Soil analysis by near infrared spectroscopy: examination of methods to solve non-linearity in regression

Yoshisato Ootake,[a] Masayuki Hioki,[a] Hiroyuki Tamaru,[b] Tsuyoshi Sato,[c] Akihiro Miyoshi[d] and Toshihiko Yoshikawa[d]

[a]*Aichi-ken Agricultural Research Centre, 1-1 Sagamine, Yazako, Nagakute, Aichi, 480-1193, Japan. E-mail: ootake-y@jasmine.ocn.ne.jp.*

[b]*Hokkaido-Asahikawa Agricultural Experiment Station.*

[c]*Nagano-ken Agricultural Experiment Station.*

[d]*Hyogo-ken Agricultural Research Centre.*

Introduction

Soil analysis is essential to make agricultural production stable and to yield a good harvest. In order to indicate the nutrient conditions in the soil, total nitrogen (T-N), total carbon (T-C), cation exchange capacity (CEC) and phosphate sorption coefficient (PSC) are often measured. Using conventional wet chemical methods to analyse these constituents requires a lot of time and labour. Also, large amounts of reagents, which are sometimes harmful, have to be used to carry out the experiments.

Near infrared (NIR) spectroscopy provides a possible alternative to save both time and labour and, also, workers can avoid exposure to such harmful reagents. Furthermore, the load on the environment can also be reduced. In this experiment, the authors developed stable calibration equations to predict constituents such as T-N, T-C, CEC and PSC for representative soil groups in Japan so that NIR soil measurements could be used at more practical locations without complicated sample preparation. Previous work carried out has shown that NIR is a promising method for analysing soil.[1–3] These authors also tried to develop stable calibration equations which could be used at the practical sites. However, there appeared to be some difficulties in developing these equations. In this report the cause of those difficulties and proposed solutions are discussed.

Experimental

228 soil samples were used, made up of the following groups: andosols (AS), Grey Upland Soil (GrUS), Yellow Soil (YS), Brown Lowland Soil (BLS), Grey Lowland Soil (GrLS) and Gley Soil (GlS). Sample pre-treatment: soil samples were air-dried and were passed through a 2 mm diameter sieve. Spectra collection: Bran+Luebbe's InfraAlyzer 500 was used. Samples were placed into a diffuse reflectance cup, then spectra were collected from 1100 to 2500 nm with a 4 nm steps. Spectra collection was duplicated. In total, 456 spectra were used for the analyses. Constituents measured: T-N,

Near Infrared Spectroscopy: Proceedings of the 9th International Conference

Table 1. Range of constituents contained in the soils which were used.

Soil group	T-N		T-C		CEC		PSC	
	max	min	max	min	max	min	max	min
AS	0.569	0.103	9.19	1.14	47.7	9.1	2400	730
GrUS	0.278	0.082	4.05	0.94	28.7	5.9	1189	128
YS	0.449	0.027	4.46	0.27	30.6	5.0	1610	135
BLS	0.510	0.064	6.99	0.62	31.9	8.3	1593	278
GrLS	0.343	0.057	3.51	0.52	21.9	5.5	1140	240
GIS	0.540	0.080	5.68	0.80	29.8	8.9	1324	110

T-C, CEC and PSC. The range of constituents contained in the soil are shown in Table 1. Data analysis was carried out using "The Unscrambler" chemometrics software (Camo AS, Norway).

Results and discussion

The first trial

The first step for developing calibration equations was done by principal component regression (PCR) using raw spectra.

The accuracy of calibration equations for each constituent is shown in Table 2. Although the accuracy of the calibration results were all insufficient, those of CEC and PSC were slightly poorer than T-N and T-C. On the other hand, as a common feature throughout the results, some curvatures, which suggest non-linearity, were seen in scatter plots (Figure 1). The prediction value of some samples was extended farther than zero to the side of minus. In this report, hereafter, we will discuss only T-N and T-C.

Trial for improving calibration equations by spectral pre-treatment

As the absorbance of the spectra varied a great deal, and this was thought to be one of the causes of calibration equations being less accurate, spectral pre-treatment was applied with multiplicative signal correction (MSC) on the data of T-N and T-C. By applying MSC, the extent of curvature appeared

Table 2. PCR results using all samples.

Constituent	Spectra	PC sused	Calibration		Predictions			
			RMSE	*r*	*RMSE*	*r*	Slope	Offset
T-N	Raw	16	0.0419	0.907	0.0461	0.882	1.013	–0.016
	MSC	15	0.0458	0.887	0.0459	0.884	1.052	–0.001
T-C	Raw	20	0.558	0.926	0.507	0.912	1.072	–0.065
	MSC	19	0.544	0.929	0.547	0.894	1.060	–0.061
CEC	Raw	15	3.20	0.897	3.17	0.857	1.117	–1.620
PSC	Raw	17	222	0.844	209	0.855	0.979	–0.730

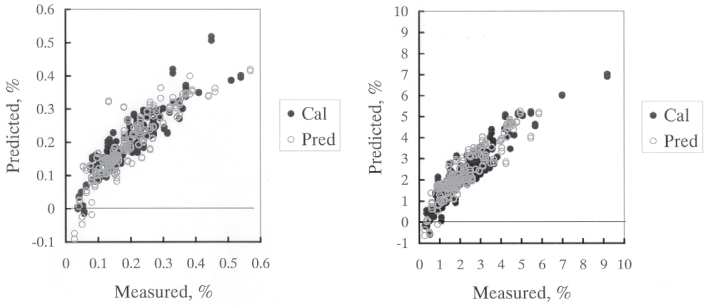

Figure 1(a). Scatter plots of T-N and T-C using raw spectra. Left: T-N, Right: T-C. Cal: Calibration set. Pred: Prediction set.

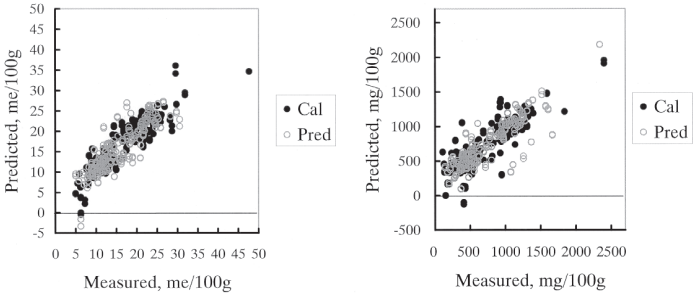

Figure 1(b). Scatter plots of CEC and PSC using raw spectra. Left: CEC, Right: PSC.

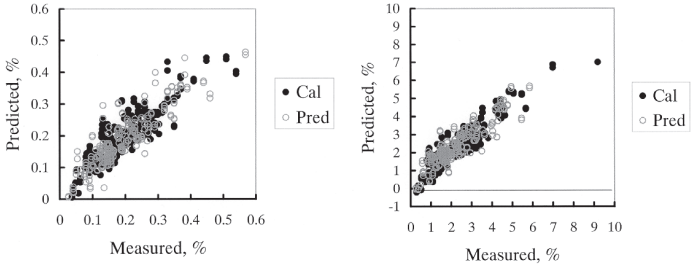

Figure 2. Scatter plots of T-N and T-C using MSC spectra. Left: T-N, Right: T-C.

to be mitigated and exceeding the prediction value toward minus almost disappeared. On the other hand, distribution of regression was rather broad (Table 2, Figure 2).

The problem of non-linearity seemed not to have been solved, even after spectral pre-treatment. Then, looking at Figures 1 and 2, there seems to be some particular sample group which exceeds to the side of minus in Figure 1 and the trend can also be seen in Figure 2.

Figure 3 shows the average spectra of each soil group. In this figure, the following points can be noted; steep slope at shorter wavelength range and the shift of the peak at around 1920 nm in AS, larger height of the peaks at around 1420 and 2120 nm and typical shoulders at the left side of the peaks, etc.

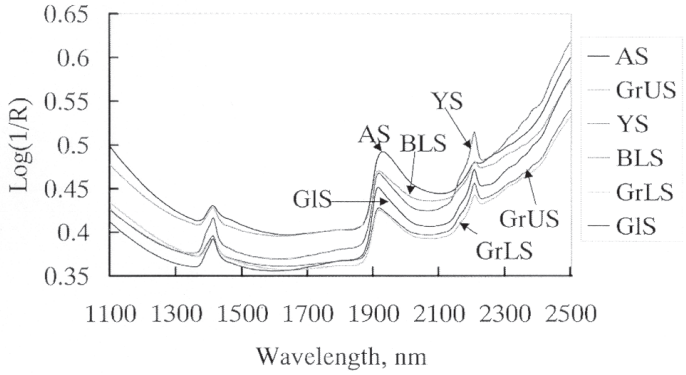

Figure 3. Average spectrum of each soil group. AS: Andosols, GrUS: gray uplands soils, YS: yellow soils, BLS: brown lowland soils, GrLS: gray lowland soils, GlS Gley soils.

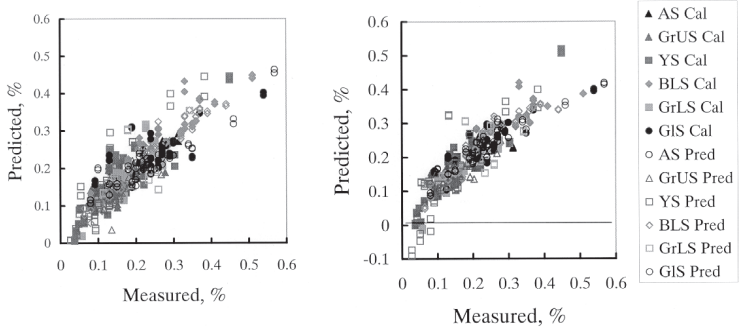

Figure 4. Scatter plots of T-N for each soil group. Left: MSC spectra, Right: raw spectra.

As it was possibile that the cause of non-linearity could be the difference in the characteristics of the soil groups, regression scatter plots were again made, separating each soil group (Figure 4).

When plotted separately, some clusters corresponding to each soil group(s) which has each trend line in the original plotting (Figure 1) became apparent (Figure 4). Typical clustering was seen for YS, BLS, AS and other soil groups.

Examination of *x*-variables

When the non-linearity problem was first found, the authors suspected that there were some errors in chemical values. However, after re-analysis, chemical data appeared to be correct.

Then we attempted to look into *x*-variables (spectral data) in relation to differences among soil groups.

Clustering according to soil group(s) was observed in PC1 vs PC2 scores plot of the PCA result from MSC spectra. Similar clustering was recognised in PC3 vs PC4 scores plot from the raw spectra. It was concluded from these results that the cause of non-linearity was the different characteristics of the soil groups (Figure 5).

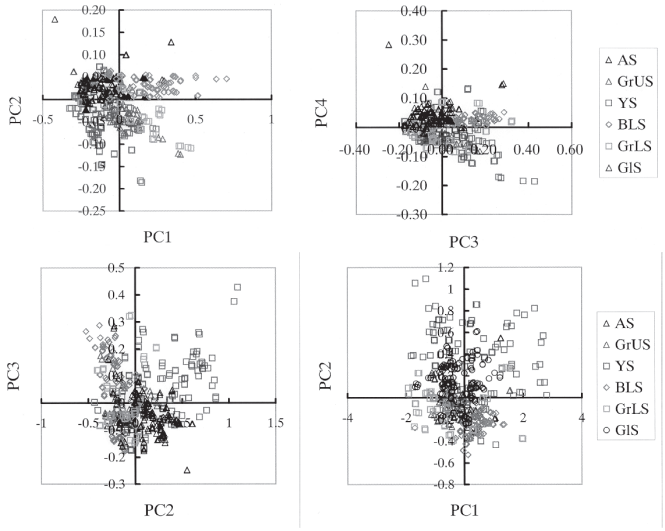

Figure 5. Score-score plots for each soil group. Upper left: MSC spectra PC1 v. PC2. Upper right: raw spectra PC3 v. PC4. Lower left: raw PC2 v. PC3. Lower right: raw PC1 v. PC2.

The similarity between two plots is thought to suggest that MSC [full: M = (M–a)/b] played a similar role with the extraction of PC1 and 2 from the raw spectra. This point can be seen in Figure 6; i.e. the patterns of loadings of PC1, of MSC and PC3 of raw, PC2 of MSC and PC4 of raw were similar, respectively.

Development of calibration equations for each soil group(s)

The authors then developed calibration equations separately for each soil group(s). The results for BLS and YS were fairly good. On the other hand, satisfactory calibration results for other soil groups such as GrUS, GrLS or GlS were not obtained. Rather, when those three groups were analysed together, better results were obtained (Figure 7, Table 3). In each case non-linearity did not appear.

Calibration development for AS was not attempted, because the number of samples of AS was too small to develop calibration equations.

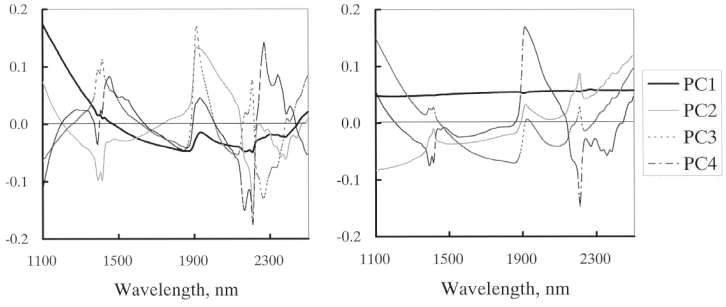

Figure 6. Plots of *x*-loadings of the first four PCs in the PCR results of Figure 4. Left: MSC spectra, Right: raw spectra

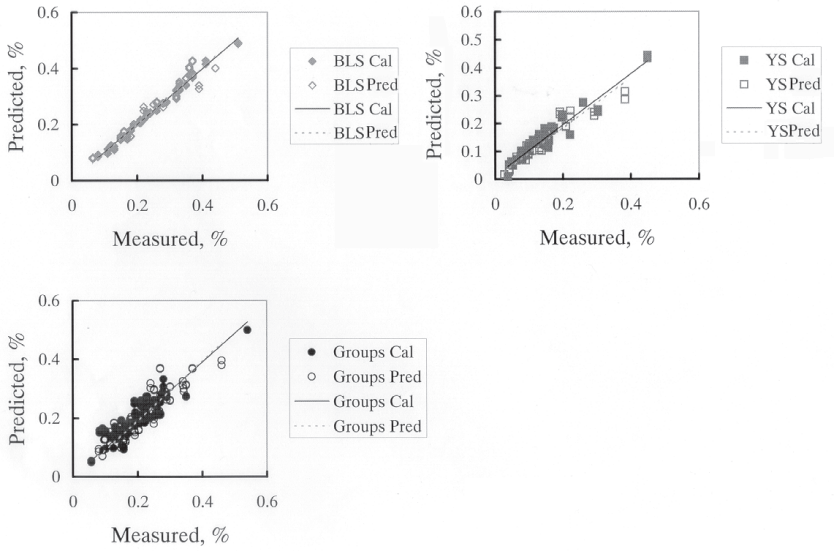

Figure 7. Scatter plots of the results calculated spearately for each soil group(s) for T-N. Upper left: BLS, upper right: YS, lower: groups of GrUS, GrLS and GlS.

Attempt to classify soil by soft independent modelling of class analogy (SIMCA)

As seen in Figure 5, different trends in spectral characteristics existed, even in the same soil group, such as YS in PC1 v. PC2 scores plot from raw spectra.

Therefore, classification of soil, on the basis of spectral characteristics, was attempted using SIMCA between YS and Groups of GrUS, GrLS or GlS. Although the classification resulst were not satisfactory, 84.2% of the YS and 93.4% of the "Groups" were classified as "closer to the correct model"[4] using raw spectra, 89.5% of YS and 90.8% of the groups using MSC, as well.

When some samples of YS which were classified as "closer to the Groups" are validated by the regression model for the "Groups", prediction results were not worse than those of the "Groups". A sam-

Table 3. PCR results of each soil group.

Constituent	Soil group	PCs used	Calibration		Prediction			
			RMSE	*r*	*RMSE*	*r*	Slope	Offset
T-N	YS	16	0.0275	0.958	0.0277	0.936	0.941	0.000
	BLS	15	0.0115	0.994	0.0251	0.970	1.030	–0.005
	Groups	16	0.0371	0.908	0.0341	0.895	1.052	–0.015
T-C	YS	18	0.415	0.942	0.307	0.909	0.933	–0.135
	BLS	8	0.267	0.988	0.328	0.977	1.046	–0.046
	Groups	15	0.414	0.918	0.456	0.871	1.043	0.010

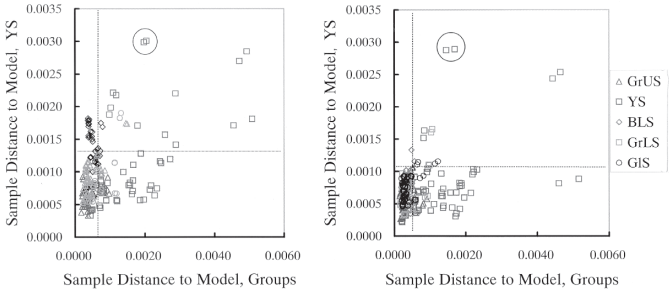

Figure 8. Cooman's plots of classification results on the models made for YS and Groups of GrUS, GrLS and GIS. Left: raw spectra, right: MSC spectra (offset correction).

ple enclosed by a circle in Figure 8 is situated apart from other samples (Figure 9). In this analysis, classification results were displayed simply by comparing distances from the "models". If the value of 95% significance limit line is considered, the classification results may be different and the sample above may have been classified as "closer to YS". [4]

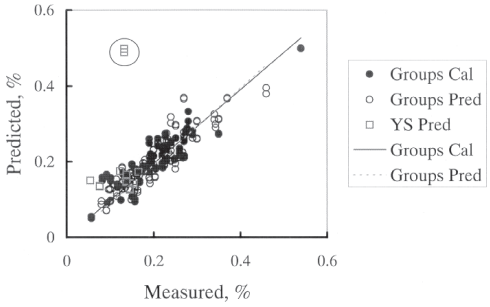

Figure 9. Scatter plot of PCR result of T-N for the "Groups" including YS which was classified to the "Groups".

Issues left for future analyses

In this report, so far, the cause of non-linearity in soil analysis and the fundamental point for the solution of it have been discussed.

However, developing calibration equations for separate soil groups was carried out only using raw spectra. Examination of wavelength ranges, spectra pre-treatments and so on have not been carried out. Also, classification by SIMCA has not been completed for all soil groups. These issues are left for future analyses.

References

1. T. Matsunaga and M. Uwasawa, *Nippon Dojo Hiryogaku Zasshi* **64,** 329 (1993).
2. A. Salgo, J. Nagy, J. Tarnoy, P. Marth, O. Palmai and G. Szabo-Kele, *J. Near Infrared Spectrosc*. **6,** 199 (1998).
3. R.K. Cho, G. Lin and Y.K. Kwon, *J. Near Infrared Spectrosc*. **6,** A87 (1998).
4. Y. Ootake and S. Kokot, *J. Near Infrared Spectrosc.* **6,** 241 (1998).