Influence of X-leverage on regression result in calibration equation development

Yoshisato Ootake
Komasuya, 2-85 Kiba, Tobishima-mura, Aichi 490-1444 Japan.
E-mail: info@nir-ai-institute.org

Introduction

The author proposed an idea shown in Figure 1 in The First Asian NIR Symposium[1] held in 2008.

The concept is that X-leverage has an important role in calibration equation development.

The author and colleagues have started a new project in which hardware and software are involved in development of a rapid and low cost method for analysis of compost and soil by near infrared spectroscopy. The software for calibration equation development introduces some new methods, based on the above idea. In this presentation the role of X-leverage in calibration equation development is demonstrated, using data from the former study.[2]

Materials and methods

Soil groups used were: Andosol (AS), Grey Upland Soil (GrUS), Yellow Soil (YS), Brown Lowland Soil (BLS), Grey Lowland Soil (GrLS) and Gley Soil (GlS). The total number was 228.

Spectra collection: a BRAN+LUEBBE InfraAlyzer 500 was used. Samples were filled into a diffuse reflectance cup, then spectra were collected from 1100 nm to 2500 nm at 4 nm intervals. Constituents analysed were originally, total nitrogen (T-N), total carbon (T-C), cation-exchange capacity (CEC) and phosphate sorption coefficient (PSC). This report, will focus on the influence of X-leverages on NIRS analysis for total nitrogen. Data analyses was carried out by the chemometrics software "The Unscrambler" Ver. 9.8 (Camo AS, Norway) and "MS Excel". Re-selection of calibration samples based on X-leverages was done after evaluation of the NIRS analysis of all samples in each soil group by cross validation. For comparison, random selection of validation samples in "the Unscrambler" was also carried out.

Results and discussion

Results for Yellow soil are included, although results for this soil type did not show high accuracy in the former study.[2]

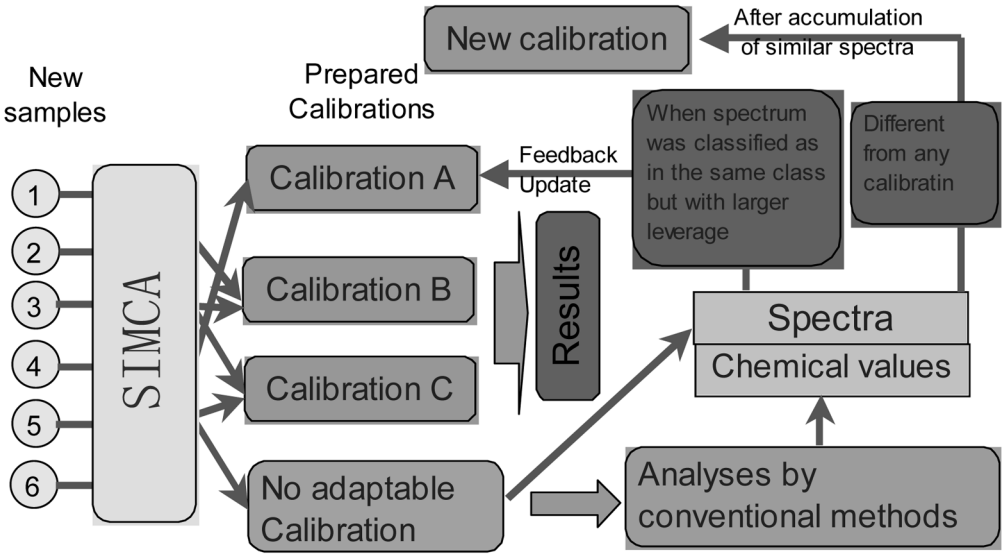

Figure 1. Diagram of adaptation of new samples to prepared calibration equations through SIMCA, and a system for the maintenance and development of calibration equations.

Figure 2 shows the effect of the selection of samples used for the calibration set, and the influence of residuals and leverages of the results of principal component regression (PCR). The upper row shows the results when calibration samples were conventionally selected on the basis of their chemical values, so that samples of which chemical values are larger or smaller are included in the calibration sets, and leverage-residual plots. The number of principal components (PCs) at which the best validation result was obtained was 8. Compared with the following two cases, the indices of accuracy were not good. With regard to the leverage-residual plots the most noticeable point is that the range of residual is larger than that of the other two, and some samples in the validation set were largely apart from calibration samples.

In soft independent modelling of class analogy (SIMCA) residuals correspond to sample-to-model-distance. This means that many spectra of validation samples should be classified to different group than the calibration samples, and this difference is thought to have affected the prediction accuracy.

The middle row of Figure 2 shows the result where "the Unscrambler" selected validation samples randomly. The prediction accuracy was obviously improved as compared to the first series of data. In the residual-leverage plots a remarkable point is the range of residual. It is about 1/30,0000 of that of the above. On the other hand the range of leverage increased to about 2. In this case the spectra of calibration samples and validation samples are much more similar to each other than in the above case.

The lower row shows the result for validation samples of which leverages (calculated by cross validation for all samples) were selected to validation samples by the X-leverage method (samples with larger leverages were selected to the calibration set). Compared to the Unscrambler random

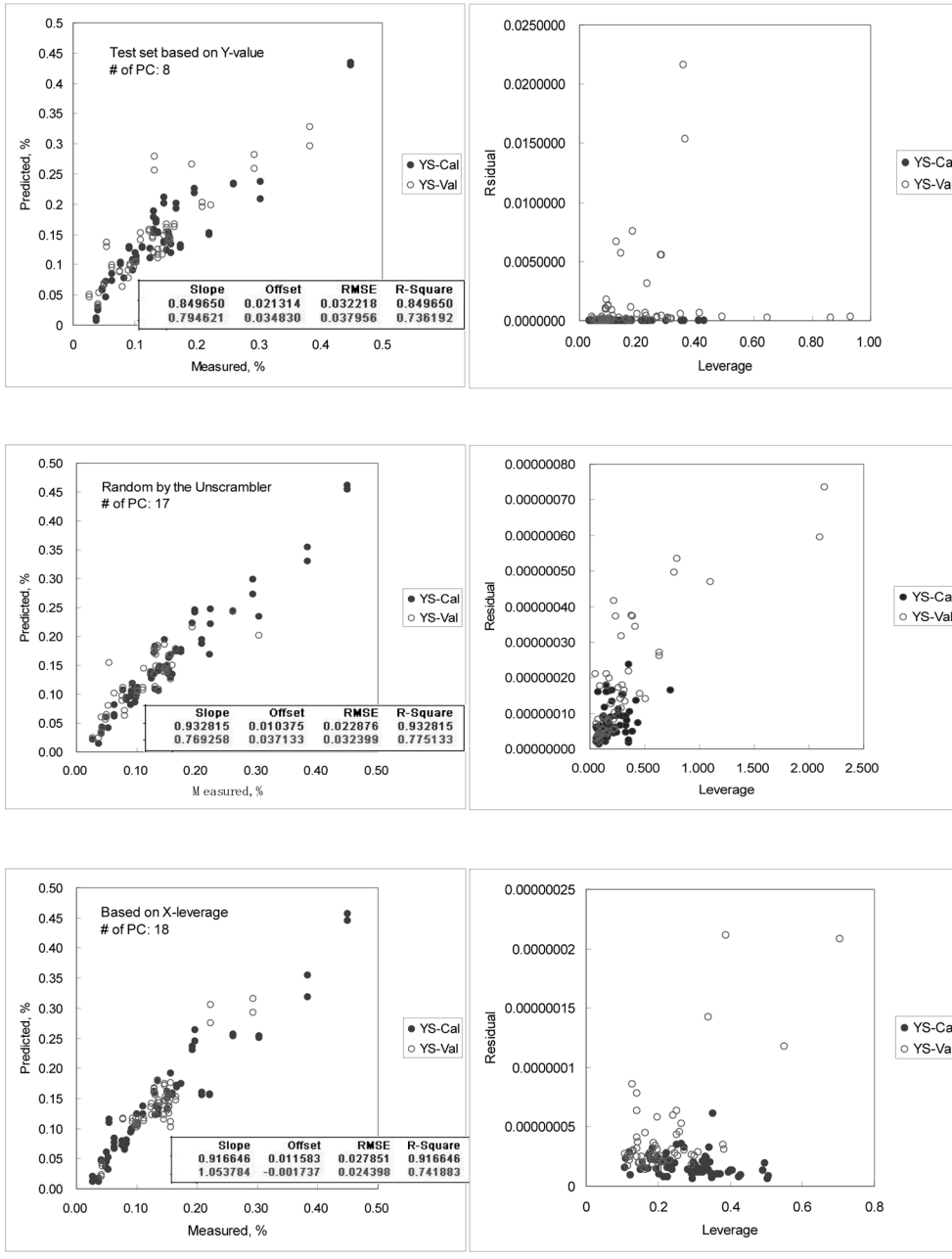

Figure 2. Results showing the effect of X-leverages in calibration equation development showing the relation between regression accuracy and leverage and residual of principal component regression (PCR) for "Yellow soil". Left: Regression plots, Right: Leverage-residual plots; Upper: Calibration samples are selected according to their chemical values; Middle: Calibration samples are selected randomly in "the Unscrambler"; Lower: Calibration samples are selected from samples of which leverages are larger. Spectra used for calibration equation development are raw spectra.

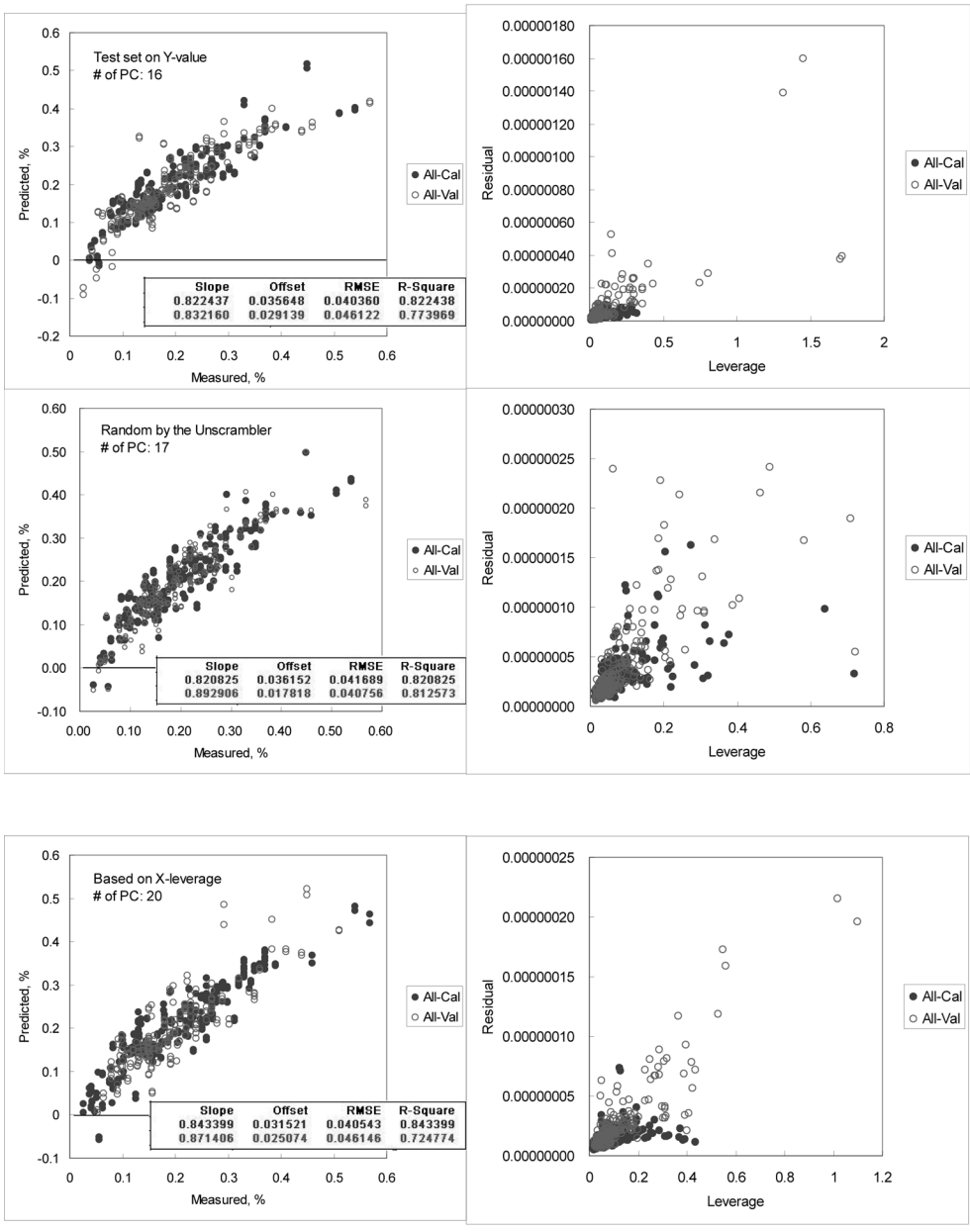

Figure 3. Results showing the effect of X-leverages in calibration equation development showing the relation between regression accuracy and leverage and residual of principal component regression (PCR) for "All soil groups". Left-Right and Upper-Middle-Lower are same as Figure 2. Spectra used are raw same as Figure 2.

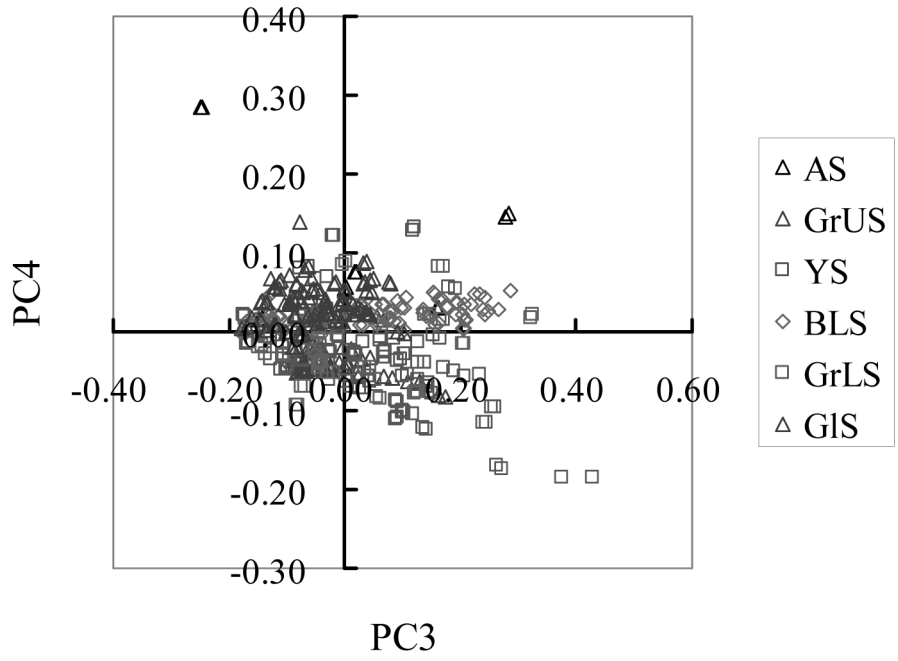

Figure 4. Score-score plots of PCR result of "All soils" above.

selection, the indices of accuracy of the validation set appeared to improve, although those for the calibration set showed no improvement over the Unscramble-selected sample set. The R-squared value was lower than that of the Unscramble series, but the *RMSEP* improved slightly. The reason of this is believed to be because the range of the values of the validation set was smaller than that of the Unscramble series. The residuals and leverages of the X-leverage validation samples ranged up to about 0.0000002 and up to about 0.7 for leverage.

The above results were achieved for the "Yellow soil", but the same trend was apparent when all samples (including Yellow soil) were included by applying the same method, but the effect of the use of leverage in the selection of calibration samples was not so clear as with the "yellow soil" data (Figure 3).

There was some decrease in the value of the residual in the Unscrambler and X-leverage series, but it was not so marked as in the case of the "Yellow soil".

The reason of this difference is thought to be as follows. Figure 4 gives the score plots of PCA analysis for PC3 and PC4. In the figure it is clear that each soil type has different orientation of its distribution.

Based on these data it is believed that the selection of calibration samples according to the X-leverage method value did not have the same clear-cut effect.

The objective of this report was to clarify the role of X-leverage in calibration equation development. The conclusions are that the effect of utilisation of X-leverage is apparent when the spectra are fundamentally similar in the samples used, but not so clear when spectra consist of

different sample types. From now, the work will focus on evaluating the effects of X-leverages in various conditions, using further data treatments, including the use of multiplicative scatter correction (MSC).

Acknowledgement

This study is carried out by the fund provided from *B*io-oriented Technology *R*esearch *A*dvancement *I*nstitution (B R A I N).

References

1. Y. Ootake, *Proceedings of 1st Asian NIR Conference*, p. 274 (2008).
2. Y. Ootake, M. Hioki, H. Tamasu, T. Sato, A. Miyoshi and T. Yoshikawa, *Proceedings of the 9th International Conference*, p. 571 (1999).