# Use of the CONDENSE algorithm to optimise calibration databases

P. Berzaghi,[a] J.S. Shenk[b] and J.W. Shenk[c]
[a]*Department of Animal Science University of Padua, Italy.*
*E-mail: paolo.berzaghi@unipd.it*
[b]*Shenk Analytical International, Port Matilda, PA, USA*
[c]*Unity Scientific, Columbia, MD, USA*

## Introduction

When maintaining a calibration database over the years, new samples are added to include new variability, that may not be accommodated by the prediction model. The criteria to select those samples for expansion of the model may vary, depending on the availability of samples or wet chemistry data, or samples may be selected based on spectral properties. This latter method has been implemented in an algorithm (Center and Select)[1] that uses spectral properties of each sample to define the similarity or uniqueness of each sample. Selection is based on the distances between samples as defined by the H matrix, calculated using spectral principal component analysis (PCA). However, prediction models are generally based on the relationship between spectra and a specific constituent (PLS-1).[2] The lack of consistency between criteria for sample selection and calibration development leads to the selection and the analysis of samples that are unnecessary, increasing costs of processing, reference analysis and NIR scanning, without adding extra useful variance to the model. Selection using PCA may lead to the selection of redundant samples. An algorithm (CONDENSE) was developed that reduces the number of samples by defining redundant samples, using PLS-1 neighbourhood distance (*ND*). Similar samples and the specific constituent are then averaged creating a new "condensed" sample. The study evaluated the performances of a forage database when redundant samples were condensed.

## Materials and methods

An alfalfa hay calibration database developed by the NIRS Consortium from 1993 to 2006 was used. The database consisted of 1047 spectra (Foss 5000) with 667, 766 and 984 chemistry values for CP, ADF and NDF respectively. A validation data set (#39) was compiled using spectra and wet chemistry values obtained during the certification process organised by NFTA from 1999 to 2008. Each year four samples of alfalfa hay are sent to the participating labs. The chemistry values are calculated as a censored average from the results of all the labs that have used NFTA reference methods. Predicting models were developed by Ucal (Unity Scientific) using the original database, and using the database after a condensation to 50 and 30% of the initial number of samples. The condense algorithm evaluates the similarity of samples based on PLS-1 *ND* value.

**Table 1.** Calibration statistics for original and condensed database.

| | | Original | Condensed 50% | Condensed 30% |
|---|---|---|---|---|
| Protein | *N* | 667 | 314 | 188 |
| | SECV | 0.74 | 0.74 | 0.77 |
| | $R^2$ | 0.93 | 0.94 | 0.94 |
| ADF | *N* | 766 | 372 | 219 |
| | SECV | 1.87 | 1.83 | 1.88 |
| | $R^2$ | 0.89 | 0.92 | 0.92 |
| NDF | *N* | 984 | 474 | 279 |
| | SECV | 1.99 | 2.09 | 2.27 |
| | $R^2$ | 0.93 | 0.93 | 0.92 |

When considered redundant, the algorithm merges and averages those samples, maintaining spectral variability, but reducing the number of samples.

## Results and discussion

The database was built over the years using different standardised instruments and including many growing seasons. It also includes spectra from different standardised instruments and wet chemistry from different laboratories. The performance of the calibration in cross-validation (Table1) has a *SECV* that may look large for this product.

The different sources of spectra and chemistry may add a type of noise to the database increasing *SECV* and decreasing *RSQ*, but this type of noise may increase robustness of the model, and using a large number of samples it may not affect overall accuracy.[3] The robustness of the prediction model may be attested to by the fact that the accuracy [*SEP*(C)] for the validation file is actually better than the *SEP* (C) estimated from the *SECV* for all three traits. The database was expanded during the years using PCA *ND* with a threshold of 0.6. Nevertheless, selecting samples with *ND*≥0.6 retained 409 samples, with only 180, 241, and 369 samples having CP,

**Table 2.** Validation using original and condensed database.

| | | Original | Condensed 50% | Condensed 30% |
|---|---|---|---|---|
| Protein | *SEP*(C) | 0.51 | 0.51 | 0.56 |
| | $R^{2-}$ | 0.98 | 0.98 | 0.97 |
| ADF | *SEP*(C) | 1.15 | 1.21 | 1.29 |
| | $R^2$ | 0.94 | 0.93 | 0.93 |
| NDF | *SEP*(C) | 1.37 | 1.36 | 1.48 |
| | $R^2$ | 0.94 | 0.94 | 0.93 |

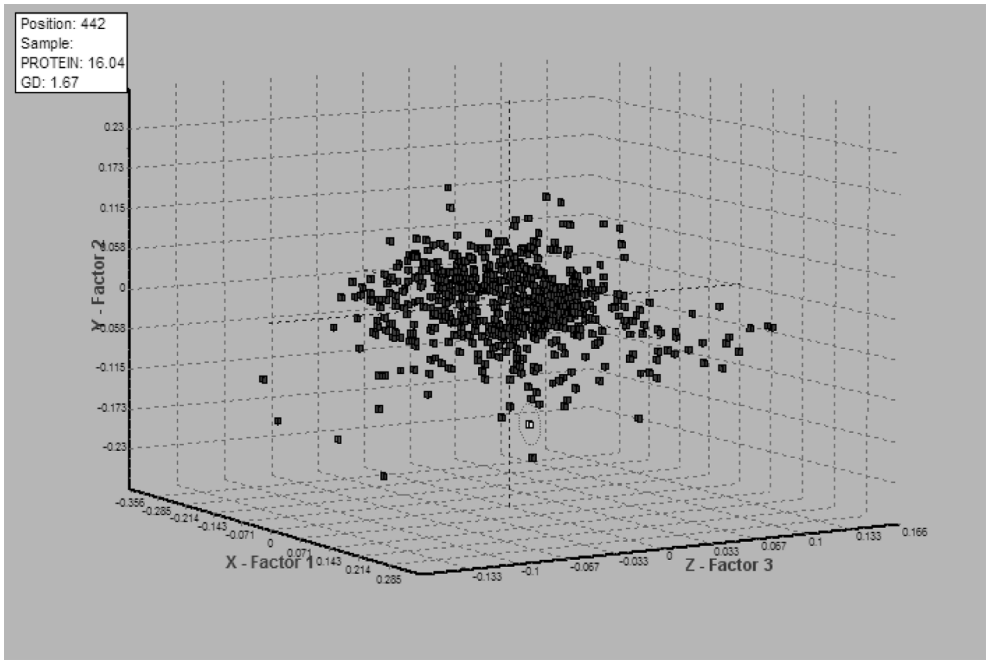


**Figure 1.** 3D PLS1 score plot for protein in alfalfa for the original (#667) database.

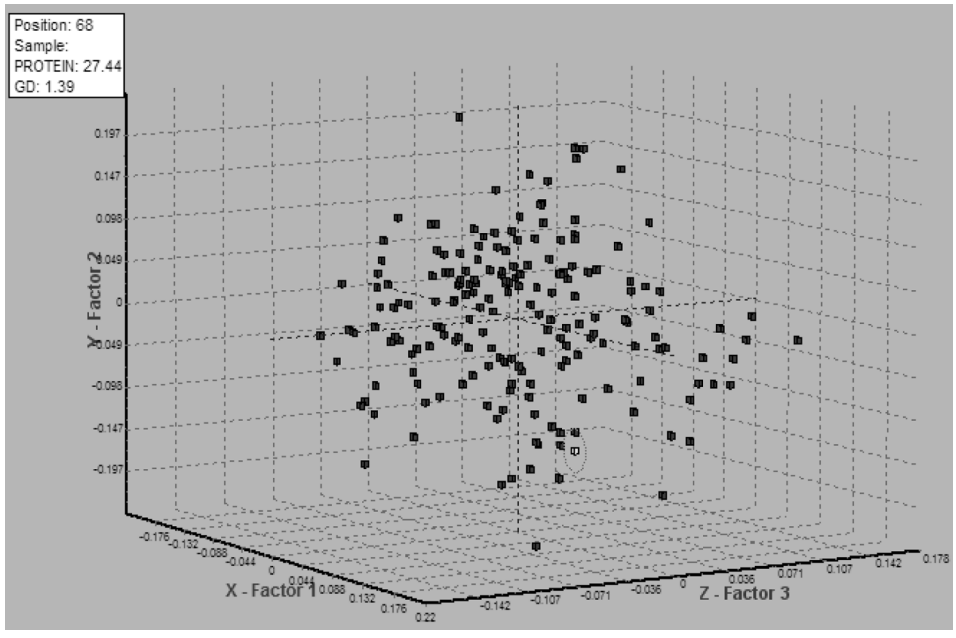


**Figure 2.** 3D PLS1 score plot for protein in alfalfa for condensed (#188) database.

ADF and NDF respectively. The CONDENSE algorithm can be set to aggregate samples by the level of PLS-1 *ND*, or by indicating the percent of reduction in the number of samples. For brevity it was tested by reducing the number of samples with reference values to half and to 30% of the original. In general removing redundant samples improved *SEC* and *SECV*, particularly reducing the original number of samples by one half. Further condensing to 30% of the original samples marginally improved the *SEC* but not the *SECV* (Table 2).

Despite the sharp reduction of samples used in the calibration, the performances of validation [*SEP*(C)] values were similar to those of the original calibrations for all three constituents. The condensed database maintains the original spectral diversity, but it has a better spectral distribution with, equally leveraged samples (Figures 1 and 2).

Also because of the reduced size of the database, it would require fewer samples for updates or expansion, when adaptation to new sources of variability is needed. This is important in networks of instruments, and in agriculture where there might be the need to add new instruments or to account for new growing seasons. In conclusion, despite the sharp reduction in the number of samples of the calibration database, CONDENSE can improve statistics of cross validation and maintain prediction performance in validation. By using the information of neighbourhood distance among samples, CONDENSE can simplify and improve the development of prediction model through a better population structure, reducing the number of samples needed for update and the costs associated with updates.

## References

1. J.S. Shenk and M.O. Westernhaus, *Crop Sci*. **31,** 1548 (2001).
2. J.S. Shenk, P. Berzaghi and J.W. Shenk, *The PLS Database Optimization Concept*, The 14th International Conference of Near Infrared Spectroscopy, Thailand (2009).
3. P. Dardenne and J.A. Fernández Pierna, *J. Near Infrared Spectrosc.* **14,** 6 (2006).