Relationships between particle size distribution and VNIR reflectance spectra are weaker for soils formed from bedrock compared to transported parent materials

B. G. Rawlins[a], S. J. Kemp[a], A. E. Milodowski[a]

[a]*British Geological Survey, Keyworth, Nottingham NG12 5GG, UK*

Running heading: *Particle size distribution and VNIR reflectance spectra*

Correspondence: B. G. Rawlins. E-mail: bgr@bgs.ac.uk

ABSTRACT

The cost of determining particle size distribution (psd) of the soil can be significantly reduced by using statistical relationships between visible and near infra red diffuse reflectance spectra (VNIR-DRS) and the proportions of the three size fractions (sand, silt and clay). The spectra contain information on the quantities of soil minerals which occur in these fractions. Statistical models for estimating psd based on a set of soil samples from common parent materials (PM) – with similar mineralogy – may provide more accurate predictions than more comprehensive, global models. The aim of this paper is to compare the performance of statistical models for the prediction of psd from VNIR-DRS for soils with differing types of parent material; specifically soils derived directly from bedrock (coal-bearing and mudstone-bearing strata) or from transported parent materials (glacial till, glacio-lacustrine deposits and alluvium) across eastern England. We assessed the accuracy of psd predictions using partial least squares regression (PLSR) models between two additive log ratios of the three size fractions and VNIR-DRS. We also formed a global PLSR model from all five soil groups. We used mean residual prediction deviation (RPD) from repeated (n=100) cross-validation to compare the performance of the models because it accounts for the magnitude of variation in the sample data. The most accurate models for the clay (RPD range 1.82–2.33) and sand fractions (RPD range 1.71–1.94) were for soils developed over the transported PM; the models for soils developed over bedrock were substantially poorer (clay RPD range 1.33–1.68; sand RPD range 1.34–1.39). The RPD values for the silt fraction models were smaller, but the same distinction between transported (better; RPD range 1.4–1.88) and bedrock derived soils (poorer; RPD range 1.15–1.25) was observed. The global model had intermediate RPD values for the three size fractions (clay=1.75, silt=1.76 and sand=1.74). Of the five groups, the soils developed from glacio-lacustrine deposits had the largest mean sand size fraction (58%), but also the most accurate models for estimation of clay and sand size fractions. Due to sedimentary transport and deposition, the mineralogy of the soils developed from Quaternary

substrates may be more homogeneous than the bedrock-derived soils, which may in part account for the more accurate models developed for the former. To date we do not have sufficient evidence to demonstrate this unequivocally.

1. Introduction

The ability of soil scientists to map the spatial variation of particle size distribution (psd) accurately at fine scales is important because psd contributes to the soil's hydraulic behaviour and water storage, its handling characteristics under tillage and its susceptibility to erosion. When psd is measured in the laboratory the results are often expressed as the proportions of three grain size-fractions (e.g. clay: $<2\mu$m, silt: 2–63μm and sand 63–2000μm) which sum to 100%. This is an example of compositional data which has constraints for certain statistical analyses (Aitchison, 1986). A large proportion of the spatial variation in psd typically occurs at scales between around 20 and 200 metres (McBratney and Pringle, 1999), so many samples and costly laboratory measurements would be required to map psd accurately using conventional methods. Scientists have shown that remote sensors – ground-based or airborne – can provide effective covariates to aid mapping of soil psd fractions including gamma radiometry (Taylor et al., 2002), geophysical measurements of electrical conductivity (Robinson et al., 2008) and near infra red reflectance spectra (Selige et al., 2006).

The visible and near infra-red (VNIR) diffuse reflectance spectrum (DRS) of a soil sample includes information on the quantities of the mineral phases it contains. It is assumed that each mineral phase or mineral coatings – iron-oxide coatings on silica (Scheidegger et al., 1993) or clay minerals (Caroll, 1958) – occur predominantly in one of the size fractions. The proportion of one size fraction – or all three fractions – in a set of soil samples are estimated from the VNIR spectra using multivariate statistical models by fitting them to laboratory measurements of psd. These multivariate models can then be used to estimate psd for other soils from the local area over which the original samples were collected. Four studies have been published where such models have been successful in predicting the proportions of particles in all three size-fractions (Chang et al., 2001; Shepherd and Walsh, 2002; Cozzolino and Moron, 2003; Sorensen and Dalsgaard, 2005).

By applying laboratory-based VNIR-DRS to a range of samples from the USA,

Europe, Africa and Asia, Brown et al. (2006) developed a global model for estimation of a variety of soil properties including percentage clay fraction using VNIR-DRS. The authors improved their estimates of percentage clay by including measurements of the sand size-fraction based on sieving. To date, no published studies have compared the performance of local and global statistical models for the estimation of all three soil texture fractions based on VNIR-DRS. Soil scientists need to know whether prediction accuracies can be substantially improved if statistical models are based on a smaller subset of soil samples when compared to a regional or global dataset (Sankey et al., 2008).

Remotely-sensed reflectance spectra have also been used to aid mapping of topsoil texture fractions at fine spatial resolutions (2 to 5 m) at farm scales (Barnes and Baker, 2000) and over small regions using airborne sensors (Selige et al., 2006; Lagacherie et al., 2008; Gomez et al., 2008). The availability of satellite-based hyperspectral data at fine spatial (30-m pixel sizes) and spectral (10-nm) resolutions (*e.g.* www.enmap.org) could provide landscape-scale covariates to substantially improve our ability to map topsoil psd when combined with ground-based measurements, in areas where topsoil is sufficiently exposed. Soil scientists need to know where in the landscape the relationships between soil psd and VNIR spectra are likely to be weak or strong – and the reasons for this – to assess the likely benefits of hyperspectral remote sensing to aid mapping psd.

The prediction of psd fractions from VNIR-DRS is likely to be more accurate if statistical models are developed and applied to groups of soils with similar mineralogy, and therefore, VNIR spectra. Of the five soil forming factors, much of the variation in soil mineralogy – and also VNIR reflectance – is likely to be explained by parent material (PM) type (Rawlins et al., 2003). This is particularly the case in areas where Quaternary substrates are the dominant PM type, such as across large parts of northern Europe, where recently formed soils have strong associations with their PM. The Quaternary parent materials comprise a range of transported materials deposited by

glaciers (till), rivers (alluvium) and the wind (aeolian deposits). Where Quaternary materials are thin or absent, soils develop directly from bedrock. It might be beneficial to develop statistical models to predict soil texture fractions using VNIR-DRS, based on their PM type.

The aim of this paper is to compare the performance of statistical models for the prediction of psd from VNIR-DRS for soils with differing types of PM; specifically soils derived directly from bedrock or from transported PM. We present statistical (partial least squares regression; PLSR) models used to estimate additive log ratios (Aitchison, 1986) of two texture components (clay:sand and silt:sand) using VNIR-DRS for groups of soils developed from five PM types in part of agricultural eastern England. We also establish a single statistical model for all the soils from the five groups. Two of the groups of PM were sedimentary bedrock; the other three represent a range of Quaternary (transported) PM types. We compare the VNIR wavelengths which are significant predictors for the soil texture fractions in the local and global calibration models and compare their prediction accuracies using independent cross validation after back transformation to the three size fractions. We seek plausible explanations to account for the differences in the accuracy of the statistical models in predicting particle size fraction for soils over the different types of PM.

2. Methods

2.1 Study region and soil sampling

The study region is the area of eastern England shown in Figure 1; the spatial distribution of the soil sampling locations are highlighted. Bedrock in the region ranges in age from Carboniferous to Cretaceous comprising coals, limestones, sandstone, siltstone, mudstone, chalk, marls and ironstones. There are a range of superficial deposits including glacial till, river and marine alluvium and a large region of lacustrine (lake) deposits formed by glacial meltwaters which predominantly give rise to Fluvisols. Soils developed from the two other parent material types are predominantly Cambisols and

Gleysols (IUSS Working Group WRB, 2006). Arable agriculture accounts for more than 90% of land use over the area from which samples were collected.

Soil sampling was undertaken across the region at a density of 1 sample per 2 square kilometres in the summers of 1994, 1995 and 1996 as part of a national-scale geochemical survey (Johnson et al., 2005). Sampling sites were chosen from alternate kilometre squares of the British National Grid by simple random selection within each square, subject to the avoidance of roads, tracks, railways, urban land and other seriously disturbed ground. At each site, surface litter was removed and soil was sampled from to a depth of 15 cm using five holes at the corners and centre of a square with a side of length 20 m by a hand auger and combined to form a bulked sample. All samples of soil were dried and disaggregated. They were sieved to pass 2 mm, coned and quartered.

We selected only those soil sampling sites (n=738) over five dominant parent material (PM) types (see Figure 1). We did this by assigning to each sampling location a PM code (based on combinations of solid or superficial geology). These PM codes were derived from digital versions of the 1:50 000 maps of bedrock geology and superficial deposits of England, part of DigMap GB of the British Geological Survey (2006). The number of soil sampling sites in each of the PM groups was as follows. For soils developed from two different types of bedrock parent material where there was little or no superficial material above the bedrock (coal-bearing strata; CM n=175 and mudstone-bearing strata; MDST n=47). For soils collected over PM types developed over thick, superficial deposits: alluvium (both marine and fluvial; ALV n=230), glacial till (TILL n=186) and lacustrine deposits (LDE n=100). The mineralogical composition – based on X-ray diffraction (XRD) analysis after removal of organic matter – for the different size fractions of selected soil samples over two of the parent material types (CM and LDE) are presented in Table 1. It is noteworthy that the soil developed over the lacustrine deposits has around twice as much kaolinite (33.6%) in the clay size fraction than the soil over the Coal Measures (16.8%), and that there is chlorite in the clay

(13%) and silt (7.1%) fractions of the Coal Measures soil, but this was not detected in the soil over the lacustrine deposits.

2.2 Measurement of diffuse reflectance spectra and redness index

Sub-samples of each soil were scanned in the visible-near infrared region (350–2500 nm) using an ASD (Analytical Spectral Devices, Boulder, CO) Agri-Spec NIR Spectrometer. In contrast to the sub-samples which were analysed to determine their psd (see below), organic matter (OM) was not removed from the sub-samples used for spectral analysis. The presence of OM – both as particulate carbon and coatings on mineral surfaces – will influence the VNIR spectra due to the occurrence of organic-related adsorption features. In some cases, the wavelengths of these adsorption features may coincide with adsorption features due to minerals in the texture fractions. This would lead to smaller regression coefficients at these wavelengths in statistical models formed between the spectra and the texture fractions. In our study, however, by not removing OM from the soil samples, the main benefit of VNIR-DRS – the rapid and cost-effective processing of samples – is preserved. In the wider context, remote sensing of soil in the VNIR region will always include adsorption features of OM in their spectra, so for its successful application, any interference caused by overlapping adsorption features must be overcome.

A 20-g subsample from each original soil sample was placed in a holder with a quartz window for scanning. Soils were illuminated and scanned from below using the spectrometer connected to an ASD muglight with an internal tungsten–quartz–halogen light source and a 12 mm spot size. Data were collected every 1 nm and every spectrum was an average of 25 readings. Each sample was scanned twice; the second scan was made after rotating the sample in its holder through 90° whilst placed on the muglight. During scanning, a Spectralon 99% reflectance panel was used to optimize and white-reference the spectrometer after scanning every set of ten samples. We checked that both sides of the Spectralon panel gave consistent baselines. Before further statistical

analysis, we obtained an average of two spectra for each sample.

In addition to the spectra, we computed the soil redness index (RI) for each sample as a potential predictor for particle size fractions. In many soil types, soil redness is dominated by the occurrence of iron-oxide minerals which form coatings on clay minerals (Carroll, 1958); so in certain soil types, RI may be strongly correlated with the proportions of the clay size fraction. The RI was computed as (Mathieu et al., 1998):

$$RI = \frac{R^2}{(B \times G^3)} \tag{1}$$

where R, G, and B represent the reflectance at the wavelength of red, green, and blue bands (700, 546, and 436 nm, respectively) recorded by the ASD spectrometer.

2.3 Particle-size analysis

The protocol for the particle size analysis was recently described in detail by Rawlins et al. (2009); here we provide a summary of its important features. Organic matter was removed from all sub-samples prior to psd determination by adding a combination of hydrogen peroxide and water to each sample and heating the mixture. Calgon solution was added to the samples to disperse them before analysis by laser granulometry. An 8 μm threshold was used for the upper limit of the clay-sized fraction instead of the conventional 2 μm; this corrects for differences in measurements by sedimentation and laser-based methods for non-spherical particles (Konert and Vandenberghe, 1997). Duplicated analyses (n=86) showed that the precision of the method was good; standard deviations were 2.1% for sand and clay, and 1.2% for silt.

The psd for each of the samples in the study (n=738) is shown in Figure 2. There are substantial differences in the mean sand and clay compositions for each of the five groups; the mean clay content varies from 23 to 38% and the mean sand content from 21 to 58% (Figure 2 and Table 2). The variation of psd within each of the five groups (standard deviations shown in Table 2) are quite similar; the TILL and CM group are somewhat less variable and, as might be expected, the ALV group (alluvial soil parent

material) has the most variable psd.

2.4 Statistical analyses

2.4.1 Additive log-ratio transformation

The compositional constraints on data with distributions that are curtailed at the limits of 0 and 1 (or 0 and 100%) induces correlations among the components, in this case the particle size fractions. Linear regression models are not limited in this way, nor are their predictions constrained to sum to 1.

Aitchison (1986) proposed a way out of this difficulty using the additive log ratio (alr). Suppose we have V variables, each with values lying between 0 and 1 and summing to 1, and that we choose $V-1$ with values for each unit $z_1, z_2, \ldots, z_{V-1}$. We can transform these to

$$q_i \;=\; \ln\left(\frac{z_i}{z_V}\right) \quad \text{for all } i = 1, 2, \ldots, V-1 \;, \tag{2}$$

where z_V is the value of the remaining Vth variable. The resulting values over all units have by definition a logistic normal distribution. This is the additive log ratio (alr) transform, and it allows us to analyse our compositional data as any other multivariate normal data.

After estimating new values $\widehat{q}_i,\ i = 1, 2, \ldots, V-1$, we want to return them to their original scale of composition, and we do so by the inverse transform, the additive generalized logistic transformation:

$$\begin{aligned} \widehat{z}_i &= \frac{\exp\left(\widehat{q}_i\right)}{1 + \sum_{j=1}^{V-1} \exp\left(\widehat{q}_i\right)} \quad \text{for all } i = 1, 2, \ldots, V-1 \\ \text{and} \quad \widehat{z}_V &= \frac{1}{1 + \sum_{j=1}^{V-1} \exp\left(\widehat{q}_j\right)} \;. \end{aligned} \tag{3}$$

As Aitchison showed, the results of this back-transformation are the same whichever variable we select as z_V.

In this study we formed partial least squares regression (PLSR) models for two alr-transformed variates and back-transformed these values to three size fractions. We note that the centre of the backtransformed distribution is equivalent to the median

on the original distribution (Pawlowsky-Glahn and Olea, 2004) not the mean, and so the backtransformed values include some bias. We then assessed the accuracy of the estimates using independent cross-validation.

2.4.2 Partial least squares regression and cross-validation

PLSR is a chemometric technique which is well-suited to multicollinear predictor variables, such as reflectance measurements in infra red spectroscopy. The predictive regression model can be represented by:

$$Y = b_0 + b_1 X_1 + b_k X_k + \epsilon \tag{4}$$

where the observed response values (Y; in this case the alr ratios) are approximated by a linear combination of the values of the spectral intensities (X), coefficients (b) - referred to as b-coefficients, and an error term (ϵ).

To determine the significant wavelengths for prediction of the alr ratios of the texture fractions, we used both the Variable Importance in the Projection (VIP) (Chong and Jun, 2005) and the PLS regression coefficients (b-coefficients; Haaland and Thomas, 1988). For an observed variable y, the VIP was calculated by:

$$VIP_k(a) \;=\; K \sum_a w_{ak}^2 \left(\frac{SSY_a}{SSY_t} \right) \tag{5}$$

where $\mathrm{VIP}_k(a)$ gives the importance of the kth predictor variable based on a model with a factors, w_{ak} is the corresponding loading weight of the kth variable in the ath PLSR factor, SSY_a is the explained sum of squares of y by a PLSR model with a factors, SSY_t is the total sum of squares of y, and K is the total number of predictor variables. The wavelength is considered important if the values of both the b-coefficients and VIP are sufficiently large. In this study, thresholds for VIP were set to 1 (Chong and Jun, 2005) and the standard deviation of the b-coefficients was applied as their threshold.

We used the *pls* package (Mevik and Wehrens, 2007) in the R environment (R Core Development Team, 2010) to form PLSR models based on the orthogonal scores algorithm. After taking alr ratios (Equation 2) using the *compositions* package (van den Boogaart et al., 2008) we fitted models to the two alr ratios for the soils from each of the five PM groups, and also to all the samples; twelve models in total. We investigated whether spectral pre-processing (first and second derivatives and Savitsky-Golay smoothing) improved model performance. In each case the untransformed reflectance data gave the best model performance so we used the original data in fitting all models. We used a truncated range (450-2450 nm) of wavelengths and the RI as predictors. We used cross validation to select the optimum number of components from which to form the models and also calculated the coefficient of determination (R^2) to assess model performance. Prior to forming each model, 10% of the samples were selected randomly and were not used in model fitting. These samples were then used to assess the model performance by forming predictions, backtransforming the alr components to proportions of the compositions (Equation 3) and calculating the root-mean-squared-error of cross validation (RMSE-CV) for between 1 and 12 model components. The RMSE-CV is calculated as:

$$\text{RMSE-CV} \;=\; \sqrt{\frac{1}{n_{\text{V}}}\sum_{i=1}^{n_{\text{V}}}\left(\widehat{z}_i - z_i\right)^2}\,, \tag{6}$$

where z_i is the measured proportion of a particle fraction and $\widehat{z}_i$ is its predicted value. We selected the optimum number of components for each PLSR model based on minimisation of the RMSE-CV.

To assess the accuracy of the selected models more thoroughly, we undertook repeated (n=100) cross-validation by randomly selecting 10% of the samples from each group and calculating the mean of the RMSE-CV and the mean of the residual prediction deviation (RPD); the ratio of the standard deviation of the validation sample set and the standard error of prediction (Equation 6). This statistic provides a useful

indication of the quality of the model because it accounts for the variation in the size fractions in the validation dataset.

3. Results and their interpretation

3.1 Regression models and psd estimation accuracy

Summary data for the PLSR models fitted to the two alr size-fraction ratios for each of the soils grouped by PM and all soils grouped together are presented in Table 3. The position of those wavelengths in the PLSR models identified as having significant predictive power – based on large b-coefficients and VIP scores – are presented in Figure 3. For both size-fraction ratios, the global group and alluvial sediment group have the largest number (n=10 or 11) of orthogonal model components probably because they represent a greater diversity of soil types than the other individual soil groups. It is notable that RI was only a significant predictor for the CM soil group – this may partly be related to the colour associated with the range and age of iron-bearing mineral phases in the coal-bearing strata. The mineral pyrite is abundant (Spears et al., 1999) in the bedrock from which these soils formed and over time this weathers to form a range of iron-bearing minerals (hematite and goethite; also present in the host rock) of varying age. Soil samples containing differing proportions of these minerals will have quite different redness features which may account for its significance as a predictor in this soil group. This may in part be due to different ageing of iron oxyhydroxides giving rise to differences in colour wavelengths (yellow, orange and red). For example, goethite reddens as it ages to hematite whilst lepidocrocite is dark brown or black and amorphous iron-oxide is between yellow and orange in colour. We cannot explain the significant wavelengths between 950 and 1050nm for the TILL and ALV models as these do not appear to relate to known absorption features in the near infra red spectrum.

The significant predictive wavelengths are consistent with colour in the visible light range (450-700nm). The H_2O adsorption bands at 750, 975, 1900–1950 and

2200nm are present as significant predictors in many of the PLSR models presumably associated with certain water-absorbing clay mineral phases. The H_2O adsorption band at 1400 nm is absent from all but the global clay:silt ratio model; this demonstrates the potential problem in forming global calibration models which may be based on predictive wavelengths that would not be justified based on a groups of local models. Significant wavelengths for other adsorption bands associated with certain clay mineral phases which are common in British soils include 2204–2211 nm (illite, kaolinite), 2340 nm (illite) and 2207 nm (smectite). Adsorption bands commonly associated with hematite (920 nm) and goethite (880 nm) are absent from all the models, although their effects through soil colour may be more significant in the visible wavelength range. There is substantially greater (n=58 wavelengths) overlap in the significant predictive wavelengths for the clay:silt size fraction ratio models (Figure 3a) for the five individual soil groups compared with the silt:sand ratio size fraction groups (n=7 wavelengths; Figure 3b). The overlapping wavelengths in the clay:silt size fraction models are dominated by wavelengths centred around the water absorption feature at 1900-1950nm (illite) and smectite (2004-2211nm); the absorption feature which is common to all models between 2407 and 2420nm may be related to adsorption features associated with carbonate minerals; chalk bedrock underlies the TILL soils in the north-east of the study region (Figure 1).

The results of repeated (n=100) 10% cross-validation for each of the models applied to the soil groups are presented in Table 2; the RMSE-CV and RPDs were calculated after back-transformation to the original three size-fractions. Overall model performance – based on RPD – is poorest for the silt size-fraction. This may be because this fraction shares a boundary with the two other size-fractions, whilst they each share only one. The intermediate, silt size-fraction is likely to comprise a larger proportion of uncommon minerals than the two other size-fractions. In terms of overall performance, the RPD values are larger (more accurate estimates) for the models relating to soils developed over the transported PM compared to those formed from bedrock. For the

clay size-fraction, the RPDs decline in the order: LDE > ALV > TILL > MDST > CM. With the exception of the last two groups which swap places in the order stated above, the same pattern also applies to the sand size-fraction. It is noteworthy that even though the LDE group has the largest mean sand-size content (58%; Table 2) it has the best overall model performance for the sand size-fraction. If, as is often the case, the sand fraction is dominated by quartz which has no absorption features in the NIR range (350-2500 nm), we might have expected model performance to be poor relative to the other groups. In the case of the silt size-fraction, soils over the transported PM again have the largest RPD values compared to those derived from bedrock: ALV > TILL > LDE > CM > MDST.

In each size-fraction, the RPD for the PLSR models developed for all soils (global) generally has an intermediate value; greater than the soils over bedrock but less than those over transported PM types. In the case of estimating clay and sand size-fractions for soil over the transported PM types, if we rely on a global PLSR model the average error of our predictions would be substantially larger than if we had developed models for each PM group (Table 2). In the case of the silt size-fraction, only the overall model performance (RPD= 1.88) for the alluvial soils is greater than that of the global model (RPD= 1.76), with particularly poor overall performance for the MDST and CM models.

4. Discussion

Previous research has highlighted the importance of PM when estimating cation exchange capacity of soil using VNIR-DRS across another part of eastern England (Savvides et al., 2010). Our analysis has shown that there are substantial differences in the performance of statistical models for prediction of particle size fractions based on VNIR-DRS for soils developed over different PM types across a large area of Eastern England. This highlights the importance of existing maps of soil PM or soil type to enhance the application of sensor-based covariates for producing digital soil maps.

We currently have no direct evidence to account for the observed differences in

VNIR-DRS model performance for the different soil groups. Here we give further consideration to three possible reasons. The first is that soils derived from transported (allochthonous) PM types are likely to be more mineralogically homogeneous than those derived from *in-situ* weathering of bedrock because the former have undergone sorting processes associated with transport and redeposition, whilst the latter have only been subject to weathering *in situ*. For example, we know from direct observation and mineralogical analyses that there is considerable variation in the lithologies (mudstones, siltstone, sandstone, coal-bearing strata) which comprise the coal-bearing strata (CM) in our study area.

A second explanation to account for the observed differences in model performance for the different soil groups concerns the relative abundance of iron-oxide coatings of soil clays which might lead to bias in the reflectance spectra; more or different types of iron-oxide coatings could diminish the clay mineral signatures and hinder VNIR-DRS model performance. We assume that total soil iron content – for which we have measurements for each soil sample from XRFS analysis (see Rawlins et al., 2009) is strongly correlated with iron-oxide concentrations. Soils developed over the coal-bearing strata contained the greatest median concentration of total iron (4.71%) and had the poorest overall model performance for size fraction prediction. However, the alluvial soils group was the next most enriched in total iron (median=3.91%) and VNIR-DRS model performance was reasonable, which confounds this theory if we assume that iron-oxides have similar associations with soil minerals in each of the soil groups (i.e. the proportion of iron-oxide coated clay mineral particles has a linear relationship with total iron content). The relationship between iron oxide coatings of minerals and VNIR-DRS warrants further investigation.

Thirdly, the presence of widely differing proportions of quartz – which has no spectral adsorption features in the VNIR range (350-2500 nm; Ferraro, 1982) – in the different size fractions would weaken the statistical relationships with VNIR-DR spectra. Although this limitation of size fraction estimation from VNIR-DRS is recognised,

it has not been widely referred to in the soil science literature. Our analyses show that soils with greater, average sand-size (quartz-dominated) fractions do not in general have weaker correlations between VNIR-DR spectra and each texture fraction; the group with the largest mean sand size fraction had amongst the best (largest) RPD values. Further research is required to provide an unequivocal, evidence-based explanation for the observed differences in the strength of relationships between psd and VNIR-DRS for soils developed over transported and bedrock-derived PM types.

Our results suggest that accuracy in mapping soil psd based on VNIR-DRS is likely to be substantially improved if local statistical models are developed compared to regional or global approaches. This is also likely to apply to the use of hyperspectral, remotely sensed data to map size fractions (Lagacherie et al., 2008) because the strength of the statistical relationships between the size fractions and spectral signatures will be similar.

Our previous research showed that the total concentration of five elements in the soil (Al, Fe, Ni, Ti and Zr) could be used to accurately estimate soil psd across the same study region with substantially smaller RMSE-CV compared to the local VNIR-DRS models (clay 4.9% versus a range from 6.1 to 8.8 %, respectively; sand 8.8% versus a range from 9.7 to 13.1% respectively). Although these estimates are more accurate the cost associated with acquisition of geochemical data makes it far more costly when compared to the spectral approach. The latter also has the advantage for the potential application of exhaustive, remotely sensed data to improve estimates at fine scales (< 10 m).

5. Conclusions

Our analyses have shown that there are substantially stronger relationships between psd and VNIR-DR spectra for topsoils developed from transported parent materials (three groups) compared to those developed directly from bedrock (two groups) at the regional scale. Based on RPD values from repeated cross-validation, the statistical models developed between the additive log ratios of the psd fractions and VNIR-DR

spectra for topsoils over the transported parent materials generally perform better than a global model developed for all five soil groups soils across the region. This has important implications for optimal strategies for mapping psd using field-based VNIR-DRS and the use of remotely sensed, hyperspectral data.

Acknowledgements

This paper is published with the permission of the Executive Director of the British Geological Survey (Natural Environment Research Council). We would like to thank: (i) staff from the British Geological Survey involved in collecting and preparing the soil samples, (ii) Andy Tye, Oliver Gould and Yinka Oshokoya who organised and undertook the spectral and psd analyses, and (iii) Doris Wagner and Ian Mounteney for the XRD analyses. We also thank an anonymous reviewer for helpful comments on the original manuscript.

References

Aitchison, J. 1986 *The Statistical Analysis of Compositional Data*. Chapman and Hall, London.

Barnes, E. M., Baker, M. G., 2000. Multispectral data for mapping soil texture: possibilities and limitations. Applied Engineering in Agriculture 16, 731–741.

British Geological Survey 2006. *Digital Geological Map of Great Britain 1:50 000 scale (DiGMapGB-50) data [CD-ROM] Version 3.14*. British Geological Survey, Keyworth, Nottingham.

Brown, D.J., Shepherd, K.D., Walsh, M.G., Dewayne Mays, M., Reinsch, T.G., 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma 132, 273–290.

Carroll, D., 1958. Role of clay minerals in the transportation of iron. Geochimica Et Cosmochimica Acta, 14, 1–16.

Chang, C.-W., Laird, D.A., Mausbach, M.J. and Hurburgh, C.R., Jr., 2001. Near-Infrared Reflectance Spectroscopy-Principal Components Regression Analyses of Soil Properties. Soil Science Society of America Journal 65, 480–490.

Chong, I.-G., Jun, C.-H., 2005. Performance of some variable selection methods when multicollinearity is present. Chemometrics and Intelligent Laboratory Systems 78, 103–112.

Cozzolino, D., Moron, A., 2003. The potential of near-infrared reflectance spectroscopy to analyse soil chemical and physical characteristics. Journal of Agricultural Science 140, 65–71.

Ferraro, J.R., 1982. The Sadtler infrared spectra handbook of minerals and clays. Heyden & Son Ltd. London. 440 pp.

Gomez, C., Lagacherie, P., Coulouma, G., 2008. Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements. Geoderma 148, 141–148.

Haaland, D.M., Thomas, E.V., 1988. Partial least-squares methods for spectral analyses.1. Relation to other quantitative calibration methods and the extraction of qualitative information. Analytical Chemistry 60, 1193–1202.

Hodgson, J.M., 1974. Soil Survey Field Handbook Describing and Sampling Soil Profiles, Soil Survey of England Wales, Harpenden, Hertfordshire (1974).

IUSS Working Group WRB, 2006. *World reference base for soil resources 2006. 2nd edition*. World Soil Resources Reports No. 103. FAO, Rome.

Johnson, C.C., Breward, N., Ander, E.L., Ault, L., 2005. G-BASE: Baseline geochemical mapping of Great Britain and Northern Ireland. Geochemistry: Exploration-Environment-Analysis, 5, 1–13.

453 Konert, M., Vandenberghe, J., 1997. Comparison of laser grain size analysis with
454 pipette and sieve analysis: a solution for the underestimation of the clay fraction.
455 Sedimentology 44, 523–535.

456 Lagacherie, P., Baret, F., Feret, J.B., Netto, J.M. and Robbez-Masson, J.M., 2008.
457 Estimation of soil clay and calcium carbonate using laboratory, field and airborne
458 hyperspectral measurements. Remote Sensing Of Environment 112, 825–835.

459 Mathieu, R., Pouget, M., Cervelle, B., Escadafal, R., 1998. Relationships between
460 satellite-based radiometric indices simulated using laboratory reflectance data
461 and typic soil color of an arid environment. Remote Sensing of Environment 66,
462 17–28.

463 McBratney, A.B. and Pringle, M.J., 1999. Estimating average and proportional vari-
464 ograms of soil properties and their potential use in precision agriculture. Precision
465 Agriculture 1, 219–236.

466 Mevik, B.H. and Wehrens, R., 2007. The pls Package: Principal Component and
467 Partial Least Squares Regression in R. Journal of Statistical Software 18, 1–24.

468 Pawlowsky-Glahn, V,. Olea, R. A., 2004. Geostatistical Analysis of Compositional
469 Data. Oxford University Press, Oxford.

470 R Development Core Team, 2010. R: A Language and Environment for Statis-
471 tical Computing., R Foundation for Statistical Computing, Vienna, Austria,
472 http://www.R-project.org.

473 Rawlins, B.G., Webster, R., Lawley, R., Tye, A.M., O'Hara, S.O., 2009 Estimating
474 particle-size fractions of soil dominated by silicate minerals from geochemistry.
475 European Journal of Soil Science 60, 116–126.

476 Rawlins, B.G., Webster, R., Lister, T.R., 2003. The influence of parent material on
477 top soil geochemistry in eastern England. Earth Surface Processes and Landforms
478 28, 1389–1409.

Robinson, D.A., Abdu, H., Jones, S.B., Seyfried, M., Lebron, I., Knight, R., 2008. Eco-Geophysical Imaging of Watershed-Scale Soil Patterns Links with Plant Community Spatial Patterns. Vadose Zone Journal 7, 1132–1138.

Sankey, J. B., Brown, D. J., Bernard, M. L., Lawrence, R. L., 2008. Comparing local vs. global visible and near-infrared (VisNIR) diffuse reflectance spectroscopy (DRS) calibrations for the prediction of soil clay, organic C and inorganic C. Geoderma, 148, 149–158.

Savvides, A. , Corstanje, R. , Baxter, S.J. , Rawlins, B.G. and Lark, R.M., 2010. The Relationship between Diffuse Spectral Reflectance of the Soil and Its Cation Exchange Capacity Is Scale-Dependent, Geoderma 154, 353–358.

Scheidegger, A., Borkovec, M., Sticher, H., 1993. Coating of silica sand with goethite: preparation and analytical idenification. Geoderma, 58, 43–65.

Selige, T., Bohner, J., Schmidhalter, U., 2006. High resolution topsoil mapping using hyperspectral image and field data in multivariate regression modeling procedures. Geoderma 136, 235–244.

Shepherd, K.D., Walsh, M.G., 2002. Development of reflectance spectral libraries for characterization of soil properties. Soil Science Society of America Journal 66, 988–998.

Sorensen, L. K., Dalsgaard, S., 2005. Determination of clay and other soil properties by near infrared spectroscopy. Soil Science Society of America Journal 69, 159–167.

Spears, D.A., Manzanares-Papayanopoulos, L. I., Booth, C.A., 1999. The distribution and origin of trace elements in a UK coal; the importance of pyrite. Fuel 78, 1671–1677.

Taylor, M.J., Smettem, K., Pracilio, G., Verboom, W., 2002. Relationships between

soil properties and high-resolution radiometrics, central eastern Wheatbelt, Western Australia. Exploration Geophysics 33, 95–102.

van den Boogaart, K. G., Tolosana, R., Bren, M., 2008. compositions: Compositional Data Analysis. R package version 1.01-1. http://www.stat.boogaart.de/compositions.

Figure captions

Figure 1 Parent material and soil sampling locations across the study region.

Figure 2 Particle-size distribution for groups of soils developed from five different parent material types across the study region. Figure 1 shows their spatial distribution. The partitions of the triangle and class names are those in the Field Handbook of the Soil Survey of England and Wales compiled by Hodgson (1974).

Figure 3 Wavelengths at which both variable importance in the projection (VIP) scores and regression (beta) coefficients are significant in partial least squares models of reflectance spectra for prediction of additive log ratios of texture fractions for soils grouped by parent material and for all soils: a) clay:silt, b) silt:sand. Bands which are significant in all five soil model groups are shown in grey.

519 Table 1 X-ray diffraction analysis of estimated mineralogic composition for three size
520 fraction separates for single selected topsoil samples developed over coal-bearing strata
521 (CM) bedrock and lacustrine deposits (LDE).

Size fraction	Sand		Silt		Clay	
Parent material	CM	LDE	CM	LDE	CM	LDE
proportion of total mass (%)	8	58	23	26	69	15
albite	na	3.2	3.9	5.4	nd	<0.5
522 anatase	na	nd	0.5	<0.5	0.7	<0.5
*kaolin	na	nd	7.5	9.1	16.8	33.6
K-feldspar	na	7.2	3.4	5.6	nd	<0.5
†mica	na	nd	27.5	13.7	58.2	57.7
quartz	na	89.6	49.9	66.1	11.3	8.2
chlorite	na	nd	7.1	nd	13	nd

523 * kaolin: undifferentiated kaolin group minerals possibly including kaolinite, halloysite
524 †mica: undifferentiated mica species, possibly including muscovite, biotite, illite and
525 illite/smectite
526 nd = not detected
527 na = not analysed
528

Table 2 Mean (%) and standard deviation (St.Dev) of three size fractions for groups of five soil samples and all samples (Global). Root mean square error of cross-validation (RMSE-CV: %) based on n=100 repeated random selection of 10% of samples based on PLSR models fitted to each group and mean residual prediction deviation (RPD).

	Clay				Silt				Sand			
	Mean	St. Dev.	RMSE-CV	RPD	Mean	St. Dev.	RMSE-CV	RPD	Mean	St. Dev.	RMSE-CV	RPD
Global	33	14	7.9	1.75	31	13	7.6	1.76	36	23	13.1	1.74
ALV	38	15	7.1	2.01	41	15	8.0	1.88	21	22	11.9	1.87
LDE	23	14	6.1	2.33	18	11	7.0	1.4	58	23	11.7	1.94
TILL	26	11	5.7	1.82	22	8	5.3	1.62	52	17	9.7	1.71
CM	38	11	8.8	1.33	32	7	5.8	1.25	29	16	11.7	1.39
MDST	33	13	7.0	1.68	34	10	7.8	1.15	32	17	11.5	1.34

Table 3 Features of partial least squares regression (PLSR) models formed between two alr ratio size fractions and VNIR-DRS for five sets of soil grouped by parent material type: a) number (n) of orthogonal PLSR components, b) coefficient of determination (adjusted R^2), c) number (n) of wavelengths (λ) in PLSR model which have VIP scores and beta coefficients greater than significance thresholds (see text).

	alr clay:silt			alr silt:sand		
Group (n)	n components	R^2	$n(\lambda)$	n components	R^2	$n(\lambda)$
Global (738)	11	0.64	523	11	0.60	217
ALV (230)	10	0.80	413	10	0.75	405
LDE (100)	8	0.86	411	8	0.77	225
TILL (186)	8	0.69	432	8	0.67	534
*CM (175)	7	0.50	304	7	0.51	245
MDST (47)	5	0.59	379	5	0.60	441

* includes redness index (RI) as a significant predictor – without the RI the maximum R^2 values of PLSR models between VNIR-DRS and the clay:silt and silt:sand fractions of the CM soils were 0.33 and 0.38, respectively.

Figure 1:

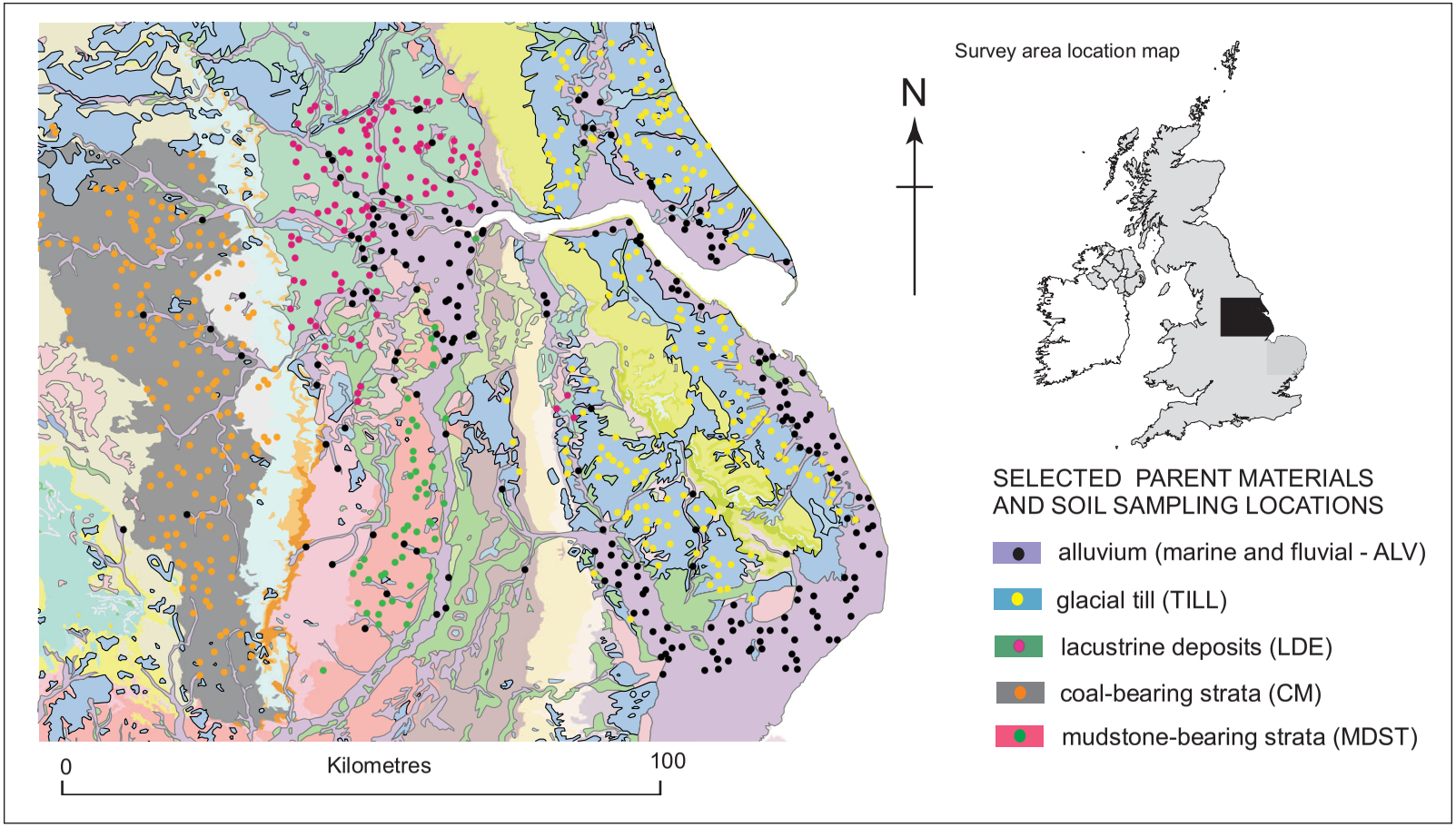

N
Survey area location map
SELECTED PARENT MATERIALS
AND SOIL SAMPLING LOCATIONS
alluvium (marine and fluvial - ALV)
glacial till (TILL)
lacustrine deposits (LDE)
coal-bearing strata (CM)
mudstone-bearing strata (MDST)
0
Kilometres
100

Figure 2:

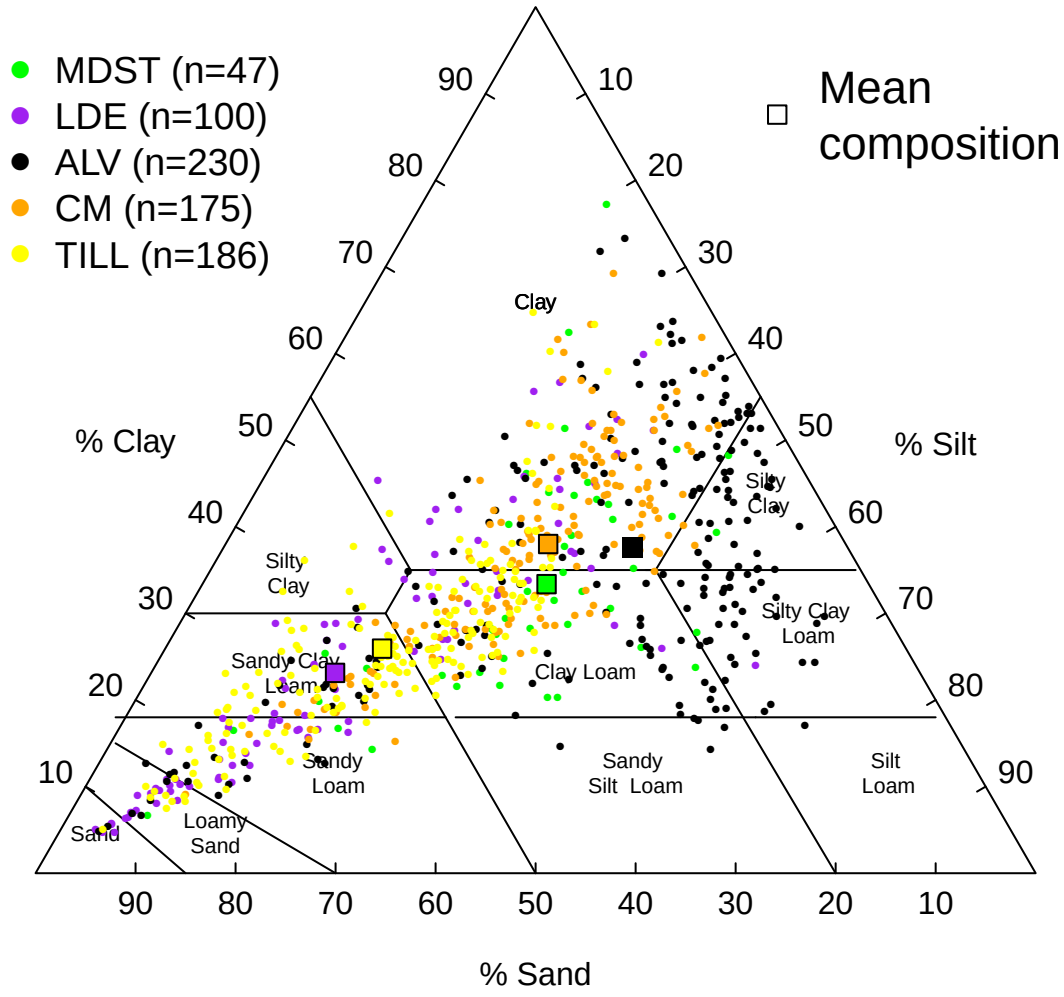
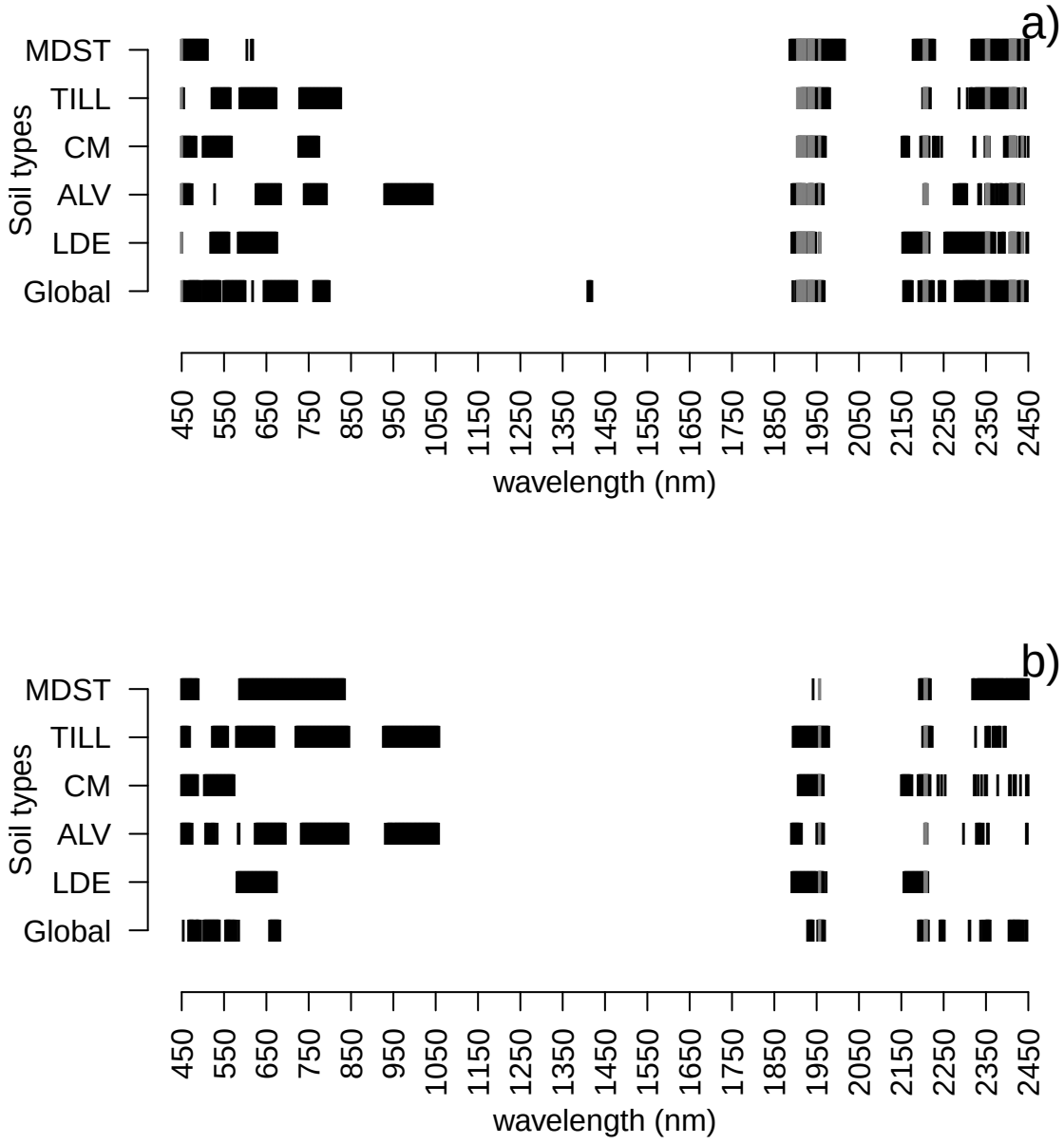

MDST (n=47)
LDE (n=100)
ALV (n=230)
CM (n=175)
TILL (n=186)
Mean composition
% Clay
% Silt
% Sand
Clay
Silty Clay
Silty Clay
Silty Clay Loam
Clay Loam
Sandy Clay Loam
Sandy Loam
Sandy Silt Loam
Silt Loam
Loamy Sand
Sand

Figure 3:


a)
MDST
TILL
CM
ALV
LDE
Global
Soil types
450
550
650
750
850
950
1050
1150
1250
1350
1450
1550
1650
1750
1850
1950
2050
2150
2250
2350
2450
wavelength (nm)
b)
MDST
TILL
CM
ALV
LDE
Global
Soil types
450
550
650
750
850
950
1050
1150
1250
1350
1450
1550
1650
1750
1850
1950
2050
2150
2250
2350
2450
wavelength (nm)