The PLS database optimization concept

J.S. Shenk,[a] Paolo Berzaghi[b] and J.W. Shenk[c]
[a]*Shenk Analytical International LLC, Port Matilda, PA, USA.*
E-mail: jsshenk@comcast.net
[b]*Padova University, Padova, Italy*
[c]*Unity Scientific, Columbia, MD, USA*

Introduction

Properly structured product databases are essential for accurate NIR analysis. The partial least squares (PLS) algorithm calculates constituent specific parameters (loadings and scores) that are used to calculate a Mahalanobis-scaled Neighbor Distance (ND) between any two spectra. The ND values can then be used to optimize the calibration database by minimizing sample redundancy and maximizing sample diversity. The purpose of this study was to show how ND can be used interchangeably to select samples to build a product database, or it can be used to condense a historic product database.

Experiment 1

Materials and methods

A file of spectra containing 1231 hay samples was collected over a 10-year period and used for this study. The database contained grass, legume, and mixed samples of hay from a broad geographical area ranging from Canada to Florida, and from the Mississippi River to the Atlantic Ocean. Each constituent was separated out into its own database, so that selection and condensing could be optimized for each constituent. To build the product database with ND, the file was divided into 8 groups of 50 samples and the 9th group was the remaining 831 samples. The ND threshold for selecting both protein and ADF samples was set at 1.0. The software used to carry out these 2 experiments was UniStar, provided by Unity Scientific, Columbia, MD.

Results and discussion

Table 1 displays the number of samples selected from each set of samples for both protein and ADF.

The first set of 50 protein samples was used to select samples from the second set of 50 samples. The ND statistic for protein selected 40 samples from the second set of 50. The ND statistic for ADF selected 33 samples from the second set. Combining the first 50 samples with the second set

Table 1. Selection of protein and ADF with an ND threshold of 1.0.

Subset	Protein		ADF	
	# Added	# Remain	# Added	# Remain
50a	50	1181	50	1181
50b	40	1131	33	1131
50c	33	1081	34	1081
50d	0	1031	0	1031
50e	0	981	0	981
50f	10	931	8	931
50g	5	881	5	881
50h	1	831	0	831
831i	49	782	51	780
Total	188		181	

of samples within each constituent was used to make a calibration to select samples from the third set of 50 and so on. The remaining samples at the end were used as an independent set.

This selecting and combining continued until 188 samples were selected for protein and 181 samples were selected for ADF. Having built the databases for each constituent, the complete file of 1231 samples was condensed to 188 samples for protein and 181 samples for ADF (Berzaghi *et al*.[1]). This provided three databases for comparisons: the original file of 1231 samples, the file built with upward selected ND samples (UpND), and the file of condensed ND samples (CondND). Comparing the actual samples selected for upward selection of both protein and ADF, there were 121 common samples selected from this process. More than 60 samples were different. This confirms the fact that to get the best accuracy, each constituent needs to be selected separately to obtain the optimum database for each constituent.

The distribution of all samples for both protein and ADF represented a typical Gaussian distribution. The range, average, and standard deviation for each population of samples is found in Table 2.

The minimum, maximum and average of each file was similar to the original file; however, the distribution of the UpND and CondND was considerably thinned out and flattened.

Table 2. Simple statistics describing the three files.

File	Protein					ADF				
	Samp.	Min	Max	Ave	SD	Samp.	Min	Max	Ave	SD
All	1231	1.85	32.5	17.14	5.1	1231	12.6	61.71	35.19	6.5
UpND	188	1.85	30.7	16.25	5.5	181	16.9	61.71	35.88	7.7
CondND	188	2.48	32.5	15.83	6.5	181	12.6	59.29	35.47	8.4

Samp. = Samples, Min = minimum constituent value, Max = maximum constituent value, SD = standard deviation of the constituents.

Table 3. Calibration statistics for the four populations of samples.

	Protein				ADF			
	Samples	Bias	*SECV*	*RSQ*	Samples	Bias	*SECV*	*RSQ*
All	1231	0.01	0.65	0.99	1231	–0.06	1.20	0.99
UpND	188	0.06	0.70	0.99	181	–0.16	1.18	0.99
CondND	188	0.00	0.61	0.99	181	0.00	1.22	0.97
Independent*	733	0.01	0.74*	0.97	729	0.04	1.25*	0.90

*Independent samples from UpND selection for Protein were 733 (782-49, Table 1) and 729 (780-51, Table 1) for ADF.

SECV = standard error of cross validation and *RSQ* = fraction of explained variance, * = *SEP*.

The calibration statistics for the protein and ADF constituents are shown in Table 3.

The standard error of cross-validation (*SECV*), bias and *r*-square values were very similar for these populations.

Standard error of prediction

The independent test of the remaining samples shows good agreement with the *SECV* errors for the original and UpND and CondND populations, but the increase in *SEP* suggest that there were additional samples in the independent set that should have been selected from the independent set, or the ND threshold may have been a little high.

Experiment 2

Materials and methods

In a separate test to look at principle component analysis (PCA) selection, the same mixed hay database containing 200 samples of protein, and ADF was used to make the PLS calibration. One hundred independent mixed hay samples were used to test which samples were needed to expand the calibration, using an ND of 1.0 for both PCA and PLS.

Table 4. Selecting samples with PCA and PLS ND.

Constituent	Cal. set	Val. set	Selected samples PCA	Selected samples PLS
Protein	200	100	34	5
ADF	200	100	34	18
Total			68	23

Cal = Calibration set of samples, Val = Validation set of samples.

Results and discussion

Table 4, shows that in a simple test of the first 200 samples in the hay file validated against the next 100 samples, that 68 samples would have been selected for laboratory reference values with PCA ND; whereas, only 23 samples were selected to be analyzed with PLS ND.

Conclusion

There are 2 reasons why PLS database management of product constituent databases with ND is important. Selecting the right samples for the database is essential to build a product database. Expanding the database with these samples can improve and maintain its accuracy. If the product database contains thousands of samples, condensing is a good procedure to clean up the database and remove redundancy. It is often best to start with a good, condensed database and expand it with new samples rather than build the database from the beginning. Both of these procedures (selection and condensing) are easily accomplished with the ND statistic.

Reference

1. P. Berzaghi, J.S. Shenk and J.W. Shenk. *Proceedings of the 14th International Conference of Near Infrared Spectroscopy*, Bangkok, Thailand, (2009).