (i.e. k = 4) soil pH is not different between these classes. Worthy of note is that 7B is different from 7D and 7C despite all these classes being associated with the Mercia mudstone, but it is similar to 7A and 7E which are not.

3.6 Comparison of FKM clustering and regression for predicting soil properties

In order to determine whether classifying the γ-ray spectrometer and EM data using FKM provides better prediction of soil properties than developing regression models we compared the $\sigma^2_{p,C}$

and $\sigma^2_{p,R}$, respectively. First we developed the regression models. The reduced models for clay and pH and the model parameters calculated from R platform are shown in Table 2. In terms of mapping clay, the best combination of ancillary data is 1mHcon, 1mPcon and K. However, 1mHcon, 1mPcon and U were selected for mapping pH.

The calculated $\sigma^2_{p,R}$ are shown in Figure 6 and for sample sizes of 20, 40 and 80. The first thing to note is that as sample size increases, $\sigma^2_{p,R}$ drops accordingly. This is consistent with the effect of sampling with regard to the FKM classes (i.e. $\sigma^2_{p,C}$). In order to compare the $\sigma^2_{p,C}$ using FKM class means and $\sigma^2_{p,R}$ using regression models, mean square prediction errors are evaluated for the same sample size (e.g. 80). In terms of predicting clay content, the regression models generated better predictions than can be achieved using class means by FKM clustering. However, when predicting pH, FKM clustering performs better than the linear regression model when $k = 6 – 9$.

4. Conclusions

A catenary sequence characterised by a broad range of textural variation and associated with Trent alluvium, Trent River terraces, and Triassic Mercia mudstone, was partitioned into k = 2-10 classes, using fuzzy k-means (FKM) analysis of four radioelement windows (K, U, Th and TC) acquired from a γ-ray spectrometer and two proximally sensed EC_a data (1mPcon and 1mHcon). The use of the FKM algorithm, along with various indices (e.g. FPI and $-dJ(\mathbf{M},\mathbf{C})/d\phi$), suggested that partitioning the data into k = 3, 7 or 8 classes and using a ϕ of 2.0 was most suitable, with the results

broadly reflecting a soil series map developed by an experienced soil surveyor using traditional morphological site descriptions and a pre-existing soil classification scheme (Palmer, 2007).

To test this we determined the mean prediction error variance ($\sigma^2_{p,C}$) of the class mean as a predictor for a soil physical (e.g. clay content) and chemical (i.e. pH) property. Using this independent approach the results indicated that k = 7 and 8 were statistically different and accounted for most of the soil variation. Prediction of soil properties by FKM analysis and regression models was also compared. The results of this analysis indicated that for a sample size of 80, the regression models were able to predict clay content, better than FKM clustering. However, when predicting pH, FKM clustering performs better than the linear regression model when k = 6 – 9. It is concluded that both the FKM and LMM methods have merit. In the case of the clustering, the approach is able to account for soil properties which have non-linearity with the ancillary data (i.e. pH), whereas the LMM approach is best when there is a strong linear relationship (e.g. clay).

In order to test this further, for example if we wanted to make predictions of soil properties across the Trent valley where conditions are homologous with our study site, we would be unable to sample with enough intensity to use kriging. However, we could collect proximally sensed γ-ray spectrometer and DUALEM-1 data, plus a set of calibration data. Then using the theory for computing the $\sigma^2_{p,R}$ for the regression and the $\sigma^2_{p,C}$ for classes of different sizes we could predict variations of the properties using regression or by using class means as predictors.

5. References

Avery, B.W., 1980. Soil Classification for England and Wales (Higher Categories). Soil Survey Technical Monograph No. 14. Harpenden.

Bezdek, J.C., 1981. Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York.

Brus, D., Lark, R.M., 2013. Soil survey. In (A.H. El-Shaarawi, ed.) Encyclopaedia of Environmetrics, 2nd Edition, John Wiley & Sons.

Cochran, W.G., 1977. Sampling Techniques. Wiley, New York.

Dang, Y. P., Dalal, R. C., Pringle, M. J., Biggs, A. J. W., Darr, S., Sauer, B., . . . Orange, D. 2011. Electromagnetic induction sensing of soil identifies constraints to the crop yields of north-eastern Australia. Soil Research, 49, 559-571.

DUALEM-421S., 2008. User's Manual. Dualem Inc., Canada.

Dudewicz, E. J., Mishra, S. N., 1988. Modern mathematical statistics. Wiley, New York.

Gooley, L., Huang, J., Page, D., Triantafilis, J., 2013. Digital soil mapping available water content using proximal and remotely sensed data, Soil Use and Management, in press.

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford Univ. Press, New York.

Guo, Y., Shi, Z., Li, H.Y., Triantafilis, J., 2013. Application of digital soil mapping methods to identify salinity management classes in coastal lands of central China. Soil Use and Management, 29, 445-456.

Hodgson, J.M. 1976. Soil survey field handbook, 2nd ed. Technical monograph (Soil Survey of England and Wales), No. 5. Soil Survey of England and Wales, Harpenden,UK.

Jung, W. K., Kitchen, N. R., Sudduth, K. A., Anderson, S. H., 2006. Spatial characteristics of claypan soil properties in an agricultural field. Soil Science Society of America Journal 70, 1387-1397.

Lark, R. M., Cullis, B. R., Welham, S. J., 2006. On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. European Journal of Soil Science 57, 787-799.

Lark, R. M., 2011. Spatially nested sampling schemes for spatial variance components: Scope for their optimization. Computers and Geosciences 37, 1633-1641.

Lesch, S. M., Corwin, D. L., Robinson, D. A., 2005. Apparent soil electrical conductivity mapping as an agricultural management tool in arid zone soils. Computers and Electronics in Agriculture 46(1-3 SPEC. ISS.), 351-378.

McBratney, A.B., Moore, A.W., 1985. Application of fuzzy-sets to climatic classification. Agricultural Forest Meteorology 35, 165-185.

McBratney, A.B., De Gruijter, J.J., Brus, D.J. 1992. Spatial prediction and mapping of continuous soil classes. Geoderma 54, 39-64.

Minasny, B., McBratney, A.B., Whelan, B.M., 1999. VESPER version 1.6, Precision Agriculture Laboratory, Sydney, Australia.

Minasny, B., McBratney, A.B., 2002. FuzME version 3.0, Precision Agriculture Laboratory, The University of Sydney, Australia.

Palmer, R.C., 2007. The soils of Nottingham Trent University farm at Brackenhurst, Nottinghamshire. NSRI unpublished research, report no. YE20006E for the British Geological Survey.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., The R Development Core Team, 2013. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-110.

Priori, S., Martini, E., Andrenelli, M. C., Magini, S., Agnelli, A. E., Bucelli, P., Biagi, M., Pellegrini, S., Costantini, E. A. C. 2013. Improving wine quality through harvest zoning and combined use of remote and soil proximal sensing. Soil Science Society of America Journal, 77, 1338-1348.

R Development Core Team 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Robinson, D.A. Binley A., Crook N., Day-Lewis F.D., Ferre T.P.A., Grauch V.J.S. Knight R., Knoll M., Lakshmi V., Miller R., Nyquist J., Pellerin L., Singha K., and Slater L., 2008. Advancing process-based watershed hydrological research using near-surface geophysics: A vision for, and review of, electrical and magnetic geophysical methods. Hydrological Processes, 20, 3604-3635.

Robinson, D. A., Abdu, H., Lebron, I., Jones, S. B., 2012. Imaging of hill-slope soil moisture wetting patterns in a semi-arid oak savanna catchment using time-lapse electromagnetic induction. Journal of Hydrology 416, 39-49.

Triantafilis, J., Lesch, S. M., La Lau, K. & Buchanan, S. M. 2009a. Field level digital soil mapping of cation exchange capacity using electromagnetic induction and a hierarchical spatial regression model. *Australian Journal of Soil Research*, **47,** 651-663.

Triantafilis, J., Kerridge, B., Buchanan, S.M., 2009b. Digital soil-class mapping from proximal and remotely sensed data at the field level. Agronomy Journal 101, 841-853.

Triantafilis, J., Gibbs, I.D., Earl, N.Y., 2013. Digital soil pattern recognition in the lower Namoi valley using numerical clustering of gamma-ray spectrometry data. Geoderma 192, 407-421.

Tye, A. M., Kessler, H., Ambrose, K., Williams, J. D. O., Tragheim, D., Scheib, A., Kuras, O., 2011. Using integrated near-surface geophysical surveys to aid mapping and interpretation of geology in an alluvial landscape within a 3D soil-geology framework. Near Surface Geophysics 9, 15-31.

Van Meirvenne, M., Islam, M. M., De Smedt, P., Meerschman, E., Van De Vijver, E., & Saey, T. 2013. Key variables for the identification of soil management classes in the aeolian landscapes of north-west Europe. Geoderma, 199, 99-105.

Verbeke, G., Molenberghs, G. 2000. Linear Mixed Models for Longitudinal Data. Springer, New York.

Webster, R., Oliver, M. A., 1992. Sample adequately to estimate variograms of soil properties. Journal of Soil Science 43, 177-192.

Woodforth, A., Triantafilis, J., Cupitt J., Malik, R.S., Geering, H., 2012. Mapping estimated deep drainage in the lower Namoi Valley using a chloride mass balance model and EM34 data. Geophysics 77, WB245-256.

1

Table 1 Euclidean centroid values of proximally sensed ancillary data clustered using FKM and for classes k = 4, 5 and 7.

Note: centroids shown for K (%), U (ppm), Th (ppm), TC (cps-counts per second), 1mPcon and 1mHcon, respectively.

Note: Number of soil samples which are members of the classes also shown.

	k = 4		*k = 5*		*k = 7*	
	Centroid values	*members*	*Centroid values*	*members*	*Centroid values*	*members*
	0.86, 1.28, 3.74, 30.51, 8, 5	57	0.83, 1.26, 3.67, 29.79, 7, 5	49	0.81, 1.25, 3.64, 29.46, 6, 4	41
	1.90, 1.93, 5.76, 50.95, 15, 9	45	1.95, 1.96, 5.86, 51.86, 15, 9	36	1.95, 1.96, 5.85, 51.84, 15, 9	21
	2.09, 2.19, 6.18, 55.38, 35, 19	45	2.12, 2.20, 6.23, 55.91, 35, 20	47	2.07, 2.18, 6.15, 54.98, 34, 19	37
	2.36, 2.23, 6.61, 60.05, 42, 27	58	2.37, 2.23, 6.62, 60.16, 42, 26	54	2.43, 2.23, 6.72, 61.18, 44, 29	23
			1.30, 1.53, 4.55, 38.96, 19, 13	19	1.20, 1.48, 4.41, 37.23, 16, 11	23
					2.31, 2.22, 6.51, 59.06, 40, 24	42
					2.02, 1.99, 5.98, 53.27, 17.36, 11.17	15

Table 2 Reduced models for different soil properties.

Properties	Selected predictors	σ^2	Covariance function	Nugget fraction	Distance parameter
Clay content	1mHcon, 1mPcon, K	31.360	Spherical	0	33 m
pH	1mHcon, 1mPcon, U	0.276	Exponential	0.3	65 m

Note: σ^2 - residual variance of the fitted regression; nugget fraction - nugget/sill; and, distance parameter - distance parameter of the exponential correlation function or range of the spherical correlation function.

Figure Captions

Figure 1 Location of study area east of Nottingham and River Trent and the Soil Series Map.

Figure 2 a) Air-photo of the pasture and arable study fields, b) spatial location of the DUELEM-1S and gamma-ray (γ-ray) spectrometry survey transects, and c) soil sample locations.

Figure 3 Spatial distribution of gamma-ray (γ-ray) spectrometry data including; a) Total count (TC – counts per second), and b) potassium (K - %), and DUALEM-1S electrical conductivity (EC_a – mS/m) of c) 1mHcon (i.e. Deep).

Figure 4 Plot of; (a) fuzziness performance index (FPI), (b) normalized classification entropy (NCE) versus classes (k = 2 to 10) and (c) fuzziness exponent (ϕ) versus – dJ (M,C)/dϕ.

Figure 5 Spatial distribution of FKM classes for k = a) 4, b) 5, and c) 7.

Figure 6 Plot of mean squared prediction error for FKM classes (i.e. $\sigma^2_{p,C}$, solid lines) and linear regression models (i.e. $\sigma^2_{p,R}$, dashed lines) of (a) clay (%) and (b) pH. Note: Sample size are labeled (i.e. N = 20, 40, 80,…, 205).

Figure 7 Plot of mean and standard deviation of soil (a) clay (%) and (b) pH.

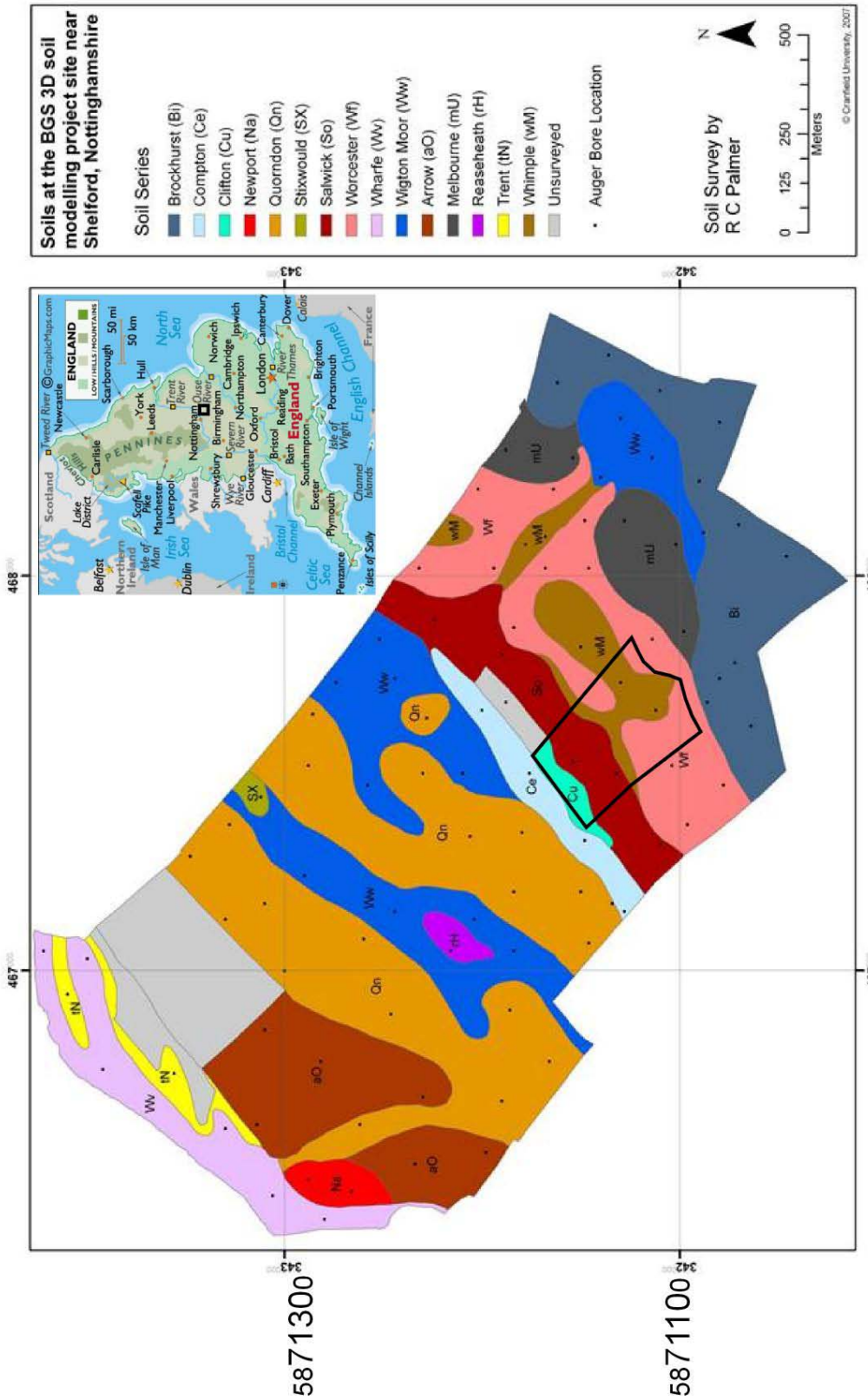
Soils at the BGS 3D soil modelling project site near Shelford, Nottinghamshire
Soil Series
Brockhurst (Bi)
Compton (Ce)
Clifton (Cu)
Newport (Na)
Quorndon (Qn)
Stixwould (SX)
Salwick (So)
Worcester (Wf)
Wharfe (Ww)
Wigton Moor (Ww)
Arrow (aO)
Melbourne (mU)
Reaseheath (rH)
Trent (tN)
Whimple (wM)
Unsurveyed
Auger Bore Location
Soil Survey by R C Palmer
N
0 125 250 500
Meters
© Cranfield University, 2007
ENGLAND
LOW HILLS, MOUNTAINS
50 mi
50 km
Scotland
Wales
Ireland
England
North Sea
Irish Sea
Bristol Channel
Celtic Sea
English Channel
France
Channel Islands
Isles of Scilly
Isle of Man
Northern Ireland
PENNINES
Lake District
Tweed River
Trent River
Severn River
Wye River
Thames River
Ouse River
Newcastle
Carlisle
Scarborough
Hull
York
Leeds
Liverpool
Manchester
Scafell Pike
Nottingham
Birmingham
Norwich
Cambridge
Ipswich
Northampton
Oxford
London
Canterbury
Dover
Calais
Shrewsbury
Gloucester
Cardiff
Bristol
Bath
Reading
Brighton
Portsmouth
Southampton
Isle of Wight
Exeter
Plymouth
Penzance
Belfast
Dublin
Chester
468
467
343
342
634600
634570
5871300
5871100

Figure 1

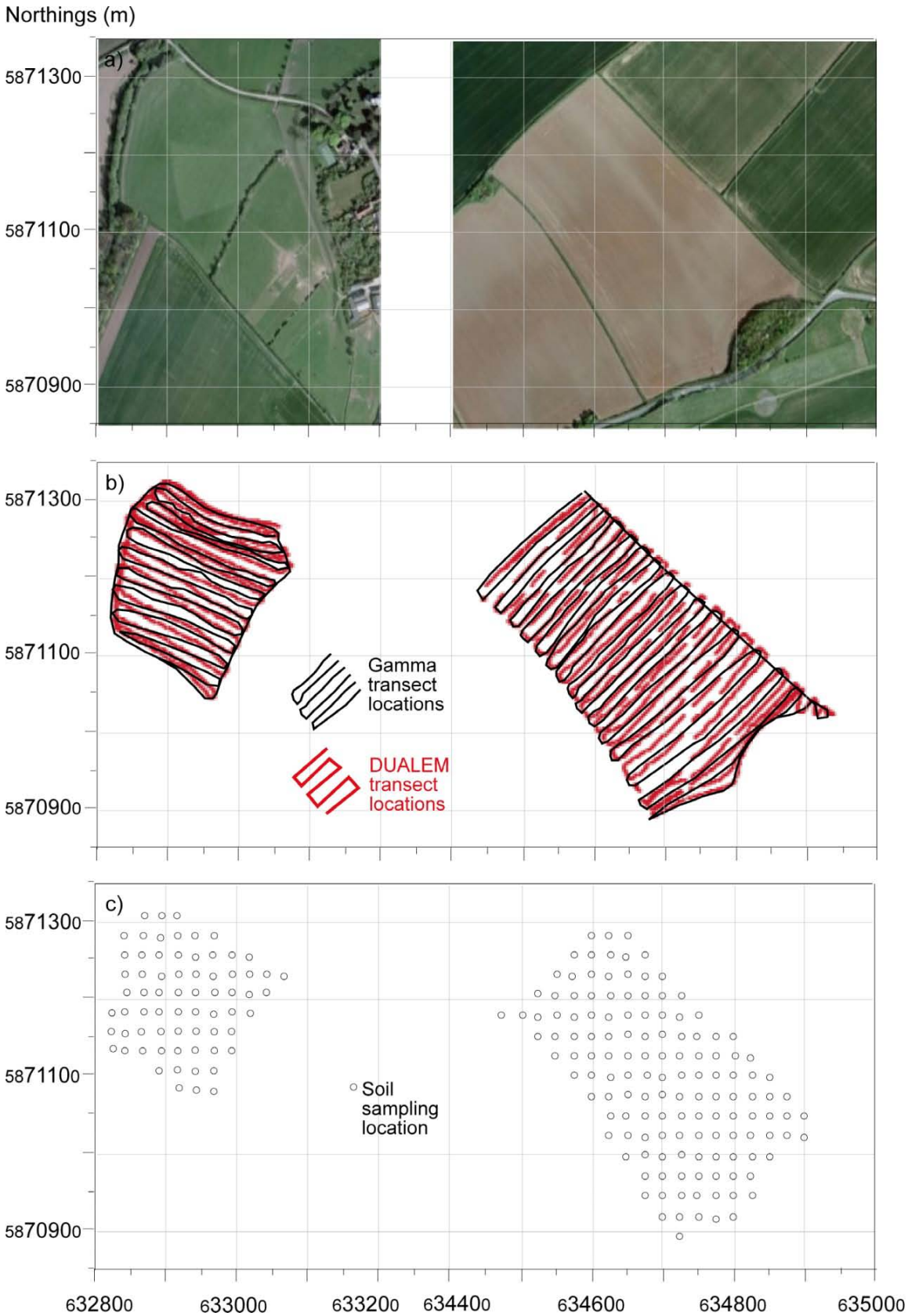
Northings (m)
a)
5871300
5871100
5870900
b)
Gamma transect locations
DUALEM transect locations
c)
Soil sampling location
632800
633000
633200
634400
634600
634800
635000

Figure 2

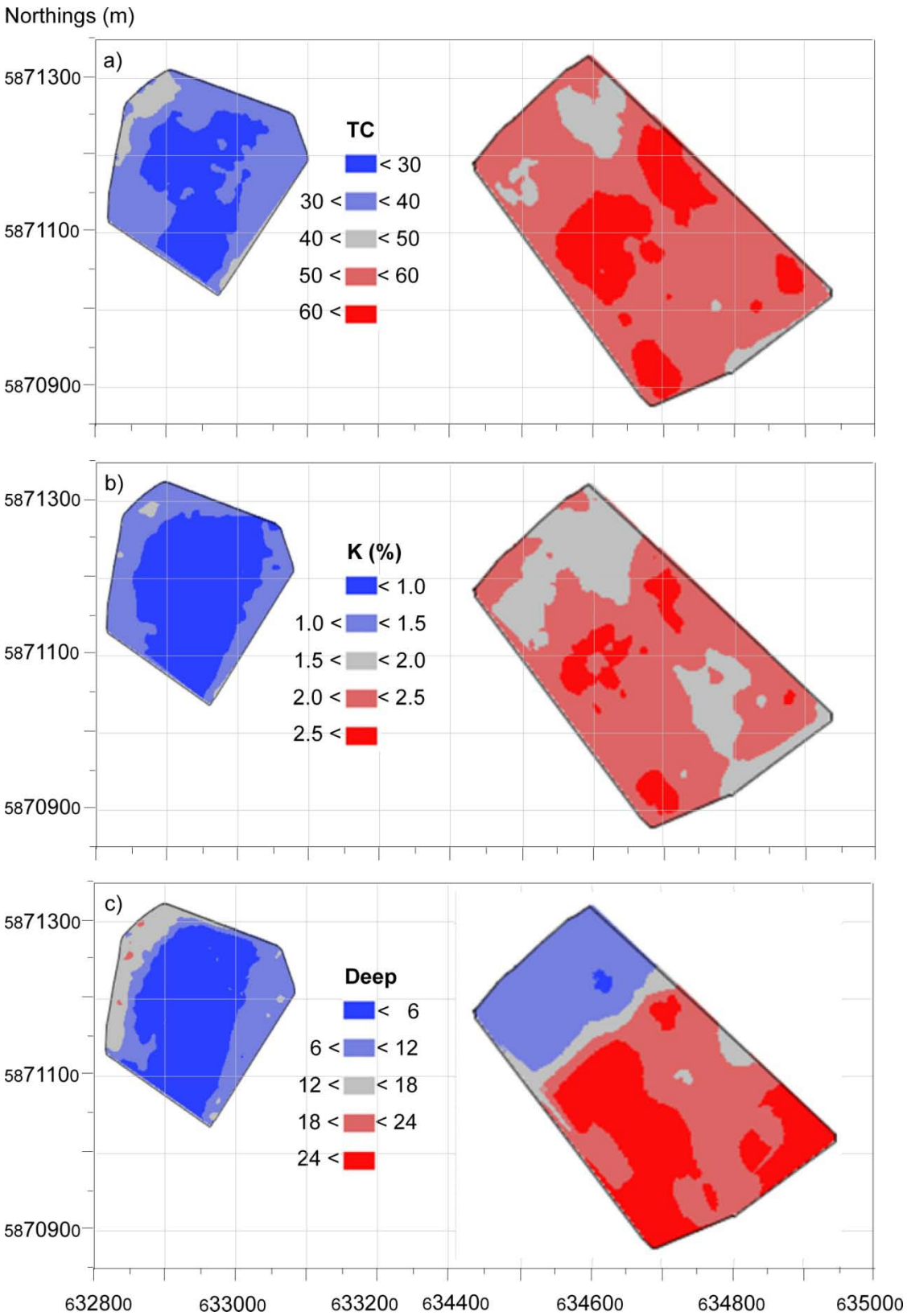
Northings (m)
a)
TC
< 30
30 < < 40
40 < < 50
50 < < 60
60 <
b)
K (%)
< 1.0
1.0 < < 1.5
1.5 < < 2.0
2.0 < < 2.5
2.5 <
c)
Deep
< 6
6 < < 12
12 < < 18
18 < < 24
24 <
5871300
5871100
5870900
632800
633000
633200
634400
634600
634800
635000

Figure 3

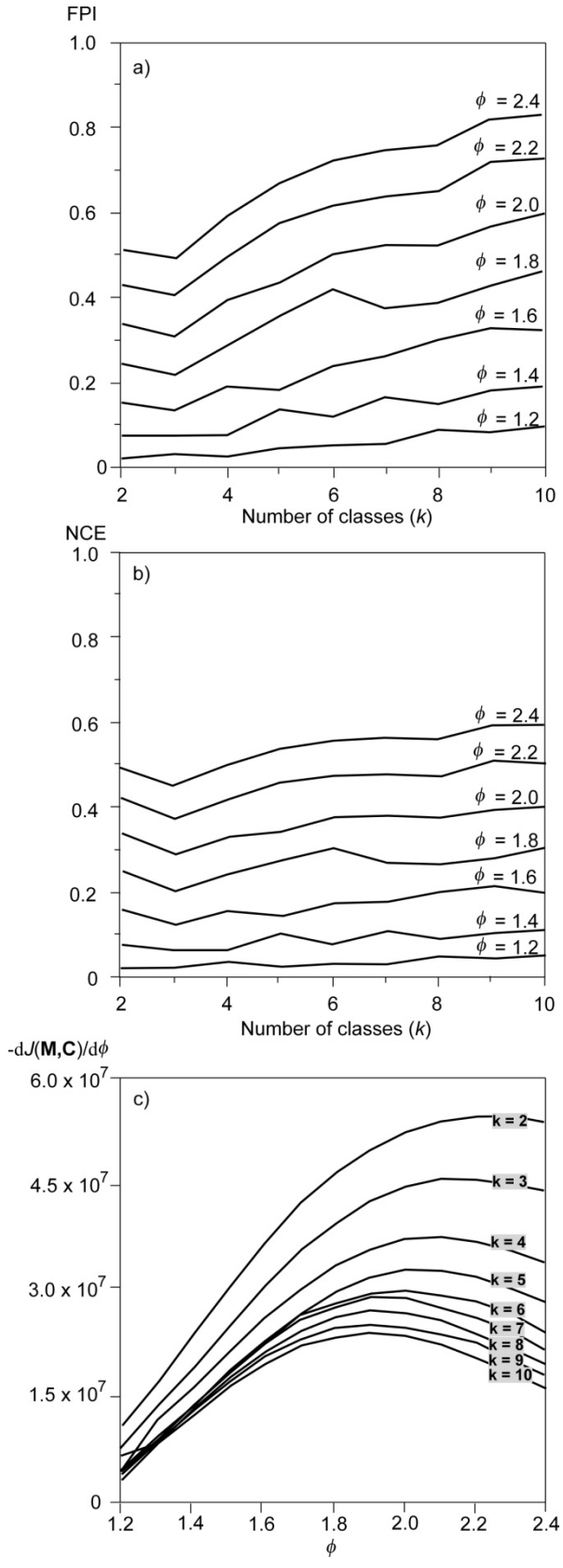
FPI
a)
ϕ = 2.4
ϕ = 2.2
ϕ = 2.0
ϕ = 1.8
ϕ = 1.6
ϕ = 1.4
ϕ = 1.2
Number of classes (k)
NCE
b)
Number of classes (k)
-dJ(M,C)/dϕ
c)
k = 2
k = 3
k = 4
k = 5
k = 6
k = 7
k = 8
k = 9
k = 10
ϕ

Figure 4

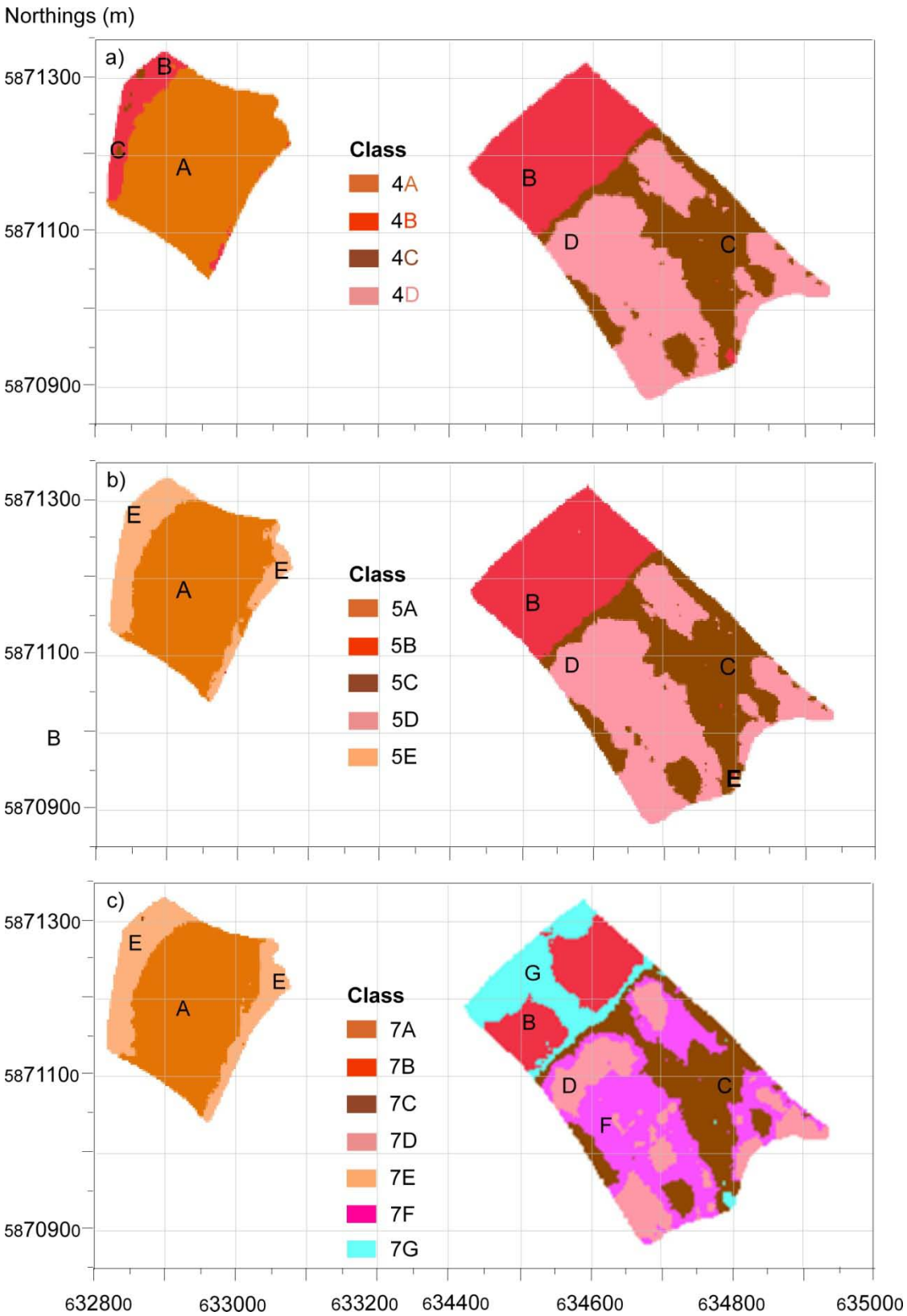
Northings (m)
a)
b)
c)
Class
4A
4B
4C
4D
5A
5B
5C
5D
5E
7A
7B
7C
7D
7E
7F
7G
5871300
5871100
5870900
632800
633000
633200
634400
634600
634800
635000

Figure 5

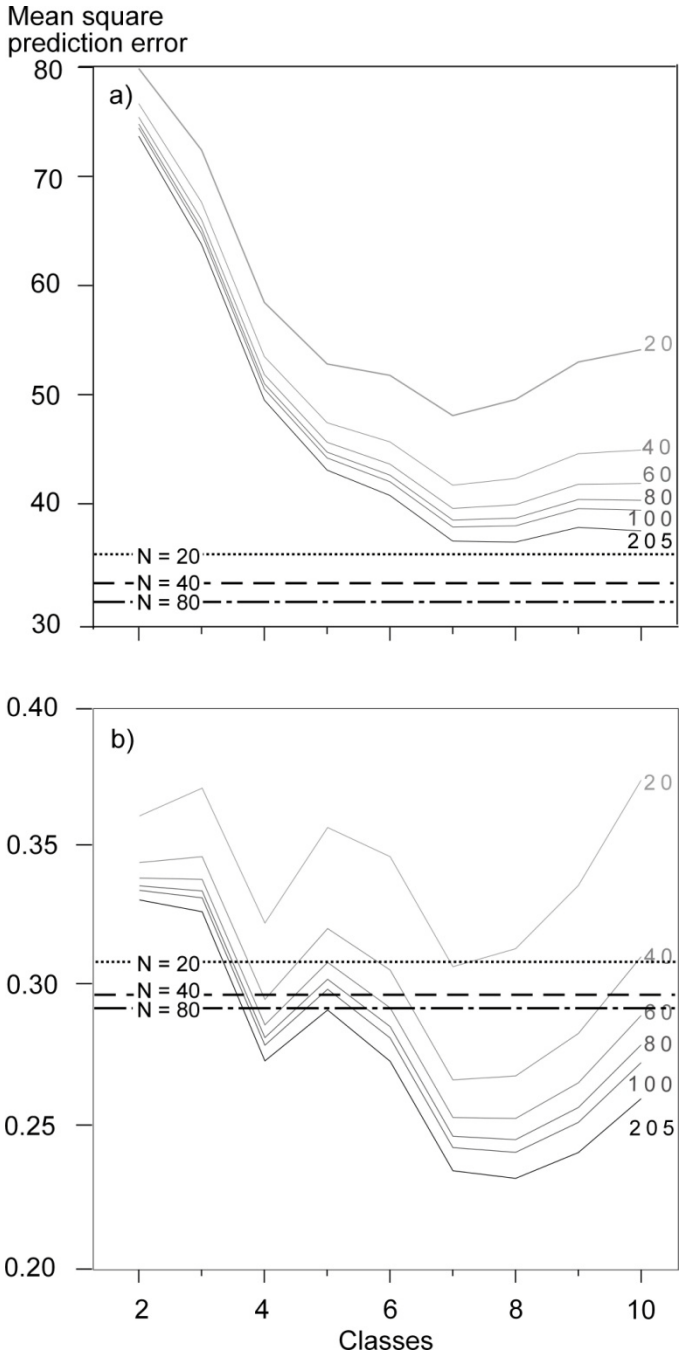
Mean square
prediction error
a)
80
70
60
50
40
30
20
40
60
80
100
205
N = 20
N = 40
N = 80
b)
0.40
0.35
0.30
0.25
0.20
2
4
6
8
10
Classes

Figure 6

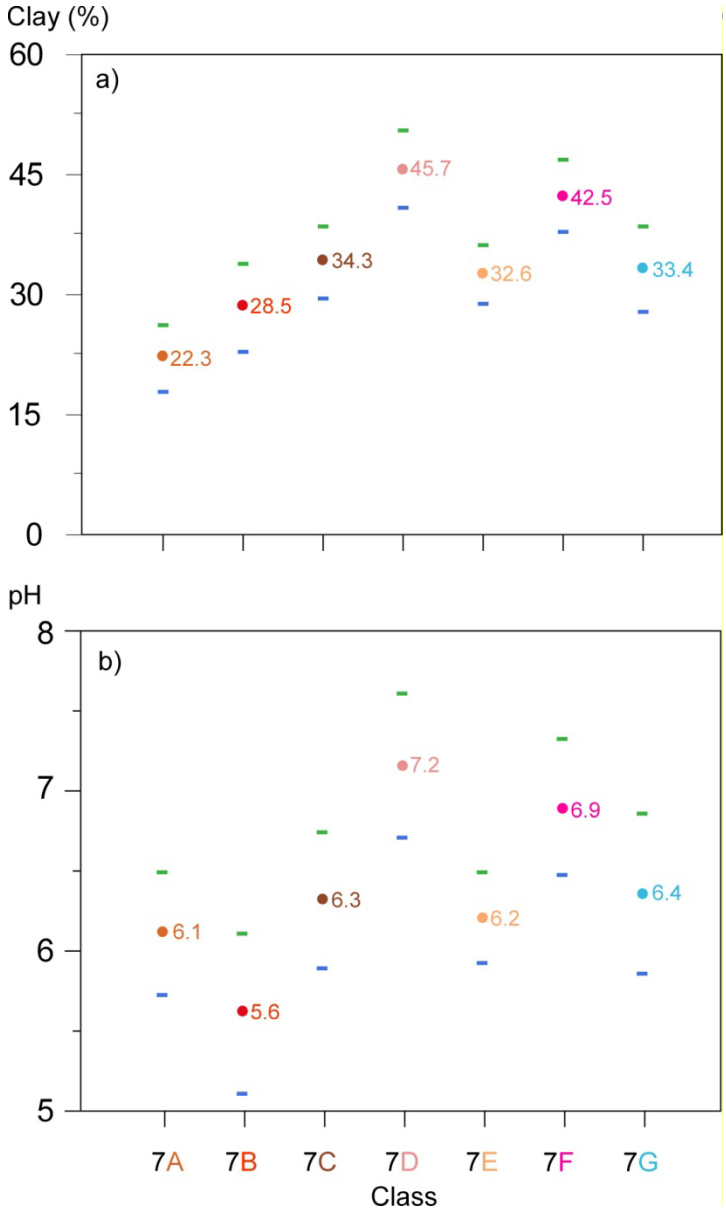
Clay (%)
60
45
30
15
0
a)
22.3
28.5
34.3
45.7
32.6
42.5
33.4
pH
8
7
6
5
b)
6.1
5.6
6.3
7.2
6.2
6.9
6.4
7A
7B
7C
7D
7E
7F
7G
Class

Figure 7