THIS PAPER HAS BEEN PUBLISHED IN: JOURNAL OF APPLIED SPECTROSCOPY 65(10) (2011) 1151-1160

CHEMOMETRIC STUDY ON THE FORENSIC DISCRIMINATION OF SOIL TYPES USING THEIR INFRARED SPECTRAL CHARACTERISTICS

Mark Baron(1), Jose Gonzalez-Rodriguez(1), Ruth Croxton(1), Rafael Gonzalez(2), Rebeca Jimenez(1)

(1) School of Natural and Applied Sciences, University of Lincoln, Brayford Pool, Lincoln, LN6 7TS, United Kingdom

(2) Departamento de Quimica Fisica y Termodinamica Aplicada, Universidad de Cordoba, Campus de Rabanales, Ed. Marie Curie, E-14071 Cordoba, Spain

Abstract

Soil has been utilized in criminal investigations for some time because of its prevalence and transferability. It is usually the physical characteristics that are studied, however the research carried out here aims to make use of the chemical profile of soil samples. The research we are presenting in this work used sieved (2mm) soil samples taken from the top soil layer (about 10cm) that were then analysed using mid infrared spectroscopy. The spectra obtained were pre-treated and then input into two chemometric classification tools: Nonlinear iterative partial least squares followed by linear discriminant analysis (NIPALS-LDA) and partial least squares discriminant analysis (PLS-DA). The models produced show that it is possible to discriminate between soil samples from different land use types and both approaches are comparable in performance. NIPALS-LDA performs much better than PLS-DA in classifying samples to location

.

Keywords - Soil Analysis, Forensic Science, Fourier Transform Infrared Spectroscopy, Nonlinear Iterative Partial Least Squares, Linear Discriminant Analysis, Partial Least Squares Discriminant Analysis

Introduction

The analysis of soil samples is of paramount importance to solve cases in which it is necessary to link the suspect to the crime scene. The use of soil for forensic purposes can be catalogued into two different areas: court and intelligence purposes. At present, a soil sample found on a suspect can be matched with samples found in the crime scene linking the individual to that particular setting. This can be easily done with the wide range of analytical techniques available and the fact that almost each soil sample is unique. Identification of soil for intelligence purposes is much more complicated and it is in this area where chemometrics can greatly help. Obtaining soil profiles that can be used to produce a map of a wide area can be very useful for the police when the crime is still developing and not at the court stage. The main basis for the comparison of sites to determine provenance is that soils vary from one place to another. This is also one of the major problems in the use of soil comparisons in legal cases, as this variation can occur both within a particular site and between sites, and the extent of this is as yet unknown.

The use of chemometrics for soil analysis is not new and has been widely used in environmental applications. Use of principal component analysis (PCA) of contaminated soil,[1, 2] partial least square regression (PLSR),[3] hierarchical cluster analysis (HCA)[4] and discriminant analysis[5] are some of the examples found in the literature in which chemometrics has been demonstrated to be a useful tool to study soil characteristics for environmental analysis. Examples on the use of chemometric tools in the forensic analysis of soil are less frequent and mainly concentrated on the use of PCA to identify and cluster the different soil types although Awiti et al. use a PLSR model for the classification of soil fertility using infrared spectroscopy.[6] Some of the work described in the literature obtains discrimination between soils using several physical, chemical and biological measures used as raw data for PCA analysis[7] or using a single type of analysis, which can be in the form of a spectrum[8] or using a compilation of chemical or physical variables obtained from the same technique such as isotopic content.[9] Combined visible and near-infrared spectroscopy have been used to classify plants using PCA combined with linear discriminant analysis (LDA). This approach is often needed as PCA reduces the data to a set of variables that are then appropriate for LDA while still retaining the chemical information needed for classification.[10]

Mid-infrared spectroscopy has been previously used for the analysis of soil samples in order to obtain information about the total organic composition of the soil,[11] organic carbon[12] or humic substances.[13] Forensic use of infrared spectroscopy (IR) for soil analysis has been suggested by Cox et al.[11] using KBr disks in combination with other techniques for the characterization of soils but no chemometric analysis was performed in order to establish soil type clustering. Elliott et al.[14] used a IR well plate reader combined with chemometrics and genetic algorithms to establish a temporal evolution and classification of soils coming from a mining area to establish whether remediation of a contaminated area had occurred and to what extent. Mid-IR spectroscopy presents all the advantages and qualities to be seriously considered in combination with

chemometric analysis. The spectral information obtained in a single soil spectrum contains information on both the organic and inorganic content and could help to classify different soil types. Modern infrared spectrometers are mobile enough to be used by rapid action units and, used in conjunction with chemometric tools, can be powerful enough to identify different soil type in an area. These are non-destructive and only require a few milligrams of sample.

A common problem identified for soil analysis is the lack of staff with the expertise and training to carry out reliable soil analysis. This problem indicates a need to find simpler methods that do not require such specialised experience. The aim of this paper is to explore the use of attenuated total reflection (ATR) spectroscopy in conjunction with chemometric tools in order to positively identify soil type from the area of Lincoln (UK) using a single measurement and a simple methodology. The novelty of this approach is that using infrared spectra with a simple chemometric analysis will create a model that enables prediction of a soil sample to at least land-use type and possibly land-use site. This information can be used for intelligence purposes when trying to locate burial sites, or identify an area in which a crime might have been perpetrated based on a sample of soil obtained from a suspect. This can provide the police with a proactive tool to gather evidence to solve an ongoing crime as well as a reactive method to obtain evidence to be used in court. The use of different tools of multivariate analysis can greatly help to reduce the spectral noise and increase clustering and discrimination. Another novel aspect of this work is the study on the use of ATR for soil analysis taking into account soil sample preparation and data pre-treatment. This is often not explained in the literature when ATR spectroscopy has been applied to soil analysis.

Methods and Materials

Soil samples were taken from 4 flowerbed, 4 woodland and 4 river bank sites in the Lincoln (UK) area. These are classed as different land-use types. Brief details of the sites and map locations are given in table 1. At each sampling site a transect was set-up using a tape measure and soil samples were taken to a depth of 10 cm along the transect at 50 cm intervals using a core sampler. Five samples were taken from each site and labelled a-e giving an overall total of 60 soil samples. The samples were air dried for 3 days, followed by removal of stones and vegetation, sieving (2 mm), grinding with mortar and pestle and finally sieving again (125 µm).

Samples were measured directly on a Golden Gate ATR accessory (Specac, Kent, UK) with a diamond internal reflection element (IRE) housed in a PerkinElmer Spectrum 100 FT-IR Spectrometer (PerkinElmer, Cambridge, UK) using PerkinElmer Spectrum v.6.1.0.0038 software for spectral manipulations. After recording the spectrum, the soil was removed from the ATR and the IRE was cleaned with a tissue moistened with an ethanol solution (80:20 ethanol-deionised water). Instrument settings used were 128 scans; 4 cm^{-1} resolution; range 4000-400 cm^{-1}. Regular background spectra were measured and the cleanliness of the lens was checked

between samples using the live spectra feature of the instrument. Soil samples were analysed in triplicate using a different portion of the sample for each replicate.

Raw spectra (.sp files) were ATR and baseline corrected (minimum value subtraction) using the routines provided within the Perkin Elmer Spectrum software. The .sp files were exported into Excel in ASCII format and either a full spectrum (400-4000 cm^{-1}) or a reduced spectrum consisting of 2 blocks (400-1850 cm^{-1} and 2400-4000 cm^{-1}) were directly imported into the multivariate analysis software package Tanagra[15] using the add-in feature in Excel. The final data matrix was either 179 x 3601 (full spectrum) or 179 x 3052 (reduced spectrum). The 179 spectra consisted of 60 flowerbed, 60 woodland and 59 river bank spectra including sample replicates. For classification, the replicates were averaged giving a matrix of size 60 x 3601 or 60 x 3052. The data set was divided into land-use type (flowerbeds (FW), woodland (W) and river bank (R)) and site within land-use type e.g. FW1, FW2 etc. This allowed two models to be created. Model 1 classifies soil samples to land-use type and here is a three class model whereas model 2 classifies to land use site and here is a 12 class model.

Two treatments were applied to the data. The first used factorial analysis (non-linear iterative partial least squares (NIPALS)) to reduce the data set before input into linear discriminant analysis (LDA). A feature selection tool known as Stepwise Discriminant (Stepdisc) analysis (forward search strategy, stopping rule: F=3.84) was used to select the most significant factors for LDA. The second treatment used partial least squares-discrimant analysis (PLS-DA). This combined technique directly accepts the sample spectra/wavenumber matrix avoiding the need for initial data reduction. It can automatically select the appropriate number of PLS components required to give the best prediction models.

Results and Discussion

Figure 1 shows the ATR spectra of examples of the different land-use soil samples. It can be seen that the general shapes of the spectra are similar although there are variations in terms of peaks present and absent and also variations in relative magnitudes of different spectral regions. Soil spectra tend to be dominated by the mineral component (mainly silicate) which is characterized by the strong absorption of Si-O bonds centred around 1050 cm^{-1}.[16] This region may also show polysaccharide bands in organic rich soils. This band is clearly evident in the 3 land-use type soil spectra shown in figure 1. Other bands, particularly in the X-H stretching region (2800-3750 cm^{-1}) may be indicative of the presence of organic material with other characteristic bands also being present in the region 1300-1900 cm^{-1}. The broad band centred around 3400 cm^{-1} is usually attributed to O-H and N-H stretching of various functional groups. The two sharp bands in the 2850-2950 cm^{-1} region are due to aliphatic C-H stretching which are more prominent in the woodland sample. The two sharp bands at 3623 and 3700 cm^{-1} are due to inner surface hydroxyl groups of clay minerals.[17] These bands are not always seen[16] in soil and extracted soil organic material but were considered as important features for differentiation of soil from mine

sites[14] and featured in the PC1 loadings spectra used for classification of clay minerals.[17] Here they are more prominent in the river bank soil and absent in the flowerbed soil. The bands in the region 1600-1700 cm^{-1} are normally attributed to several functional groups including aromatic C=C and C=O. The variation across the full spectrum suggests the potential for building multivariate classification models of land use type and possibly location.

The appropriateness of applying a particular data pre-treatment method needs to be considered carefully. A number of different methods exist and one needs to be aware of the issues relating to the measurements in selecting the most appropriate methods. In near-infrared spectroscopy of solid samples techniques such as multiplicative scatter correction (MSC) and standard normal variate (SNV) are often employed to correct for variation in path length and particle size. With ATR the path length is only dependent on the wavelength and the refractive index of the sample and only extends several micrometers into the sample (approximately 1.2 µm at 1000 cm^{-1} for a diamond IRE assuming 45^{o} internal reflection angle). This therefore means that the application of methods such as MSC and SNV are not necessary and, if used, are likely to distort the data by introducing artefacts. The main source of non-chemical variation in spectra seen with ATR is due to the contact between the sample and diamond ATR element; however, with the Specac ATR accessory a constant pressure can be applied thereby minimising contact variation between samples. Although we have applied an ATR correction here, it merely corrects for the variation in penetration depth observed with wavelength. It serves simply to remove this spectral distortion from all spectra in a constant way and to amplify the bands at short wavelength. Some baseline variation was observed between spectra in the central spectral region and so a simple minimum value (single-point) baseline correction was applied to remove offset differences.

NIPALS produces factors (latent variables) similar to those of principal component analysis (PCA) (table 2) but with a much faster processing time. As with PCA, NIPALS can be used for exploratory data analysis looking for hidden data structures within a dataset. It can also be used as a method of data reduction prior to the use of supervised learning tools such as linear discriminant analysis (LDA). This approach is commonly used with PCA[17,18] and offers the advantages of significant data reduction allowing the use of tools such as LDA (not possible with the initial dataset) and providing orthogonal variables which removes problems due to variable co-linearity observed in spectral data. Table 2 summarizes the variance explained by the first 10 NIPALS factors for the 400-4000 cm^{-1} data set with soil sample replicate spectra and with average spectra. 77.15% and 79.58% of the spectral variance is captured by factors 1 and 2 with 96.5 % and 97.35 % captured by the first 7 factors and so any chemical differences detected in the spectra that relate to soil classification would be expected to be modeled from these factors. The reduced data set shows that 99.35% of the variance is captured by the first 7 factors with factors 1 and 2 accounting for 83.35%. This shows there is little difference between factorial analysis of the three variants of the dataset used in this study.

The score plot for factors 1 and 2 (figure 2) shows good separation of the 3 land-use groups of spectra. The flowerbed samples are more tightly grouped with woodland and river bank being more dispersed. The boundary between the woodland and flowerbed groups is more defined than that between the flowerbed and river groups suggesting a greater potential for misclassification between these two groups. Considerable overlap between the 3 groups is seen along both factors. Within the woodland and river bank groups there is also significant separation of sites along both factors with factor 1 separating W1 from the other W sites and R2 and R3 from sites R1 and R4. Factor 2 separates W and R sites although R3 is placed with W sites along this factor. Closer inspection of the flowerbed group also reveals sub-groups relating to samples from the same site. If the land use groups are processed separately the grouping of locations is more noticeable for all 3 types. Figure 3 illustrates this for the river bank data. R3 and R2 are well separated from each other and the other two sites, R1 and R4 as observed in figure 2. R1 and R4 show some dispersion relating to location but the boundary region shows significant mixing of the two sites. Land-use differences (table 1) may be responsible for the greater dispersion observed for these sites because of the different types of water courses found in the Lincoln region. This pattern indicates that variance in the data can be related to land-use type and site within a land-use type group. Figure 3 also shows that replicates from the same soil sample are often closely matched relative to other samples from the same site although this is not always the case showing that intra-sample variation is generally less than variation between different soil samples from the same site. An average of the triplicate spectra can therefore be used as representative of single soil samples.

Variation between sample replicates is due to inhomogeneity within the sample both in terms of chemical composition and particle size. Well mixed samples with relatively large particle sizes can give rise to poor repeatability due to the very small sample area of the diamond IRE, the short penetration depth of the evanescent field and the application of pressure to the sample needed to bring about contact between sample and the lens. The procedure used here ensures that a finely divided representative sample is prepared.

Variation between replicates can be assessed using the precision index (PI) defined as follows:

$$PI = 100 \frac{\sum_{N} \frac{RMSD_i}{\mu_i}}{N} \qquad (1)$$

where N is the number of wavebands, *RMSD* is the root-mean-squared deviation across the replicates for band i and μ is the mean absorbance in the band.[19]

PI provides a single measure that allows replicate spectra to be compared against a threshold for acceptance. A threshold of 3% has been applied for acceptance of reflectance spectra from soil samples.[19] PI values as low as this were not achieved with the spectra reported here due mainly

to the low signal-to-noise ratio seen above about 1600 cm^{-1} leading to high values of *RMSD* relative to the mean.

A measure of precision variation within the spectral range between a set of replicate spectra can be seen using the coefficient of variation for the root mean square deviation of the replicates (CV(*RMSD*)) defined as follows:

$$CV(RMSD_i) = 100\frac{RMSD_i}{\mu_i} \qquad (2)$$

where the terms are as defined for equation (1). *CV*(*RMSD*) spectra can be plotted that indicate how the precision between replicates varies across the spectral range. The *CV*(*RMSD*) spectrum for a set of triplicate flowerbed spectra is shown in figure 4. Variation between spectra is less than 5% for the fingerprint region but then increases with a significant change in the spectrum between 1900-2400 cm^{-1}. The region centred around 2000 cm^{-1} is particularly noisy because of the absorption by diamond in this region. A threshold of 5% for the fingerprint region only was therefore used for retention of triplicates and the subsequent use of the average value in the dataset.

The use of replicates distorts the results of cross validation methods as these become split between calibration and test data sets giving over optimistic prediction scores. Representative average spectra were therefore used for the classification models. The score plot for NIPALS factor 1 and 2 for average spectra is shown in figure 4. These two factors account for 92% of the variance.

The PCA loadings spectra (figure 5) show that the variance captured by all 5 components can be mapped onto spectral bands in the ATR spectra. PC1 shows a number of the characteristic bands described earlier including the SiO-H hydroxyl bands at 3700 and 3650 cm^{-1} seen with clay minerals.[17] The region 1850-2400 cm^{-1} is dominated by noise in the FTIR spectra which appears in all the loadings spectra. Although this is generally a region of little interest in mid-IR spectroscopy, the PC1 loadings spectrum shows a feature in this region centred around 1900 cm^{-1}. This region also contributes significantly to the variable importance in the projection (VIP) measure in Partial Least Squares Discriminant Analysis (PLS-DA) and is discussed later.

A NIPALS-LDA classification model was generated to classify samples according to land use type (model 1). The number of NIPALS factors used as input was optimized by evaluating the prediction performance of models with a re-sampling cross-validation method in which the dataset is divided into 10 groups of equal number (10 groups of 6 spectra). The classification model is created on 9 groups and then tested on the tenth group which is repeated until all 10 groups have been tested. This generates a test set of 60. Cross-validation provides a more realistic prediction error rate for use in optimization than other methods such as resubstitution and is suited for evaluating models created on small data sets where separate calibration and

test sets are not feasible. Figure 6 shows that prediction error rate decreases as the number of factors used in the model is increased up to 7 factors. The error decrease is not uniform and shows that factors 1, 2, 5 and 7 have dramatic greater effect in decreasing prediction error rate. Model 1 with only factors 1 and 2 gives a prediction error rate of 25% which is reduced to 16.7% when factor 5 is added. A stepwise discriminant (Stepdisc) analysis (forward search strategy, stopping rule: F=3.84) selected factors 1-7 as being the most significant for classification by linear discriminant analysis. When all are used, the model gives a prediction error rate of 8.3%. As expected from figure 5 factors 8-10 are not selected as being significant for the model. Table 3 shows the contingency table for the 7 factor NIPALS-LDA model 1. This data can be transformed into parameters that characterize the model of recall, precision and accuracy. Recall is the ratio of true positives to total true positives and false negatives. Precision is the ratio of true positives to the total classified as belonging to that group. Accuracy is the ratio of the sum of the true positives and negatives to the total number of examples. The data from table 3 is shown in this form in table 4. The model appears to be slightly better for woodland samples giving a recall and precision of 95%. This is explained by the greater overlap between river and flowerbed samples as shown in figure 2 however given the size of the sample there is little difference between the performance of the model for the 3 groups.

A second NIPALS-LDA model was created classifying to land-use type site (model 2). Stepdisc selected factors 1-6 as being the best factors for discrimination which is in agreement with figure 6. The contingency table (table 5) for the model using these factors shows an acceptable prediction accuracy for this number of groups given the overlap between groups seen in the score plots (figures 2 and 3).

An advantage of using PLS-DA is that the spectral variables can be used as direct input and so model output can be directly related to spectra. When applied to the 400-4000 cm^{-1} data set a model was created that selected 5 PLS components (stopping rule: redundancy in Y(Rd.Y) =0.025) giving a prediction error rate of 10%. Table 4 compares the performance of the model with the 7 factor NIPALS-LDA model. Overall performance is similar but they provide slightly different results. Flowerbed examples have the lowest recall in both models but 100% precision in the PLS-DA model. Classification of the river examples has 100% recall but the lowest precision. As with the NIPALS-LDA model1 these features can be explained by the greater overlap between flowerbed and river examples seen in the score plots.

PLS-DA model 2 gives a prediction error rate of 31.7 % (table 6) obtained using 14 PLS factors (stopping rule: Rd.Y=0.025). A 5 factor model gives a high prediction error rate of 56.7%. Such high error rates even using a high number of PLS factors suggests that PLS-DA may not be an appropriate tool when attempting to create a many class model.

Figure 6 shows how the PLS-DA model performance responds to the different number of factors used. Clearly this behavior is poor, showing almost a linear relationship over the first 10 factors at

high levels of error rate. This behavior has been observed previously[16] and it is suggested that this could be the result of cross-validation methods being inappropriate for the optimization of PLS-DA models. A bootstrap optimization is recommended. Resubstitution, bootstrap and cross-validation all gave similar results, although resubstitution gave slightly lower error rates as expected. Model accuracy can be improved by inclusion of a large number of PLS factors; however, such models are likely to include spectral noise and not just spectroscopic variance due to chemical differences between samples. It is therefore advisable to create models with the minimum number of factors. The NIPALS-LDA model demonstrates behavior more expected with a plateau being reached at 6 factors with a reasonable level of error rate.

PLS-DA generates a variable importance in the projection (VIP) score that enables comparison of the relative value of the different variables in the model. As a general rule a variable with VIP>1 is considered important to the model and should be retained whereas a variable with VIP<0.8 can be removed without affecting the performance of the model. Figure 7 shows a VIP spectrum compared to ATR spectra of the 3 land use types. A large proportion of the spectral range is above 0.8 with a considerable amount above 1. The VIP spectrum coincides with a number of features in the IR spectra particularly in the fingerprint region (400-1500 cm^{-1}). Another noticeable feature is the region between 1850-2400 cm^{-1} which is dominated by noise as discussed earlier and yet the VIP scores suggest that some of the variance seen here is contributing to the model. This was also seen with the PC1 loadings spectrum in figure 5. Examination of the VIP scores for the 14 factor model 2 showed this region to dominate and so the model is achieving its best discrimination by including this region. Based on these observations this region of the spectrum was removed giving a dataset consisting of two blocks, 400-1850 cm^{-1} and 2400-4000 cm^{-1} and referred to as the reduced dataset.

NIPALS factors from the reduced dataset show 99.35% of the variance captured in the first 7 factors (table 2). The factor 1 vs 2 score plot (figure 8) shows similar distribution to figure 2 which is expected as most replicates are well grouped and so replicate averages generally lie in similar positions in relation to each other. Site examples are clearly seen grouped together with the R3 group clearly separate from the other river bank groups.

NIPALS factors 1-6 give a prediction error rate of 8.3% (table 7) for a NIPALS-LDA model 1. The performance of the model is identical to that for the full data set showing that removal of the central spectral region appears to neither improve nor degrade the model.

The PLS-DA model also performs very well (table 9) and is comparable with the NIPALS-LDA model on the reduced data set and the PLS-DA model 1 on the full data set in further support that the data reduction does not appear to improve model performance.

Similar observations are observed with the NIPALS-LDA model 2 on the reduced data set in that similar results are obtained to the full data set (table 8). This model gives a prediction error rate of 23.3% which is not surprising upon closer inspection of figure 8 given the obvious overlap

between sites within a given group and the impact of a single misclassification with only 5 spectra per class. The misclassification errors are consistent with the patterns observed in figure 8, for example, 3 R1 samples are classed as FW1, FW4 has excellent precision (no false positives) but 1 FW4 sample is classed as an FW1 example. The information in the factor 1 vs 2 score plot is a good indicator of the likely success of developing a useful classification model. Further refinement of this model is needed to include a greater number of sites of the different land use types and to also determine what constitutes representative data per site. Here we have simply taken 5 sampling points in close proximity but a rationale for sampling and representation of location in the model needs to be developed.

The PLS-DA model 2 (table 8) is consistent with the previous model 2 on the full data set in requiring a large number of PLS factors to achieve a comparable level of error rate to the corresponding NIPALS-LDA model. The 5 factor model gives an error rate of 56.7% clearly indicating the difficulty PLS-DA has with the many-class problem. Reduction of the data appears to have little effect on the model performance.

It is known that PLS-DA has difficulty with multiclass problems and is best employed as a two-class modeling technique[18]. Here we have demonstrated that it performs very well on a three- class land use type problem but badly on a twelve class land use site problem. Methods to enable the use of PLS-DA on these problems have been proposed such as pairwise comparisons between classes which would mean running 66 two class classifications for the 12 class problem. This will become more complicated as more sites are added to the model, moving away from our aim to create a simple protocol for classifying soil samples. We therefore propose that the NIPALS-LDA approach used here offers the potential to be developed further as a tool in the classification of soil samples from their IR spectra.

Conclusion

It has been demonstrated that it is possible to build a classification model that can discriminate between land use type soils using only ATR spectra as input. Both NIPALS-LDA and PLS-DA achieved comparable performance in modelling 60 soil spectra to the 3 land-use types. An adaptation of this model to classify to land-use site was also relatively successful using NIPALS-LDA however this was not the case with PLS-DA which performed poorly and is therefore an inappropriate tool for a many class problem.

A simple data pre-treatment is proposed in only the ATR and baseline corrections are applied. Data reduction to remove a spectral region with poor signal-to-noise ratio appears to make little difference to the performance of the model.

Representation of sites is also another important consideration in moving towards a model that can be validated and used on real samples. Five independent samples per site gave a reasonable model but a limitation of this work is only having four sites per land use type. Further work will develop a sampling rationale per site with a greater number of sites incorporated into the model. If this level of discrimination can be achieved such a model would be of even greater benefit for criminal intelligence purposes and enable regions to be mapped to the level of site. It may well be found that mid-IR spectroscopy alone is insufficient to achieve this and so combination with other spectroscopic techniques such as near infrared and visible reflection spectroscopy along with other soil parameters such as organic/inorganic composition and amount of carbonate may be necessary.

The NIPALS-LDA tool seems to offer a simple and effective approach for modelling this multi-class problem. However models at different levels will offer a more organized and systematic approach as the dataset increases. A model that discriminates land use type followed by a model that then discriminates sites within land use type could be one approach. The use of classification tree tools could be another and will be explored in further work.

References

1. M. Andersson, RT Ottesen and M. Langedal, Geoderma **156,** 112 (2010).
2. L.I. Tsikritzis, J. Radioanal. Nucl. Chem **261**, 1, 215 (2004).
3. A.M. Mouazen, B. Kuang, J. De Baerdemaeker and H. Ramon, Geoderma **158,** 23 (2010).
4. J.L. Perez Pavon, A. Guerrero Pena, C. Garcia Pinto and B. Moreno Cordero, J. Chrom. A **1137**, 101 (2006).
5. M. Stemmer, K. Roth and E. Kandeler, Biol. Fertil. Soils **31**, 294 (2000).
6. A.O. Awiti, M. G. Walsh, K. D. Shepherd and J. Kinyamario, Geoderma **143**, 73 (2008).
7. V.F. Melo, L.C. Barbar, P.G.P. Zamora, C.E. Schaefer and G.A.Cordeiro, Forensic Sci. Int. **179**, 123 (2008).
8. N.C. Thanasoulias, E. T. Piliouris, M.S.E. Kotti and N.P. Evmiridinis, Forensic Sci. Int **130**, 73 (2002).
9. S. Dragovic and A. Onjia, Journal of Environmental Radioactivity **89,** 150 (2006).
10. T. Borregaard, H. Nielsen, L Nørgaard and H. Have, J. Agric. Eng. Res. **75**, 389 (2000).
11. R.J. Cox, H.L. Peterson, J. Young, C. Cusik and E.O. Espinoza, Forensic Sci. Int. **108,** 107 (2000).

388 12. D. Solomon, J. Lehmann, J. Kinyangi, B. Liang and T. Schäfer, Soil Sci. Soc. Am. J. **69**, 107
389 (2005).

390 13. A. Vergnoux, M. Guiliano, R. Di Rocco, M. Domeizel, F. Theraulaz and P. Doumenq,
391 Quantitative and mid-infra-red changes of humic substances from burned soils,
392 Environmental research (article in press).

393 14. G.N. Elliott, H. Worgan, D. Broadhurst, J. Draper and J. Scullion, Soil Biol. Biochem. **39** 2888
394 (2007).

395 15. R. Rakotomalala, Proceedings of EGC'2005, RNTI-E-3 **2,** 697 (2005).

396 16. A.N. Fernandes, M. Giovanela, V.I. Esteves and M.M. de Souza Sierra, J.Mol. Struct. **971,** 33
397 (2010).

398 17. M. Ritz, L. Vaculikova and E. Plevova, Appl.Spectrosc. **64,** 1379 (2010).

399 18. R. Brereton, *Chemometrics for Pattern Recognition* (Wiley, Chichester, UK, 2009)

400 19. M. Cohen, R.S. Mylavarapu, I. Bogrekci, W.S. Lee and M.W. Clark, Soil Science **172,** 469
401 (2007).

402

403

404 **Figure captions:**

405 Figure 1: Average ATR spectra for all site 1 soil samples obtained from the 3 land use types.

406 Figure 2: NIPALS factor 1 (55 %) versus factor 2 (22%) score plot for 400-4000 cm^{-1} data set with
407 replicates. FW flowerbed, W woodland, R river soil types.

408 Figure 3: NIPALS factor 1 (65%) versus factor 2 (21%) score plot for river sites only (including
409 replicates).

410 Figure 4: CV(RMSD) spectrum for a set of triplicate flowerbed spectra.

411 Figure 5: Factor loadings plot for NIPALS factors 1-5 on average data set.

412 Figure 6: Cross-validation % prediction error rate as a function of the number of NIPALS/PLS
413 factors used in the classification model.

414 Figure 7: VIP values for PLS-DA model 1 compared to full spectra from the 3 land use types used
415 in the model. A is VIP values. B is a river; C is a flowerbed; D is a woodland soil sample.

416 Figure 8: NIPALS factor 1(57%) vs factor 2 (26%) score plot for the reduced data set.

Tables

Table 1. Description and identification of the land use sites and the different soil types used in this study

Sample	code	GB National Grid reference	Brief description of the location
woodlands	W1	SK946696	urban park, close to a lake area
woodlands	W2	SK952687	urban park, close to a pond area, near a road
woodlands	W3	SK954686	urban park, close to a road
woodlands	W4	SK965689	urban park, close to a lake area
flowerbeds	FW1	SK985717	flowerbed close to an urban park in the city centre
flowerbeds	FW2	SK966714	flowerbed in the University Campus close to student accommodation and close to a railway line
flowerbeds	FW3	SK967713	flowerbed in the University Campus and close to a railway line
flowerbeds	FW4	SK968709	flowerbed in a roundabout close to the city centre
riverbank	R1	SK967691	riverbank close to a urban park with little vegetation
riverbank	R2	SK963709	riverbank close to the university campus and close to a railway line with little vegetation
riverbank	R3	SK966719	Riverbank close to a pumping station with little flow of water
riverbank	R4	SK971705	Riverbank close to an industrial state in the city

Table 2: NIPALS factors 1-10 for 400-4000 cm^{-1} data set. Also shown for comparison are the PCA results for PC1-10

Factor	NIPALS Eigen value			NIPALS Explained variance (%)			PCA Average	
	Replicates	Average	Average reduced	Replicates	Average	Average reduced	Eigen value	Explained variance (%)
1	1984	2036	1700	55.09	56.54	56.64	2036	56.54
2	794	830	802	22.06	23.04	26.71	830	23.04
3	293	291	264	8.16	8.19	8.81	291	8.19
4	170	147	117	4.72	4.09	3.90	147	4.09
5	134	119	75	3.73	3.30	2.49	119	3.30
6	70	56	19	1.95	1.54	0.62	56	1.54
7	28	27	7	0.77	0.74	0.25	27	0.74
8	10	10	4.4	0.28	0.27	0.15	10	0.27
9	9	8	3.8	0.26	0.22	0.13	8	0.22
10	6	7	1.6	0.18	0.19	0.05	7	0.19

Table 3: Contingency table for NIPALS-LDA model 1 using factors 1-7 obtained from 400-4000 cm^{-1} data set. Error rate = 8.3%

	FW	W	R	Sum
FW	17	1	2	20
W	1	19	0	20
R	1	0	19	20
Sum	19	20	21	60

Table 4: Comparison of NIPALS-LDA and PLS-DA model 1 on 400-4000 cm^{-1} data set.

	NIPALS-LDA			PLS-DA		
Number of factors	7			5		
% error	8.3			10		
	Recall	1-precision	Accuracy	Recall	1-precision	Accuracy
FW	0.85	0.1053	0.8667	0.75	0	0.9167
W	0.95	0.05	0.9667	0.95	0.05	0.9667
R	0.95	0.0952	0.9000	1	0.2	0.9167

Table 5: Contingency table for NIPALS –LDA model 2 using factors 1-6 from 400-4000 cm^{-1} data set. Error Rate = 21.7 %

	FW1	FW2	FW3	FW4	W1	W2	W3	W4	R1	R2	R3	R4	Sum
FW1	5	0	0	0	0	0	0	0	0	0	0	0	5
FW2	0	4	0	0	0	0	0	0	0	1	0	0	5
FW3	1	0	3	0	0	0	1	0	0	0	0	0	5
FW4	1	0	0	4	0	0	0	0	0	0	0	0	5
W1	0	0	0	0	4	0	0	0	0	1	0	0	5
W2	0	0	0	0	1	3	1	0	0	0	0	0	5
W3	0	0	0	0	0	0	5	0	0	0	0	0	5
W4	0	0	0	0	0	1	0	4	0	0	0	0	5
R1	0	1	0	0	1	0	0	0	3	0	0	1	5
R2	0	0	0	0	0	0	0	0	0	5	0	0	5
R3	0	0	0	0	0	0	0	0	0	0	5	0	5
R4	0	0	0	0	0	0	0	0	3	0	0	2	5
Sum	7	5	3	4	6	4	7	4	6	6	6	2	60

435 **Table 6: Comparison of NIPALS-LDA and PLS-DA model 2 on 400-4000 cm^{-1} data set.**

	NIPALS-LDA		PLS-DA	
Number of Factors	6		14	
% error	21.7		31.7	
	Recall	1-precision	Recall	1-precision
FW1	1	0.2857	1	0
FW2	0.8	0.2	0.8	0.4286
FW3	0.6	0	0	1
FW4	0.8	0	1	0.2857
W1	0.8	0.3333	0.4	0.5
W2	0.6	0.25	0.2	0
W3	1	0.2857	1	0.5
W4	0.8	0	1	0
R1	0.6	0.5	0.4	0.7143
R2	1	0.1667	1	0.2857
R3	1	0.1667	1	0.1667
R4	0.4	0	0	1

436

437

438 **Table 7: Comparison of NIPALS-LDA and PLS-DA model 1 on reduced data set.**

	NIPALS-LDA			PLS-DA		
Number of factors	6			5		
% error	8.3			8.3		
	Recall	1-precision	Accuracy	Recall	1-precision	Accuracy
FW	0.85	0.1053	0.8667	0.8	0	0.8667
W	0.95	0.05	0.9667	0.95	0.05	0.9667
R	0.95	0.0952	0.9000	1	0.1667	0.9333

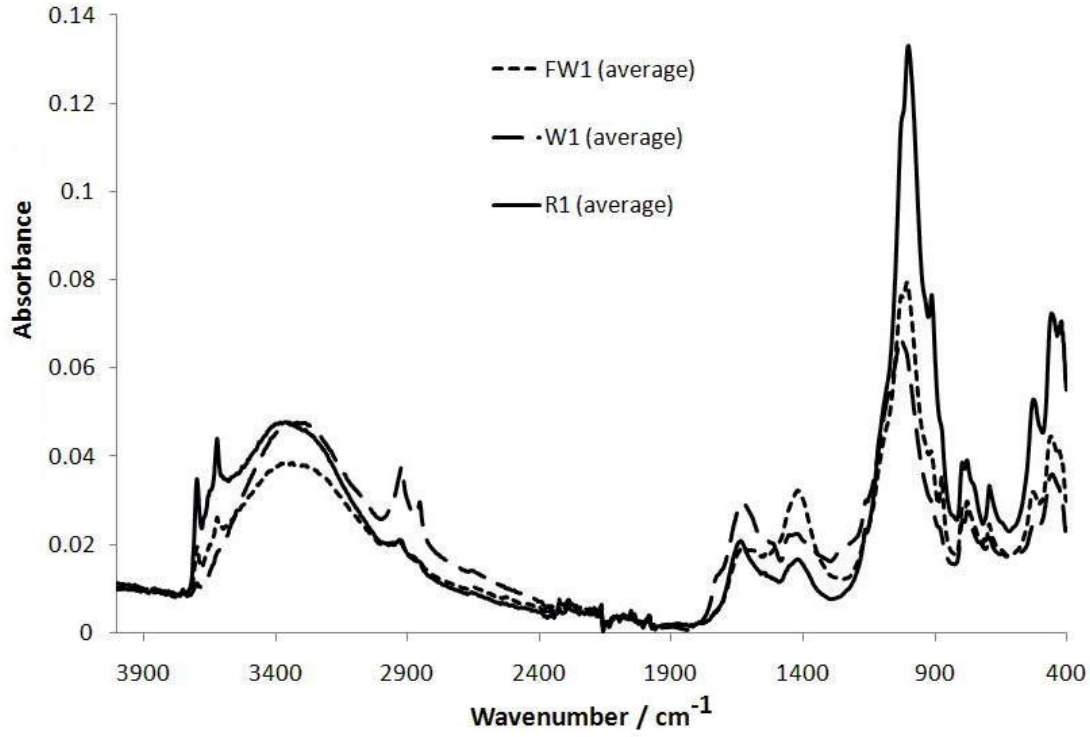
439

440 **Table 8: Comparison of NIPALS-LDA and PLS-DA model on reduced data set**.

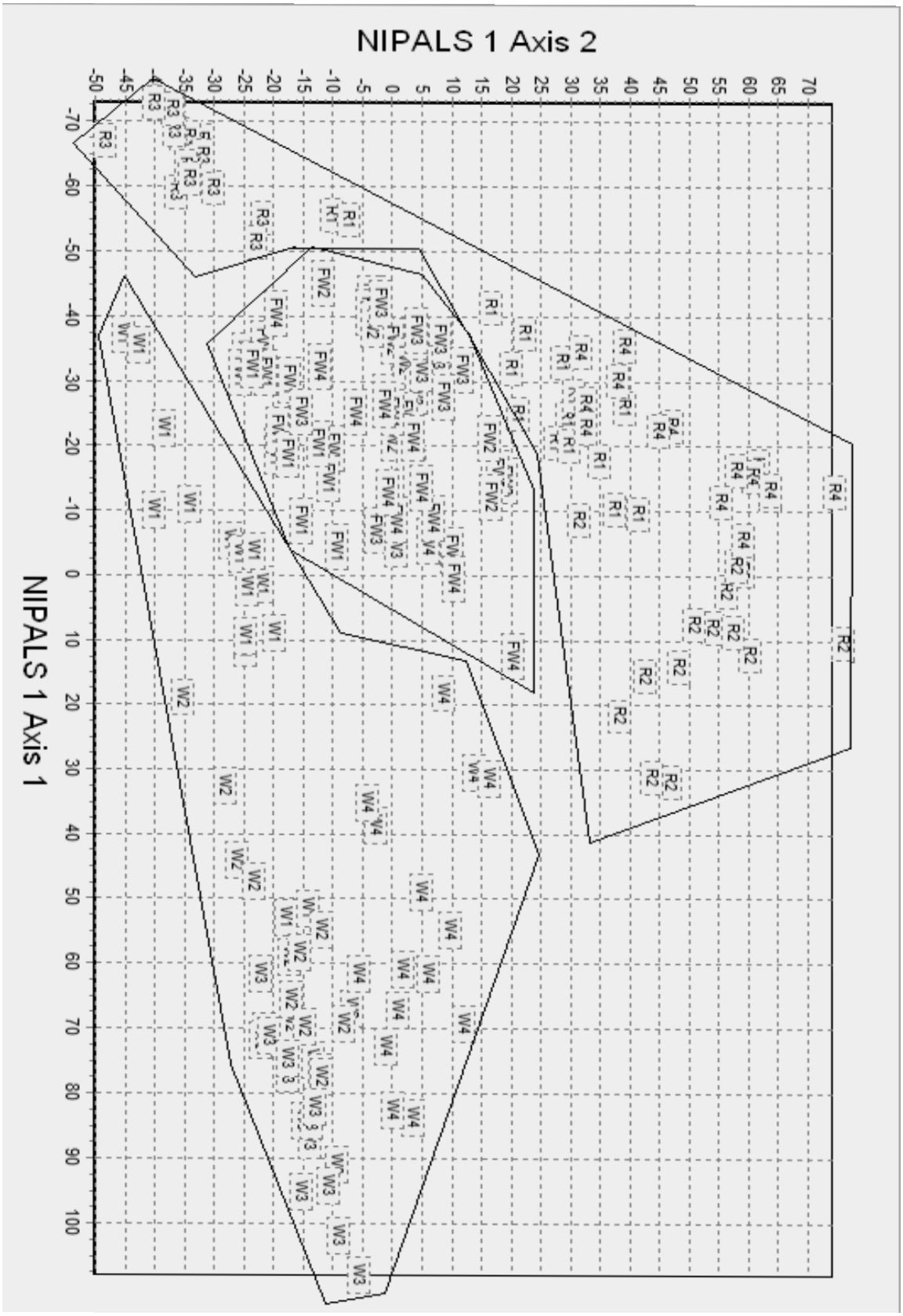
	NIPALS-LDA		PLS-DA	
Number of Factors	6		15	
% error	23.3		31.7	
	Recall	1-precision	Recall	1-precision
FW1	1	0.2857	0.8	0.2
FW2	0.8	0.2	0.6	0.5714
FW3	0.6	0.25	0.6	0
FW4	0.8	0	1	0.1667
W1	0.6	0.25	0.4	0.5
W2	0.6	0.4	0.6	0.4
W3	1	0.2857	0.6	0.4
W4	0.8	0	0.8	0
R1	0.6	0.5	0.4	0.5
R2	1	0.1667	0.8	0.2
R3	1	0.1667	1	0.2857
R4	0.4	0	0.6	0.4

441

Figure 1


0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Absorbance
FW1 (average)
·W1 (average)
R1 (average)
3900
3400
2900
2400
1900
1400
900
400
Wavenumber / cm-1

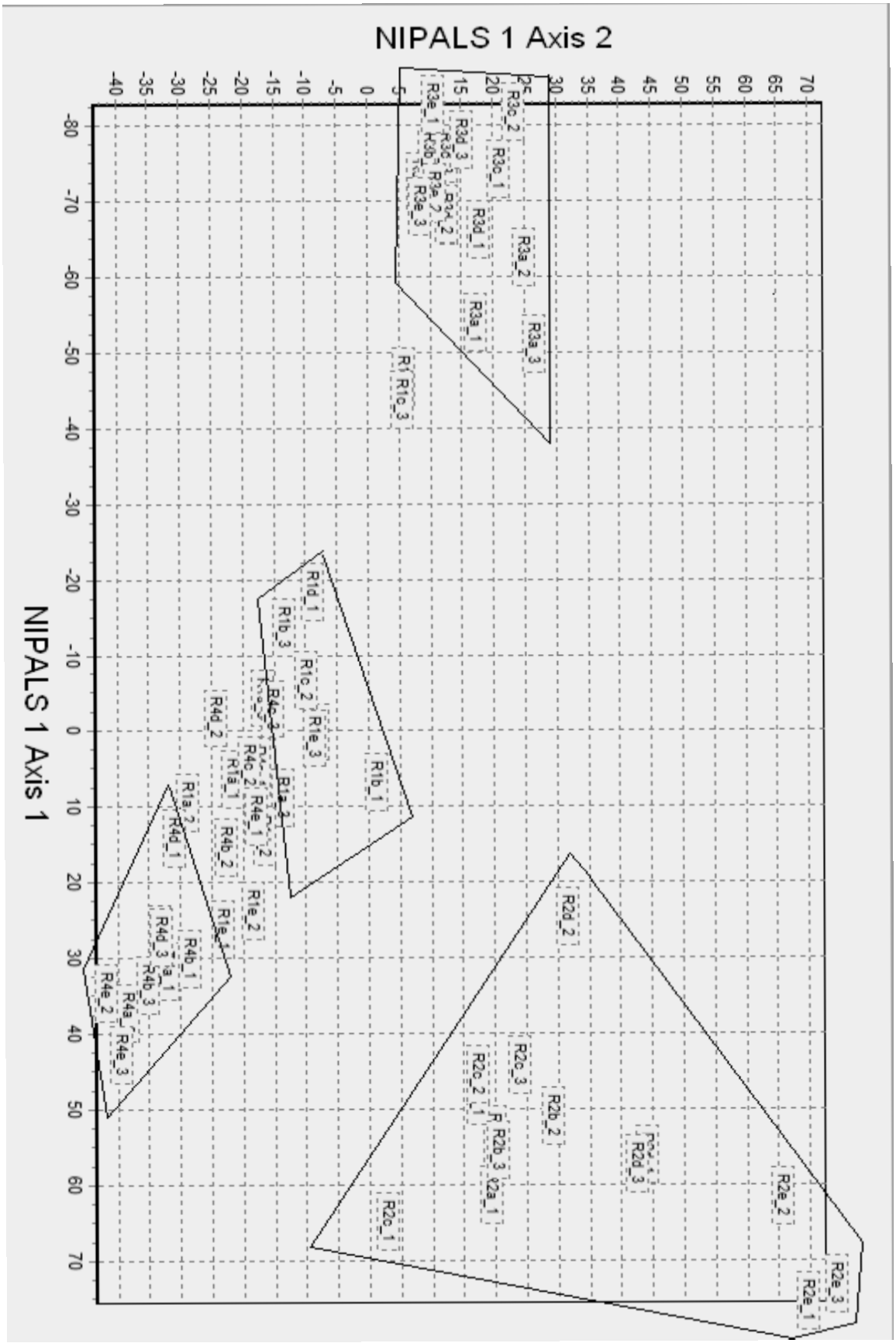
458 Figure 2



459

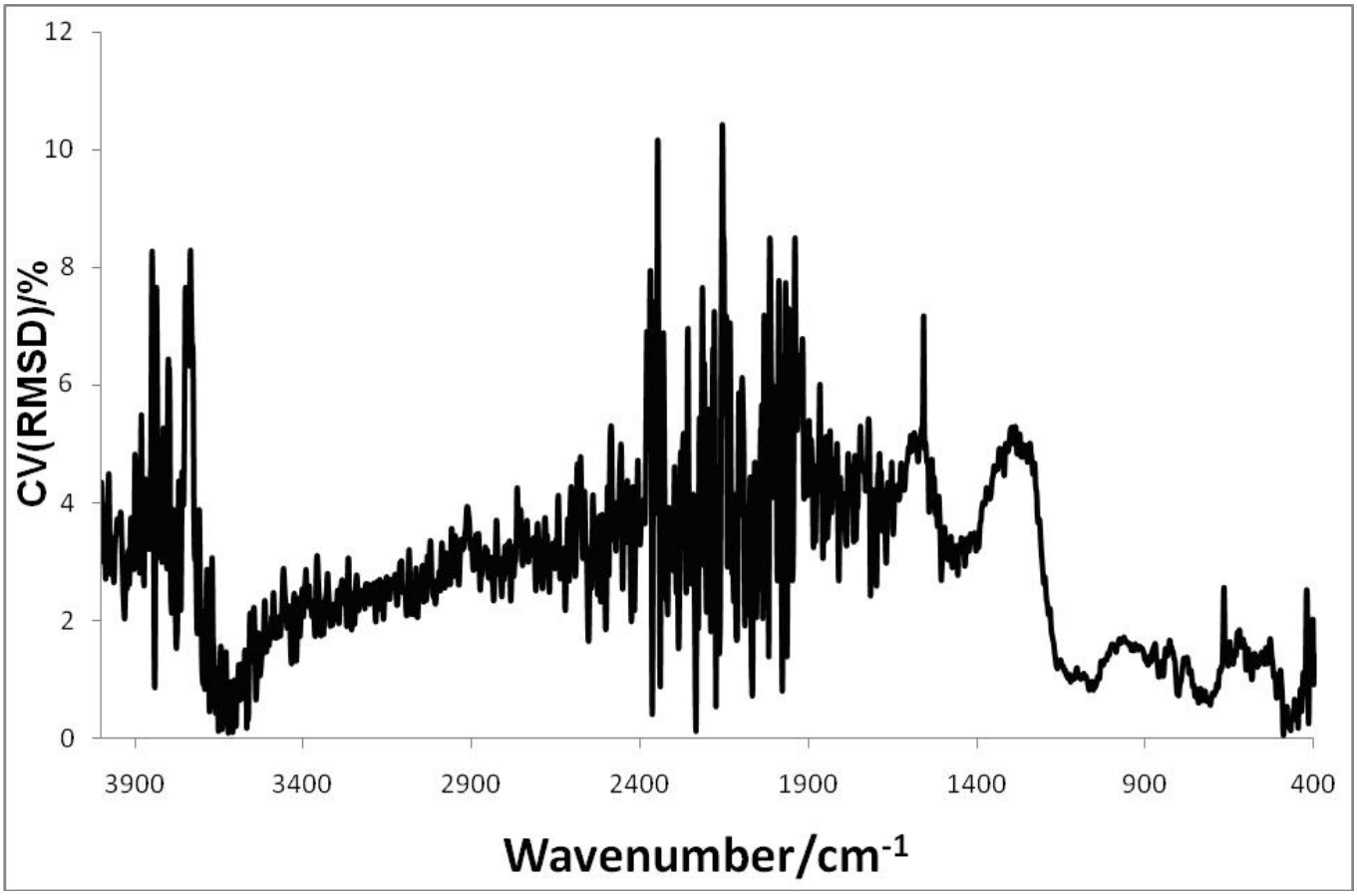
460

461 Figure 3

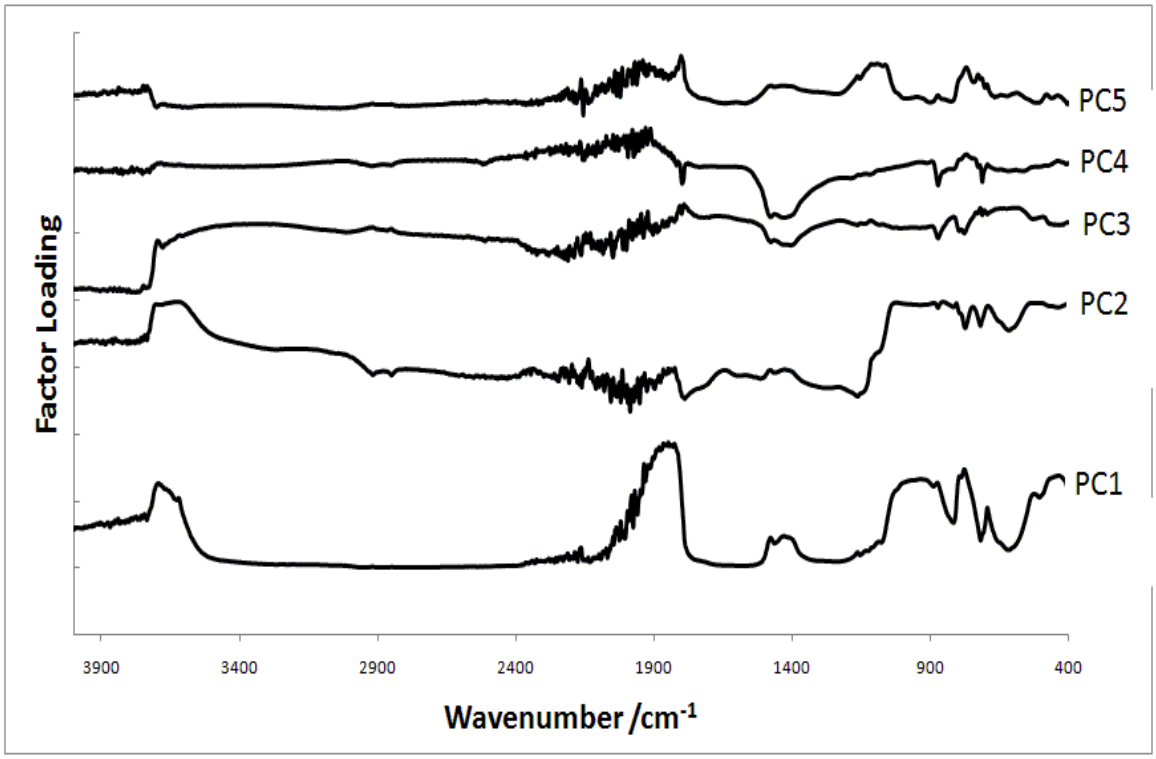


462

Figure 4


12
10
8
6
4
2
0
CV(RMSD)/%
3900
3400
2900
2400
1900
1400
900
400
Wavenumber/cm-1

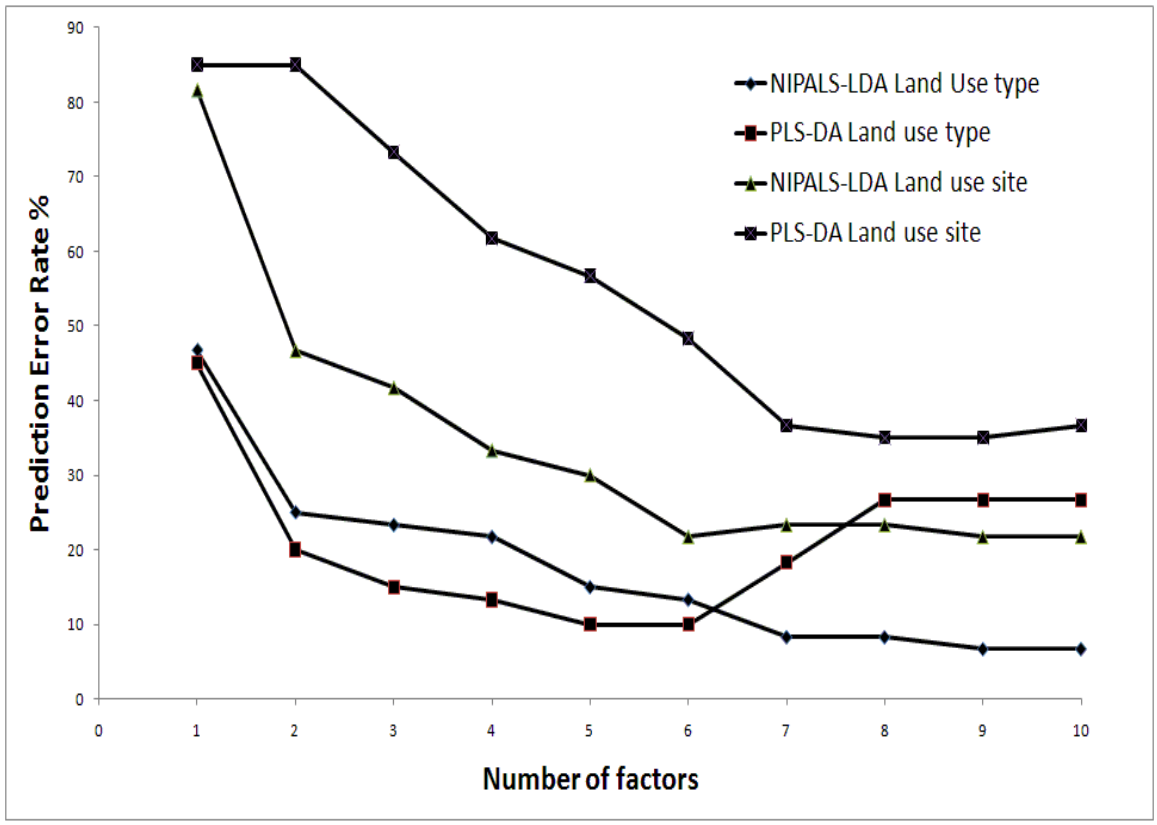
477 Figure 5



478

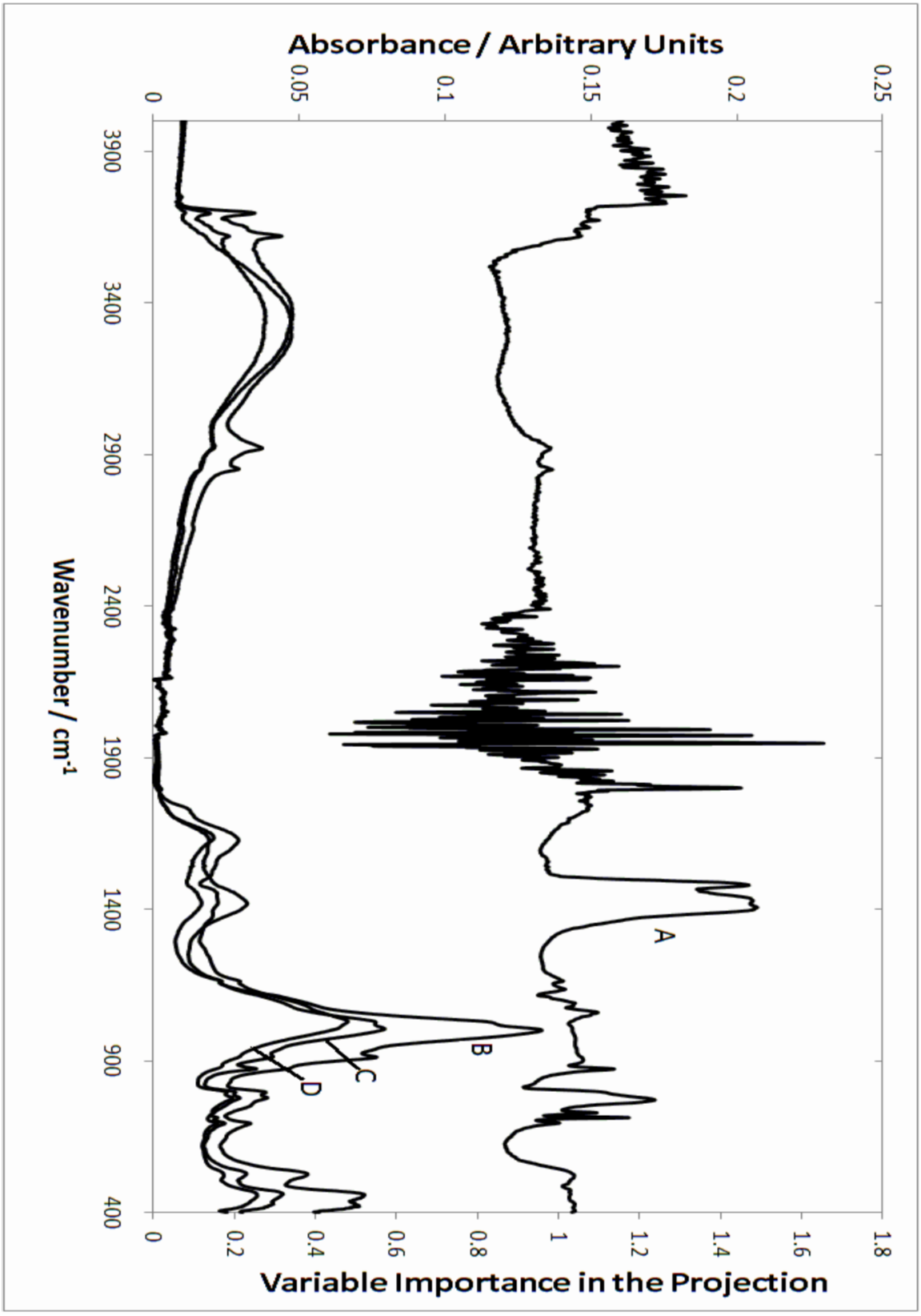
479

480 Figure 6



481

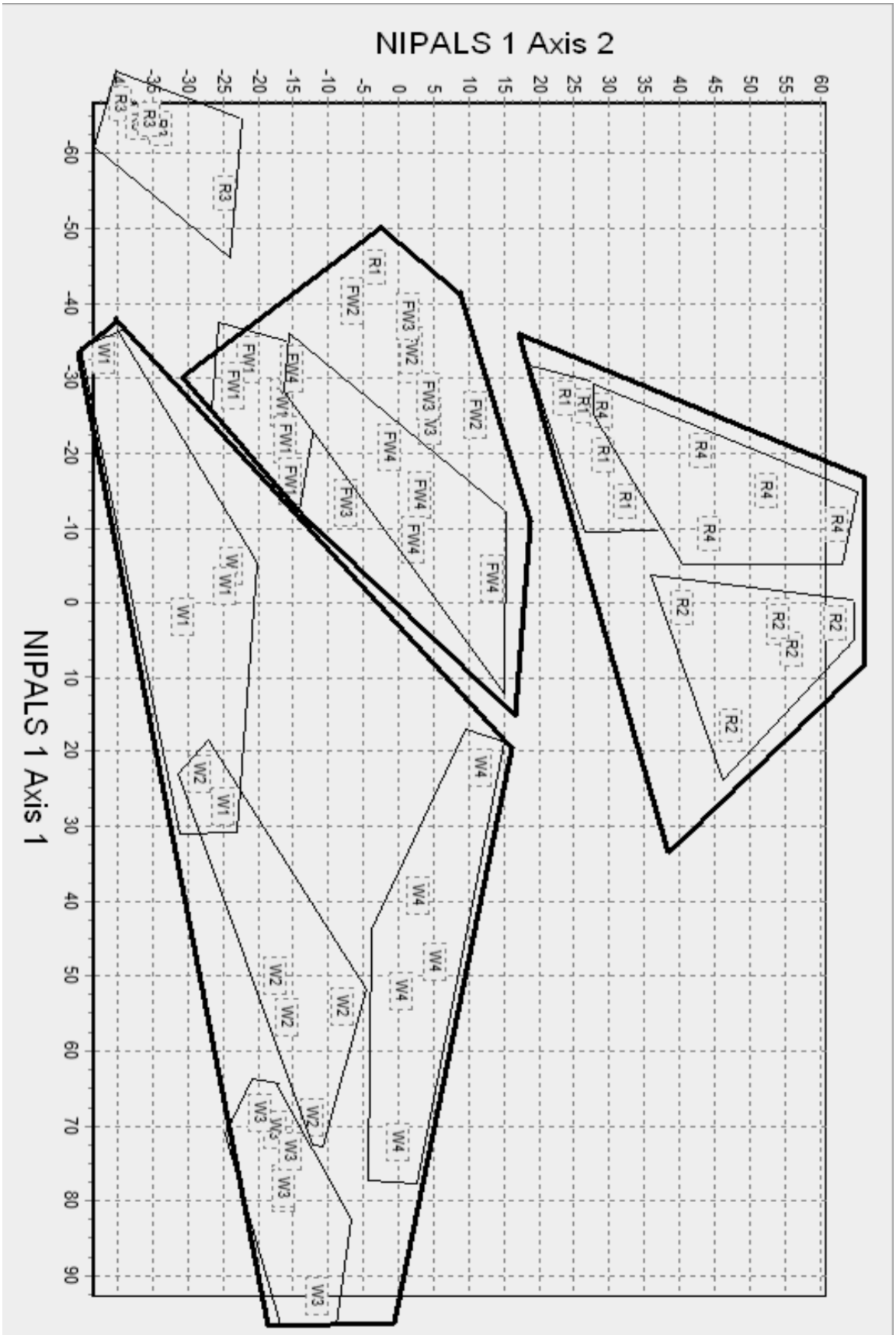
482 Figure 7



483

484

485 Figure 8



486