Interpretation of the cause of the non-linearity problem seen in the NIR measurement of soil

Yoshisato Ootake

Aichi-ken College of Agriculture, 1-2 Namimatsu, Miai, Okazaki-shi, Aichi, 444-0802, JapanE-mail: ootake_yoshisato@ybb.ne.jp

Introduction

In NIR 99, the author reported that the non-linearity problem was seen in the scatter plots of regression for some nutrition indices in several kinds of soil groups, showing that its curvature was caused by the difference of spectral characteristics of each soil group. Also the possibility of the solution by classifying the soil groups according to their spectral characteristics was suggested.[1]

However, the cause of the curvature was just pointed out by raising the distribution pattern of score-score plots of PC-3 v. PC-4 in raw spectra, and PC-1 v. PC-2 in MSC spectra in which clustering of each soil groups can be observed. In this report, therefore, further analyses about the cause of non-linearity are carried out. Also a possibility to remove curvatures not developing calibration equations on each soil group but using all soil groups together was pursued.

Experimental

Soil groups used: andosol (AS), Grey Upland Soil (GrUS), Yellow Soil (YS), Brown Lowland Soil (BLS), Grey Lowland Soil (GrLS) and Gley Soil (GlS) of which total number is 278. Sample Pre-treatment: Soil samples were dried and were passed through 2mm diameter sieve. Spectra collection: BRAN+LUEBBE's InfraAlyzer 500 was used. Sample was filled into a diffuse reflectance cup, then spectra were collected from 1100nm to 2500nm with 4nm step. Spectra collection was duplicated. In total, 456 spectra were used for analysis. Although constituents analysed were, originally, total nitrogen (T-N), total carbon (T-C), cation-exchange capacity (CEC) and phosphate sorption coefficient (PSC), discussion of the causes of the curvature is carried out just for total nitrogen in this report. Data analysis was carried out by the chemometrics software "The Unscrambler" (Camo AS, Norway) and "MS Excel".

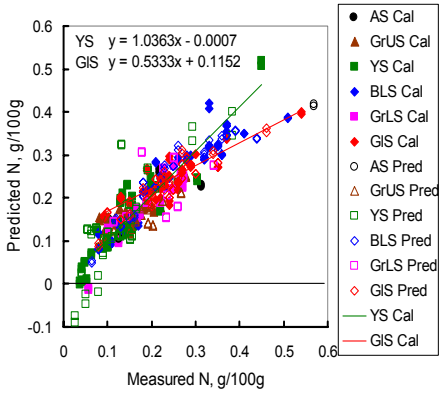

Figure 1. Scatter plots of PCR result for T-N indicating two groups of which regression lines are critically different

Near Infrared Spectroscopy: Proceedings of the 11th International Conference.

www.nirpublications.com

Results and discussion

Factors which cause curvature in regression scatter plot

In Figure 1, it can be observed that two soil types form distributions of which regression lines have critically different angles, i.e. YS and GlS. In the figure, the slope of all calibration samples is 0.822. While the slope of YS is 1.036, for GlS it is 0.533. So, in this report, analyses will be performed mainly for these two soil groups. When curvature of regression plot occurs, there would be three types shown as following figure.

In a case that chemical value of grey line distributes at a lower range and another one (black line) distributes at higher range, shape of regression plot will curve toward upper left direction (a). Opposite case is "c", curving toward down right direction. When chemical values are similar, shape of regression plot will be X shape or just broad. The type of curvature in this report is "a".

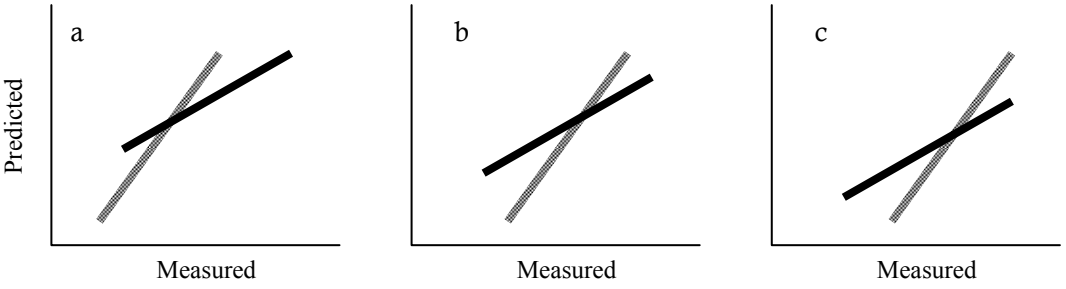

Figure 2. Three types of regression plots that would be formed by two sample groups that have different regression slopes

As to one of the factors causing curvature, the shift of chemical value, range of T-N in YS shifts to the left on abscissa from the GlS by about 0.1 g/100g in Maximum, and 0.04 g/100g in Minimum (Table 1). In NIR-99, besides the author's presentation, some people gave presentations on soil analyses. In many of them similar curvature of regression scatter plots was more or less found.[2,3,4] Cause of the curvatures in those presentation was thought to be able to be explained by the same manner as this report. Direction of curvature was toward upper left same as the author's report in all reports. This means that there are commonly one or some soil groups that contain relatively less chemical constituents and have steeper slope in regression line. However, whether this phenomenon was just coincidence or something necessity was not clarified.

Table 1. Range of Total Nitrogen in each soil group, g/100g

	AS	GrUS	YS	BLS	GrLS	GlS
Max	0.569	0.278	0.449	0.510	0.343	0.540
Min	0.103	0.082	0.027	0.064	0.057	0.080

Then how was the difference of slopes made?

Analyses for causes of the difference of slopes

When PCR results from raw spectra were traced from PC-1, until the number of PCs reached 3, slope of regression line remained almost zero. At PC-3 value of the slope of total samples became 0.13, and then 0.43 at PC-4, afterward the slope became steeper as the number of PCs went larger. As to YS and GlS, at PC-4 the slope of regression lines became apparent for the first time. Also significant difference of the slopes of regression lines appeared here (Figure 3). It seems that this difference continued to be remained to even the final PC where the best validation (prediction) result was obtained.

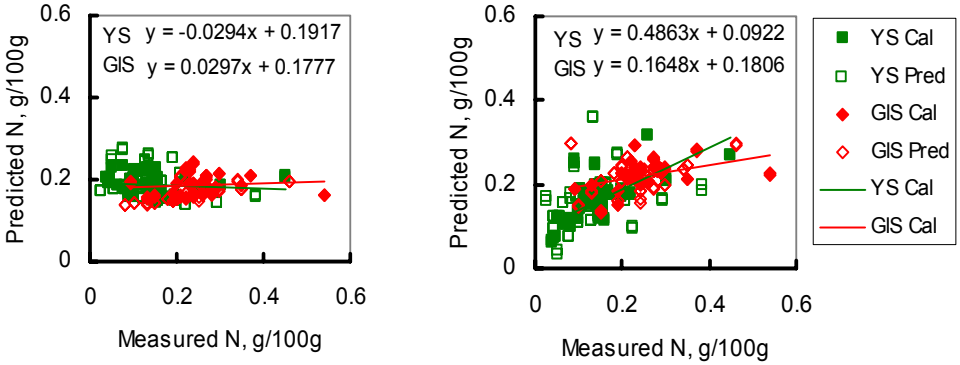

Figure 3. Scatter plots of regression results on T-N for YS and GlS Number of PCs used for calibration, Left: 3, Right: 4

In the previous report,[1] the author pointed out that cluster pattern of each soil group varied in the score-score plot of PC-3 v. PC-4 and this had the relation with "the cause of the curvature". However, what a 2D scores plot can do is to make the difference among objects visual, and it does not necessarily explain the role of scores in making a regression equation.

Then, looking into Figure 4 again, along the axis of PC-4 distribution of YS much larger than that of GlS. As predicted "y" is calculated as the product of y-loading and score, distribution of scores and the values along ordinate have correlation, or they are arranged in the same order. Here, range of the distribution of YS is much larger than that of GlS. On the other hand, ranges of the y-value do not differ so much. As a result, the slopes of regression lines of two soils came to be different, steep YS and almost flat GlS (Figure 3, right). As to PC-3, relationship between scores and chemical values appears to vary

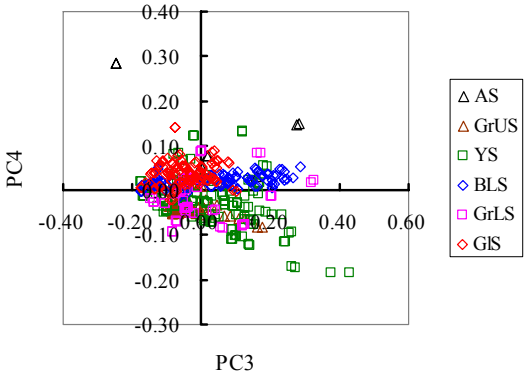

Figure 4. Score-score plots of each soil group on PC-3 v. PC-4

depending upon soil group. For example, in YS they have negative correlation and in GlS positive. Other soil groups may have positive correlation.

Table 2 shows relationships among the scores, y-values (T-N) and y-loadings. At most of PCs of lower number, ranges of the distribution of scores were larger at YS than GlS. It is critical that the PC-4 is the first appearing important PC. At PC-5, the difference of slopes became larger (no figure). Going up to higher PCs condition became more complicated, analyses could not be completed in this report.

Table 2. Maximum and minimum value of scores and the difference between them on each PC, and Y-loading at each PC, for the PCR result using whole spectra range

		YS			GlS		All
	Min	Max	Differ[1]	Min	Max	Differ[1]	Y-load[2]
PC- 1	-1.813	2.803	4.616	-1.648	1.217	2.865	-0.003
PC- 2	-0.433	0.859	1.292	-0.240	0.363	0.603	-0.065
PC- 3	-0.176	0.261	0.437	-0.160	0.102	0.262	0.289
PC- 4	-0.124	0.121	0.246	-0.021	0.067	0.087	1.149
PC- 5	-0.093	0.106	0.199	-0.051	0.047	0.098	0.586
PC- 6	-0.036	0.104	0.140	-0.022	0.017	0.038	-1.004
PC- 7	-0.030	0.034	0.063	-0.016	0.017	0.033	-2.901
PC- 8	-0.052	0.029	0.081	-0.017	0.015	0.033	-0.683
PC- 9	-0.024	0.019	0.043	-0.008	0.010	0.018	1.459
PC-10	-0.010	0.012	0.023	-0.005	0.011	0.016	6.554
PC-11	-0.011	0.016	0.027	-0.022	0.008	0.030	-1.823
PC-12	-0.008	0.006	0.014	-0.010	0.004	0.014	-2.850
PC-13	-0.012	0.013	0.025	-0.006	0.004	0.010	-4.932
PC-14	-0.008	0.007	0.015	-0.005	0.007	0.012	-3.210
PC-15	-0.006	0.004	0.010	-0.005	0.004	0.009	2.153
PC-16	-0.003	0.007	0.100	-0.004	0.003	0.007	0.730
T-N[3]	0.038	0.449	0.412	0.092	0.540	0.448	

1: Difference between Maximum and minimum values of scores
2: Y-loading
3: Value of total nitrogen of samples used for calibration

Trial for improving non-linearity problem

Difference in the slopes of YS and GlS that is one of the causes of the non-linearity in regression plots was clarified that it can be explained by the difference in the range of distribution of scores of two soils. Also, in making the difference of the distribution of scores, PC-4 was thought to act an important role at the beginning of the calibration equation development. Therefore, removal or

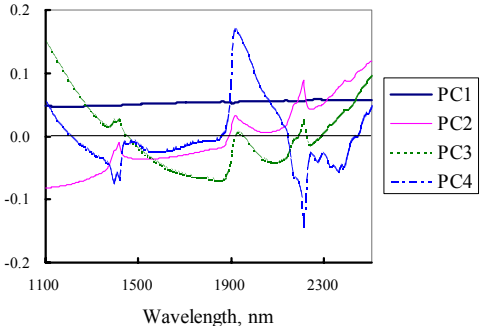

Figure 5. Plots of x-loadings of the first four PCs in the result of Figure 1

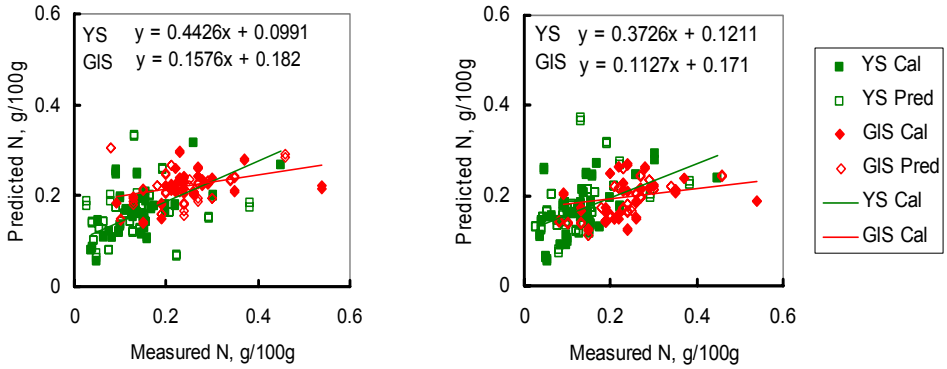

Figure 6. Scatter plots of regression for YS and GIS using four PCs
Left: Negative part of Figure 5 is removed in calibration development
Right: Positive part of Figure 5 is removed

mitigation of the difference in the distribution of scores between YS and GIS was tried.

Looking at x-loadings of PC-4, there are two distinct parts, positive around the region from about 1900 to 2100 nm, and negative part from about 2100 to 2300 nm.

At first, the author thought that negative part should be removed, as the extension of scores of YS to the negative direction was significant.

Result using four PCs is shown in Figure 6 (left). Although the difference of slopes was a little mitigated, it still remains. Then calibration development was again tried by removing positive part of Figure 5. Again, although the difference of the slopes was a little mitigated, it still remains. When final result is seen, non-linearity problem has been visually a little improved although it is not so good. Significant difference would be predictions samples that exceed to minus side on the ordinate in Figure 1. The difference of the slopes of YS and GIS was also a little mitigated as compared with Figure 1. However, this result would not be able to satisfy our needs. And also, this method that remove spectral range that is supposed to be obstacle for calibration development seems to be not effective in order to eliminate the difference of slopes among sample groups.

Table 3 shows the changes of scores and y-loadings at each PC during the development of calibration equation in the same way as Table 2. In the table, "difference" of YS at PC-4 fairly decreased, and y-loading became negative. However, scores or y-loadings of other PCs changed much resulting to show similar non-linearity problem. Deletion of some wavelength range appears to be effective in order to remove or mitigate factors that cause curvature in the scatter plots of regression. However, the effect of it is thought to be limited because he effect seems to be compensated by the changes at other factors. Consequently practical solution would be the calibration development on previously classified sample groups separately.

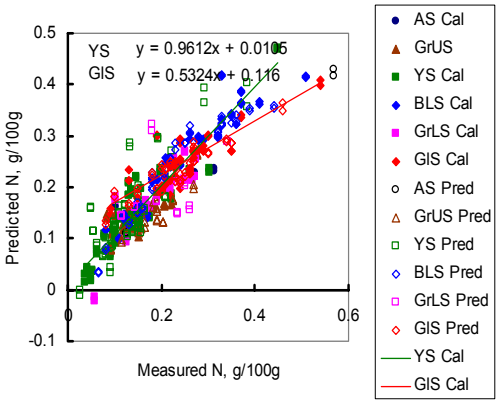

Figure 7. Scatter plots of PCR result of which process spectra range from 1900 to 2050 nm was removed.

Table 3. Maximum and minimum value of scores and the difference between them on each PC, and Y-loading at each PC, for the PCR result removing 1900 – 2100 nm range

	YS			GlS			All
	Min	Max	Differ[1]	Min	Max	Differ[1]	Y-load[2]
PC- 1	-1.730	2.603	4.333	-1.545	1.103	2.648	-0.004
PC- 2	-0.421	0.859	1.280	-0.240	0.355	0.595	-0.069
PC- 3	-0.170	0.247	0.417	-0.155	0.100	0.255	0.316
PC- 4	-0.071	0.117	0.189	-0.066	0.027	0.093	-1.492
PC- 5	-0.140	0.041	0.181	-0.017	0.039	0.056	1.810
PC- 6	-0.060	0.046	0.106	-0.023	0.023	0.046	-0.201
PC- 7	-0.031	0.044	0.074	-0.012	0.017	0.029	1.078
PC- 8	-0.017	0.020	0.036	-0.007	0.011	0.018	-3.543
PC- 9	-0.023	0.018	0.041	-0.011	0.009	0.019	-4.100
PC-10	-0.009	0.012	0.021	-0.012	0.010	0.021	3.244
PC-11	-0.007	0.008	0.015	-0.004	0.009	0.013	3.431
PC-12	-0.009	0.012	0.021	-0.017	0.005	0.023	-0.075
PC-13	-0.011	0.008	0.019	-0.007	0.003	0.010	-6.076
PC-14	-0.005	0.004	0.009	-0.003	0.005	0.008	6.165
PC-15	-0.003	0.004	0.008	-0.002	0.002	0.004	-1.291
PC-16	-0.005	0.003	0.008	-0.004	0.005	0.009	-3.511
PC-17	-0.004	0.003	0.007	-0.001	0.001	0.003	8.360
PC-18	-0.002	0.003	0.005	-0.002	0.002	0.003	17.499
T-N[3]	0.038	0.449	0.412	0.092	0.540	0.448	

1,2,3: See footnote of Table 2.

References

1. Y. Ootake, Near Infrared Spectroscopy: Proceedings of the 9th International conference, 571 (2000)
2. D.F. Malley, P.D. Martin, L.M. McClintock, L Yesmin, R.G. Eilers and P. Haluschak, Near Infrared Spectroscopy: Proceedings of the 9th International conference, 579 (2000)
3. J.B. Reeves, III and G.W. McCarty, Near Infrared Spectroscopy: Proceedings of the 9th International conference, 587 (2000)
4. X. Han, S. Seo, W. Park and R. Cho, Near Infrared Spectroscopy: Proceedings of the 9th International conference, 667 (2000)