Noise robustness comparison for merging large datasets

S.A. Roussel,[a] C.R. Hurburgh, Jr. [b] and D.B. Funk[c]

[a] *AGROMETRIX. Cemagref, 361, rue JF Breton, BP5095, 34033 Montpellier, CEDEX 1, France. roussel@agrometrix.com*

[b] *Iowa State University Grain Quality Laboratory. 1563 Food Sciences Building, Ames, IA 50011, USA. tatry@iastate.edu*

[c] *USDA Grain Inspection, Packers and Stockyards Administration, 10383 N. Executive Hills Blvd, Kansas City, MO 64153, USA*

Introduction

Near Infrared (NIR) spectroscopy is commonly used to assess the composition of whole grains in a rapid and non-destructive way. Thus, large datasets have been gathered through the years on several NIRS instruments. Those datasets include different types of grain samples with regards to varieties, crop years, environmental/weather conditions, etc. Accurate generic calibration models require that these datasets be merged.

However, before combining the data coming from different NIR instruments into a single larger set, the optical differences between spectra must be assessed and corrected if required. Inter-instrument measurement variations can be considered as noise in the spectra.

The objective of this paper is to quantify the *noise robustness* of different calibration models and determine noise stability areas where optical standardisation (before calibration) is not required before merging spectral datasets. The concept of *model robustness* is not straightforward. Several definitions of robustness can be found in the literature, but none has been issued by the International Standard Organisation. The notion of robustness is the capacity of a model to remain stable under small perturbations.[1] The second objective of the paper is to design a simulation procedure to assess the robustness of multivariate models.

Experimental

Whole corn samples were scanned in transmission by Infratec Grain Analyzers, which are monochromator-based near infrared spectrometers manufactured by FOSS/Tecator (Höganäs, Sweden). The spectra contain 100 wavelengths in the 850–1050 nm range, with a 2 nm resolution. The Grain Quality Laboratory of Iowa State University (ISU-GQL), Ames, Iowa, has been collecting corn samples at harvest, during the years 1987–2002. Thus, the corn database comprises the variations in genetics, geography, growing conditions and physical characteristics of US Middle-West corn. Reference values for the moisture content were obtained by air-oven method,[2] performed by ISU-GQL. The corn moisture range was restricted to [8%; 25%], with an average of 14.7% and a standard deviation of 3.8%. The moisture database consisted of 3699 samples in total, split into a 3289 sample calibration set and a 410 sample test set, using the venetian-blind technique. The moisture distribution of both sets was similar.

Theory

Different types of noise

The different types of noise that can be observed in NIT measurements were defined in cooperation with the Grain Inspection, Packers and Stockyards Administration from the US Department of Agriculture (GIPSA–USDA). In order to test the model robustness, the seven identified noises were simulated and introduced in the test set:

1. random noise (RND): uncorrelated Gaussian noise was simulated by generating normally-distributed values between 0 and 0.01% OD RMS
2. multiplicative noise (MLT): the spectra were multiplied by a constant to simulate path length variations
3. baseline shift (BLS): a baseline offset was added to the spectra to simulate wavelength-independent gain variations. Wavelength shift (WLS): Monochromator wavelength-axis shift was simulated by shifting the spectra after a Spline interpolation between the original spectral wavelengths
4. wavelength stretch or shrink (WLSt): monochromator wavelength-axis stretch/shrinkage was simulated by re-sampling the spectra after a Spline interpolation, maintaining the centre of the spectra at 950 nm

Stray light (STL): The effects of stray light can be simulated by transforming the individual components of an absorbance spectrum (A) using the following equation:

$$A_{noisy} = -\log\left(10^{-A} + Straylight * \overline{T}/100\right) \quad (1)$$

5. Where A is the absorbance spectrum, and $\overline{T}$ is the mean transmission value of the spectrum. Bandwidth variations (BDW), simulated by convolving the spectra with functions to broaden or sharpen the spectra. The bandwidth function is assumed to be Gaussian and the result of the difference between two different functions is used to compute the noisy spectra.

Experimental designs

Two experimental designs were applied to assess and compare the model robustness. First, a full 2-level factorial design was carried out to identify the noises that have a significant influence on model standard error of prediction (*SEP*). 128 (7^2) simulations were carried out, using the noise levels shown in Table 1. The confidence interval was computed using σ estimated by bootstrapping.[3] Second, a surface response design was performed with the significant noises, to determine the noise stability areas for each calibration model as well as assess the behaviour of the model performance in every point of the noisy domain.

Table 1. Factor levels for the 2 experimental designs.

	Full factorial screening design		
Noise type	**Minimum**	**Centre point**	**Maximum**
1. Random (RND)	–1 OD RMS [(a)]	0 OD RMS	1 OD RMS
2. Multiplicative (MLT)	–0.2 (=26 mm) [(b)]	0 (=30 mm)	0.2 (=34 mm)
3. Baseline shift	–1 OD [(c)]	0 OD	1 OD
4. Wavelength shift	–1 nm	0 nm	1 nm
5. Wavelength stretch	–1 nm	0 nm	1 nm
6. stray light	0	0	1%
7. Bandwidth	6nm	FWHM[(d)]=7nm	8nm

(a) Optical Difference Root Mean Square
(b) simulated path length
(c) Optical Difference
(d) Full Width at Half Maximum

Models

- Multivariate models were built on 3700 calibration near infrared spectra to predict the moisture content of whole corn kernels: linear multivariate models: partial least squares (PLS)[4] regression with three different pre-processing techniques: (1) mean-centred spectra (PLS), (2) standard normal variety [5] (*SNV*-PLS) to remove scattering effects, and (3) selecting a subset of wavelengths using genetic algorithms [6] (GA-PLS), since the more parsimonious the model, generally the more robust;[7]
- local multivariate models: locally weighted regression (LWR)[8], with mean-entered spectra (LWR) and SNV pre-processed spectra (SNV-LWR);
- Non-linear multivariate models: Three-layer feed-forward artificial neural networks (ANN)[9] were trained using error-gradient back-propagation algorithms and dynamic learning. The inputs were the first principal components computed on the normalised spectra. In order to avoid overfitting, two training strategies were applied: (1) a *weight-decay method*,[10] called *regularisation learning,*[11] that tries to force the weights towards zero to *smooth* the neural networks (r-ANN) and (2) a *pruning technique,*[12] that iteratively pruned the least important neurons to optimise both the network structure and its generalisation performance (p-ANN).

Robustness index

The robustness index represents the stability area of the model. It is computed as follows:

Robustness index (RI) : surface where $SEP < 2 \, . \, SEP_{\text{original}}$ (2)

This is then the area in which the determined noises would not create more than twice the original standard error. The important point is that the broader the area or RI, the more stable is the model, and therefore the more desirable for use in large databases with larger noise potential.

Results

Full factorial screening design Significant effects

Figure 1 is a "Pareto chart", showing the effect of the individual noises by model type. The *y* axis is the average effect of each factor (noise) on the *SEP*, computed using the screening design results. The green horizontal lines show the confidence interval (computed by bootstrapping), to determine if the noise effects are significant. Figure 1 (h) shows the noise influence in average for all the models. Among the seven noises tested, three proved to have a significant influence on the overall model performances : the baseline shift (BSL), wavelength shift (WLS), and multiplicative noise (MLT), adding more than 0.1% point moisture to the *SEP* ($SEP_{\text{original}} \approx 0.4\%$ of moisture) .

Model comparison

ANN were more sensitive to baseline shifts [Figure 1.(d,e)], whereas PLS and LWR were more significantly influenced by the multiplicative noise and the wavelength shift [Figure 1.(a,c,f)].

SNV pre-processing makes the models insensitive to multiplicative effects and baseline shifts [Figure 1.(b,g)], by construction, as explained in the following equations:

$$SNV(x) = \frac{x - \bar{x}}{\sigma(x)} = SNV(x + \text{constant}) = SNV(x \times \text{constant}) \quad (3)$$

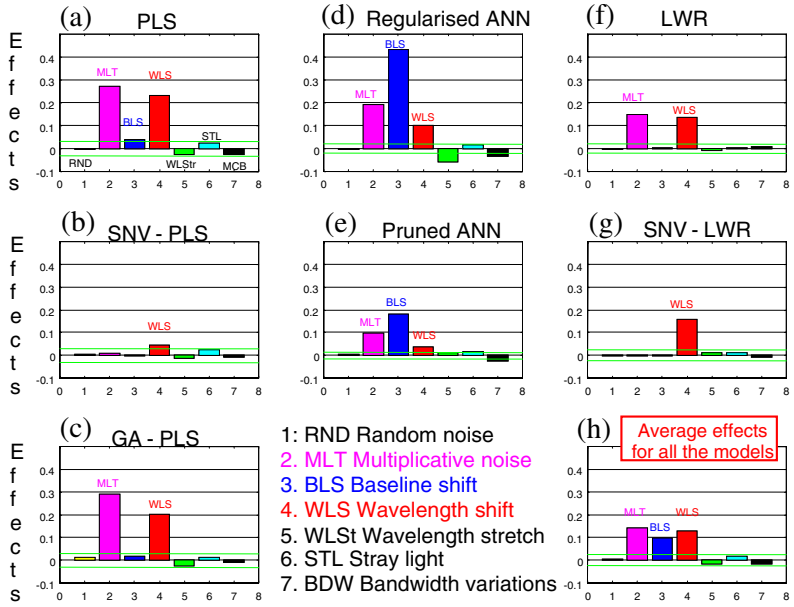

Figure 1. Pareto charts : average effects of the 7 noise types on the model SEP

Surface response design

The three significant noises—baseline shift, wavelength shift, and multiplicative noise—were tested using a surface response design, as shown in Figure 2. In this case, the experimental design output (*SEP*) was also computed in six "star points", in order to model the *SEP* in the entire noisy domain.

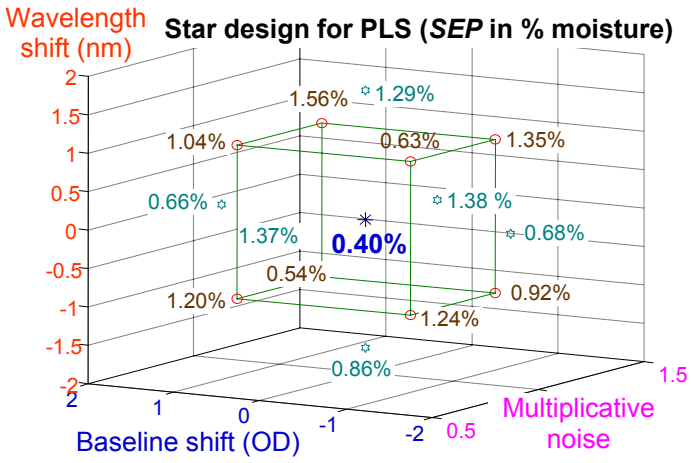

Figure 2. Surface response design : star design results for PLS model.

The behaviour of the multivariate model performance can be modelled within the area of interest for noise variations (Figure 3), based on the surface response design simulations. The dashed area corresponds to the Robustness Index computation, where the *SEP* is lower than twice the original *SEP*.

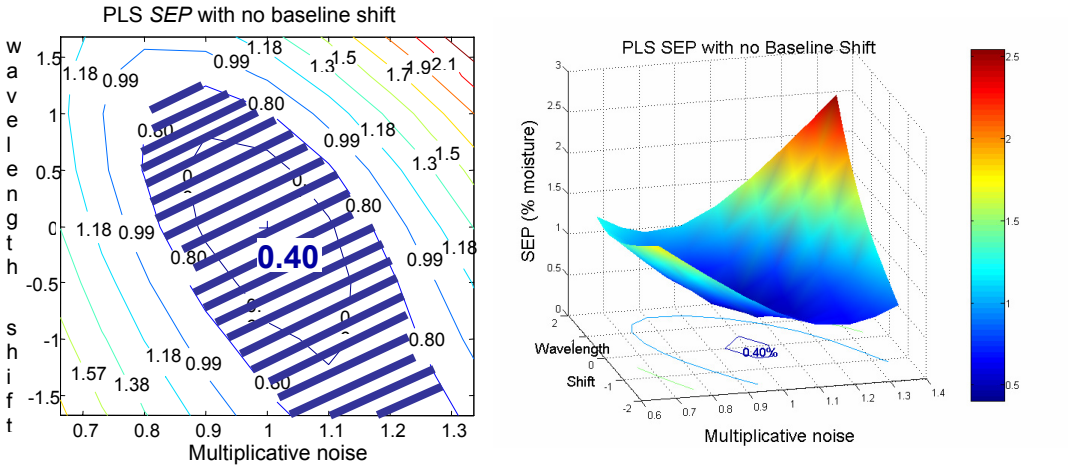

Figure 3. PLS model performance based on the surface response design simulations.

Model comparison

Figure 4 compares the robustness index of all the models; the higher the index, the more robust the models. The *SNV* pre-processed models are clearly the most robust ones. PLS model robustness can be highly improved by a post-regression bias correction (dashed bars in Figure 4), whereas the lack of robustness of ANN and LWR models are mainly due to a pure scattering (no post-regression correction improvement).

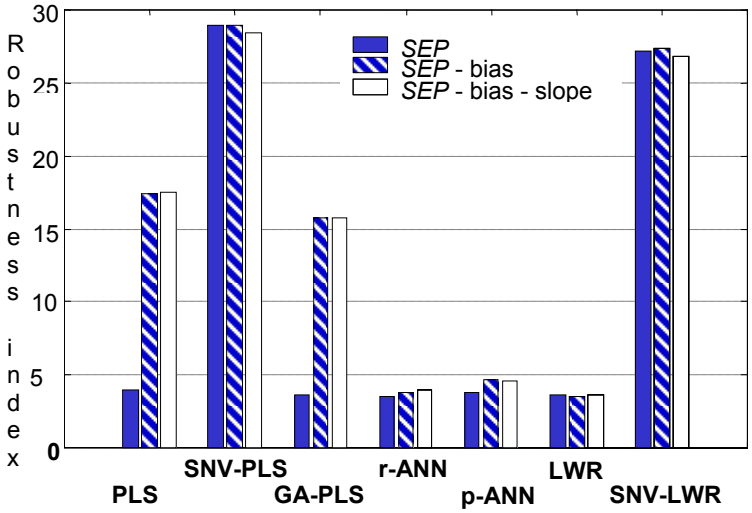

Figure 4. Robustness index comparison for all the calibration models.

Conclusions

This method relies only on simulations to provide a direct assessment of the model robustness, when dealing with noisy spectroscopic measurements. The experimental design procedure helps determine the noises which have a significant effect on model performances, using as lower simulations as possible. Furthermore, surface response designs are able to provide 2^{nd} order polynomial modelisation of the model performance behaviour within the noisy area. When merging spectral data coming from various instruments, the differences between these spectra correspond to the types of noise described in this study. Thus, the computation of the robustness index as well as the examination of the stability area is a decision-support system, allowing the user to decide whether an optical standardisation is required before merging datasets. It also helps identify the most suitable pre-processing techniques and models to apply to the merged dataset, depending on their robustness towards the main sources of noise.

References

1. Y. Vander Heyden, A. Nijhuis, J. Smeyers-Verbeke, B.G.M. Vandeginste and D.L. Massart, *J. Pharma. Biomed. Anal*. **24(5/6),** 723 (2001).
2. FGIS, *Moisture Handbook*, Chapter 4. Federal Grain Inspection Service, Washington, DC, USA (1986).
3. N.M. Faber, *Chemom. Intell. Lab. Syst*. **49(1),** 79 (1999).
4. P. Geladi and B.R. Kowalski, *Anal. Chem. Acta* **185,** 1 (1986).
5. R.J. Barnes, M.S. Dhanoa, , S.J. Lister, *Appl. Spectrosc*. **43,** 772 (1989).
6. D. Jouan-Rimbaud, D.L. Massart, R. Leardi, and O.E. de Noord, *Anal. Chem*. **67,** 4295 (1995).
7. M.B. Seasholtz and B. Kowalski. *Anal. Chim. Acta* **277,** 165 (1993).
8. Z. Wang, T. Isaksson, and B.R. Kowalski, *Anal. Chem*. **66,** 294 (1994).
9. E. Fiesler and R. Beale, *Handbook of Neural Computation*. IOP Publishing Ltd and Oxford University Press, Inc. (1997).
10. A. Frogh and J.A. Hertz, in *Advances in Neural Information Processing Systems*, 4, Ed. by J.E. Moody, S.J. Hanson and R.P. Lippman. Morgan Kaufman Publishers, San Mateo, CA, USA, p. 102 (1992).
11. F. Girosi, M. Jones,and T. Poggio. *Neural Comp*. **7,** 219 (1995).
12. G. Thimm , E. Fiesler E, in *IDIAP-Research Report 97-03*, Dalle Molle Institute for Perceptive Artificial Intelligence, Martigny, Valais, Switzerland (1997).