(Article begins on next page)

Minimisation of instrumental noise in the acquisition of FT-NIR spectra of bread wheat using experimental design and signal processing techniques

Journal:	*Analytical and Bioanalytical Chemistry*
Manuscript ID:	ABC-01329-2010
Type of Paper:	Original Paper
Date Submitted by the Author:	20-Jul-2010
Complete List of Authors:	Foca, Giorgia; Universita di Modena e Reggio Emilia, Dipartimento di Scienze Agrarie e degli Alimenti Ferrari, Carlotta; Universita di Modena e Reggio Emilia, Dipartimento di Scienze Agrarie e degli Alimenti Sinelli, Nicoletta; Universita di Milano, Dipartimento di Scienze e Tecnologie Agrarie e Microbiologiche Mariotti, Manuela; Universita di Milano, Dipartimento di Scienze e Tecnologie Agrarie e Microbiologiche Lucisano, Mara; Universita di Milano, Dipartimento di Scienze e Tecnologie Agrarie e Microbiologiche Caramanico, Rosita; CRA, Unità di Ricerca per la Selezione dei Cereali e la Valorizzazione delle varietà vegetali Ulrici, Alessandro; Universita di Modena e Reggio Emilia, Dipartimento di Scienze Agrarie e degli Alimenti
Keywords:	Chemometrics / Statistics, Agriculture, Foods / Beverages, IR spectroscopy / Raman spectroscopy, Spectroscopy/Instrumentation, Software

Minimisation of instrumental noise in the acquisition of FT-NIR spectra of bread wheat using experimental design and signal processing techniques

G. Foca[a], C. Ferrari[a], N. Sinelli[b], M. Mariotti[b], M. Lucisano[b], R. Caramanico[c], A. Ulrici[a*]

[a] Dipartimento di Scienze Agrarie e degli Alimenti, Università degli Studi di Modena e Reggio Emilia, Padiglione Besta, via Amendola 2, 42100 Reggio Emilia, Italy.

[b] Dipartimento di Scienze e Tecnologie Agrarie e Microbiologiche, Università di Milano, via Celoria 2, 20133 Milano, Italy.

[c] CRA - Unità di Ricerca per la Selezione dei Cereali e la Valorizzazione delle varietà vegetali, Via Forlani 3, 26866 S.Angelo Lodigiano (Lodi), Italy.

* *alessandro.ulrici@unimore.it*

Abstract

Spectral resolution (R) and number of repeated scans (S) have a significant effect on the S/N ratio of FT-NIR spectra, but the optimal values of these two parameters have to be determined empirically for a specific problem, considering separately both the nature of the analysed matrix and the specific instrumental setup. To this aim, the instrumental noise of replicated FT-NIR spectra of wheat samples was modelled as a function of R and S by means of the Doehlert design. The noise amounts in correspondence to different experimental conditions were estimated by analysing the variance spectra derived from replicate measurements with two different signal processing tools, Savitzky-Golay filtering and Fast Wavelet Transform, in order to separate the "pure" instrumental noise from other variability sources, which are essentially connected to sample inhomogeneity. Results confirmed that R and S values leading to minimum instrumental noise can vary considerably depending on the type of analysed food matrix and on the different instrumental setups, and helped in the selection of the optimal measuring conditions for the subsequent acquisition of a wide spectral dataset.

***Keywords* -** FT-NIR, Doehlert design, noise removal, Fast Wavelet Transform, Savitsky-Golay filtering

Introduction

Over recent years, we have worked on the development of classification procedures to predict the technological quality of bread wheat as a whole, based on NIR spectra measured on flour [1-3]. The results obtained until now, however, highlighted the complexity inherent to this kind of application. In fact, many pieces of information reflecting different wheat quality-related aspects have to be extracted from the NIR spectra, and the contribution of some of these relevant features can be very low and hidden into a relatively large amount of not pertinent information and noise. Thus, before proceeding to the sampling of a wide dataset to be analysed by means of different multivariate classification methods, including the use of feature selection procedures, it is extremely important to define objectively the optimal instrumental measuring conditions, in a manner to gain spectra as much informative as possible. This aim can be reached to a great extent by improving the S/N ratio, also considering that the specific nature of the food matrix to be analysed and the particular instrumental setup have their own influence.

It is well known in the literature [4] that significant variations on the S/N ratio of FT-NIR spectra can be induced by varying the spectral resolution (R) and on the number of repeated scans (S). The spectral resolution is connected with the frequency of sampling of the instrumental signal along its spectral range: the higher is the number of sampled points, the higher the resolution of the signal. However, the increase of the number of sampled points is a double-edge weapon, since together with the signal resolution, also the sampled instrumental noise (which generally has high frequency) increases.

On the other hand, the NIR spectrum acquired by means of a FT spectrophotometer is obtained by averaging a number of successive scans of the sample. As a consequence, by increasing the number of repeated scans, S, it is generally known that the instrumental noise decreases of a factor equal to the square root of S. Unfortunately, the choice of the possible highest number of repeated scans is not always convenient, since the time necessary for the analysis increases too, hence a proper compromise has to be found.

Usually, the optimal values of R and S are defined on the basis of the experience and sensitivity of the analyst who, for each specific problem, choses the best values of these parameters in an empirical way, fixing them *a priori* or by a fast trial-and-error procedure, which is generally based only on the visual inspection of the resulting spectra.

At variance, in the present research work, we have decided to define objectively the optimal values of the acquisiton parameters, by extracting the amount of instrumental noise with proper signal processing techniques and then by modelling it as a function of R and S. In addition, we have also considered different operative conditions, i.e., different physical forms under which the sample (bread wheat) has been analysed (i,.e., grit, wholemeal flour and white flour), and different instrumental techniques (i.e., different instrument accessories) for signal acquisition. In order to find the conditions corresponding to the minimum value of noise, this was modelled as a function of the factors R and S using the Doehlert experimental design.

The noise values were estimated by analysing the variance spectra obtained from replicate measurements. Since different variability sources contributed to the variance spectra, including not only the high-frequency instrumental noise (whose minimisation was the aim of the present work), but also lower-frequency components, which are essentially due to sample inhomogeneities, two different signal processing techniques, i.e., Savitzky-Golay smoothing filters and Fast Wavelet Transform were used to isolate and to extract the “pure” instrumental noise component from the “whole” variance spectra. Then, multilinear regression was applied to model the instrumental noise; this allowed to build response surfaces that were used to locate, within the experimental domain, the operating conditions (R and S values) leading to the minimum noise values in correspondence to each kind of matrix analysed by each instrumental apparatus.

Material and methods

Sampling and instrumentation

Two different bread wheat types were analysed in this work, which were selected among 153 different types within the Italian bread wheat production of the harvesting year 2007. The samples were selected on the basis of their intermediate values for the main technological parameters that are usually considered to define wheat quality [5]. Selection was performed with the help of a PCA model calculated on the matrix composed by 322 replicate measurements of the 153 different bread wheat samples, for which the values of the following seven parameters have been measured: Hectolitre weight, Falling Number, Protein content, Alveographic W, Alveographic P/L, Stability and sodium-dodecyl sulphate (SDS) sedimentation volume. In particular, samples selection was done considering the Q vs. T^2 plot of a 4 PCs model (85.53% explained variance) calculated on autoscaled data (Fig. 1). Two samples were chosen as representative of the whole dataset, belonging to two different quality categories, within the 95% confidence limits of the PCA model: one sample showing low Q and intermediate T^2 values, and the other one showing low T^2 and intermediate Q values.

The samples were submitted to NIR analysis under three different physical forms: grit, wholemeal flour and white flour, in order to consider all the different forms under which the wheat is usually analysed by NIR spectroscopy in the cereal manufacturing process. The analyses were performed in two independent laboratories equipped with two spectrophotometers from the same manufacturer and model, i.e., two Bruker MPA Multi Purpose FT-NIR Analyzer spectrophotometers. Each laboratory acquired the spectra on a different sample, in order to verify the possibility to obtain reliable results even on samples having different compositional characteristics. The two FT-NIR instruments were equipped both with three sampling tools: Fiber Optic probe (FO), Integrating Sphere (IS) and Transmission Unit (TU). In order to lower the variability due to possible differences in the optical path passed by the radiation, the weight of each aliquot of sample under examination was set at a constant value, between 10.0 and 80.0 g, depending on the physical form of the analyzed sample and on the sampling tool used for spectral measurements.

Doehlert experimental design

Among other possible experimental designs, the Doehlert design [6] was chosen since it allows to consider different levels for each factor (in particular, in our case five levels for S and three levels for R in an hexagonally-shaped domain) leading to the possibility to model the response function with a quadratic surface, based on only seven different experimental conditions. For our two factors S (number of scans) and R (resolution), the postulated model is:

$$\mathrm{IN} = b_0 + b_1\,\mathrm{S} + b_2\,\mathrm{R} + b_{11}\,\mathrm{S}^2 + b_{22}\,\mathrm{R}^2 + b_{12}\,\mathrm{S}\,\mathrm{R} \qquad \text{(eq. 1)}$$

where the response IN stands for Instrumental Noise.

The five levels for S were equally spaced in the interval 20 ÷ 200 scans (Fig. 2). As it was previously stated, the higher is the number of scans, the lower the noise; however, we have chosen 200 scans as the highest level of S, in order to maintain an acceptable acquisition time by a practical point of view. The three levels of R were varied over 2, 4 and 8 cm^{-1} for all the experiments, except for grit and wholemeal flour analysed by TU instrumental setup, for which our previous experience suggested to use higher values (4, 8 and 16 cm^{-1}). Samples measured as grit were not analysed by FO because it was not possible to guarantee an acceptable reliability of the obtained measurements, due to the interface between sample and the fiber surface, whose diameter (3 mm) is too small compared to the grit dimensions and shape (the analysed sample should be in direct contact with the FO surface). Therefore, globally the Doehlert design was applied to eight different sample form/instrumental setup combinations (both TU and IS on grit, wholemeal and white flour and FO on wholemeal and white flour).

The NIR spectra were acquired for all the seven experimental conditions (*points*) of the design, replicated three times in random order. Each spectrum acquisition was repeated three times by each laboratory (each time with different sample aliquots) for a total of nine spectra for each point and for each sample form/instrumental setup combination, leading globally to 1008 NIR spectra.

Noise estimate

The estimation of the noise was obtained starting from the variance spectra of the 3 repeated measurements. These signals highlight the presence of two kinds of “noise” contributions: a high-frequency component (ascribable to the “pure” instrumental noise) and a low-frequency one, which is essentially due to the differences in the specific nature of the sample (e.g. variability of particle size or moisture content among different sample aliquots). In Figure 3 some examples of variance spectra are reported.

The differences due to the sample inhomogeneity cannot be eliminated even by properly adjusting the instrumental conditions by means of the experimental design, hence the low frequency variance contribution has not to be considered in the noise minimisation process. High frequency and low frequency variances were separated by means of two different signal processing techniques: Savitsky-Golay filtering (SG) and Fast Wavelet Transform (FWT).

The Savitzky-Golay filtering [7, 8] is a widespread smoothing processing, where a polynomial function of a given order is fitted to the signal in a sliding window having an odd number width. The polynomial function is moved point by point across the signal and, each time, the experimentally measured value of the midpoint of the window is substituted with the corresponding value calculated with the polynomial. By varying the order of the polynomial and the width of the window it is possible to modulate the extent of the smoothing, taking care to not distort the signal shape.

The FWT [9-11] operates on an individual discrete signal by convolving it with a couple of filters (called the High-pass and Low-pass decomposition wavelet filters, i.e., Hi_D and

Low_D, respectively), thus splitting it into two orthogonal subspaces, called vectors of approximations (retaining only the low frequency content of the signal) and of details (which collects the high frequency content), respectively. The two wavelet filters are orthogonal, since frequencies "kept" by the Lo_D are not "kept" by the Hi_D (and *vice versa*), and complementary, since the original signal can be perfectly reconstructed from the approximations and details vectors, by applying the proper couple wavelet reconstruction filters (Lo_R and Hi_R). Then, the decomposition procedure can be repeated to a further decomposition level, applying the same two filters to the approximations vector. In this way, sharp and coarse properties of the signal are captured and disjointed in different sub spaces (vectors of wavelet coefficients, called 'blocks'), at different levels of resolution. In the present application, the approximations vector at the lowest level of decomposition has been kept as the low frequency component of the variance signal, i.e. as the component that is not related to instrumental noise, but to the variations in the specific nature of the sample.

A proper Matlab function, namely *noise_estimate*, was written to subtract the low variance contribution from the variance spectrum – using both SG filtering and FWT – to obtain the instrumental noise spectrum. The suitability of the chosen conditions for SG filtering (i.e., size of the filtering window and order of the fitting polynomial) and for FWT decomposition (i.e., type of wavelet and decomposition level) with respect to the convenient noise extraction was veryfied by visual inspection for each combination sample form/instrumental setup. The output figure of *noise_estimate* applied to the NIR spectra measured in the experimental conditions corresponding to the Doehlert point (0.5; -0.866) of white flour sample analyzed by the integrating sphere is reported in Figure 4. Figure 4a is an example of a poor extraction of instrumental noise from the variance spectrum, due to the choice of a not optimal decomposition level in the wavelet domain, while Figure 4b is an example of good separation of instrumental noise from low frequency variations. The subtraction of the low variance contribution from the variance spectra was repeatedly done by means of *noise_estimate* until satisfactory results (i.e., optimal filtering conditions) have been reached.

The preliminary tests permitted to set the polynomial order of the SG interpolating function equal to 2, as the best compromise between eliminating the low frequency variance without loosing spectral information. As for the wavelet functions for FWT decomposition, different wavelets from the *symlet* wavelet family (*sym*) were tested, since these often resulted the most appropriate ones to analyse the NIR spectral shapes [3]. Finally the *sym8* wavelet was chosen, which is a nearly symmetric wavelet whose filters are 16 points long [12].

In order to maintain the range of the interpolating function approximately constant with respect to the wavenumber values of the spectra, the number of points for SG filter varied on the basis of the R level. For the same reason, also the maximum decomposition level of FWT varied with the R level (Tab. 1).

Once *noise_estimate* was punctually applied to the different variance spectra sets, then a further Matlab function was developed (*noise_estimate_fast*), which was employed to collect a comprehensive summary of the signal processing conditions found for each variance spectra set and to calculate the corresponding pure instrumental noise value. In fact, *noise_estimate_fast.m* provided an output array containing the best SG and FWT conditions, a figure reporting into three subplots the original signals, the low variance contribution signals and the noise signals, and finally the mean of the absolute values of noise, that was considered as the response variable to be minimised.

Both functions were developed in Matlab language ver. 7.0, using some of the funtions available in the Wavelet Toolbox ver. 3.0.

Results and discussion

The first visual inspection of noise behaviour was done by means of 3D scatter plots of the response values obtained on the 3 replicates of the 7 Doehlert points. For some of the sample form/instrumental setup combinations (e.g., for grit analysed by integrating sphere, which is reported Figure 5) consistent differences among replicates were observed. The reason for this behaviour is probably ascribable to the fact that noise is a completely stochastic variable, so that its values are not as much reproducible as those of “ordinary” responses, where the deterministic component is generally much greater than the stochastic one. Thus, even though the mean value of the three replicate measurements for each experimental condition accounts for the instrumental noise values that are encountered on the average, it could be more interesting to estimate the maximum instrumental error that can affect the measurements. In fact, since noise has to be minimised, the maximum value can be considered as a pessimistic estimate of the noise amount, therefore allowing to choose the most convenient experimental conditions even in the worst cases. Based on this considerations, it was decided to create the response surfaces using both the average value and the maximum value of the three replicates. As it is shown in Figure 6, the shapes of the response surfaces can slightly change when using average or maximum values for multilinear regression. However, in all the considered conditions (sample form/instrumental setup) it was observed that the overall behaviour and (most important) the location of the minimum values of instrumental noise within the analysed experimental domains are coincident when using average or maximum values .

The optimal measuring conditions, i.e. those leading to the minimum instrumental noise values, were estimated through the computation of the multilinear regression models described in eq. 1 for all the sample form/instrumental setup combinations, considering separately the data measured in the two independent laboratories, thus modelling the noise estimates obtained from the two different instruments of the two laboratories by separate response surfaces. Then, the eight response surfaces obtained for each sample form/instrumental setup combination (2 for SG-FWT × 2 for mean-max × 2 for lab1-lab2) were compared to pick out, as the best overall compromise, the experimental condition (Doehlert point) leading to the minimum of noise (Figure 7). Even if the two laboratories analysed different samples, the trends of the corresponding response surfaces – obtained on the same sample form measured with the same instrumental setup – generally converged. In Table 2, the optimal experimental conditions for spectra acquisition are reported, in terms of Resolution and number of Scans, obtained for all the sample form/instrumental setup combinations. Despite the fact that it is usually assumed that the S/N ratio increases with decreasing the resolution and with increasing the number of scans, it has to be noticed that the results do not ever confirm this rule. In general, the combinations between the high R level and the medium-high S level or between the medium R level and the high S level worked as best.

Conclusions

This work confirmed the convenience of applying experimental design techniques to find the optimal measuring conditions in FT-NIR analysis. In fact, the obtained results demonstrate that the amount of noise into the spectra depends not only on the resolution and on the number of scans with which the signals are acquired, but also on the physical nature of the sample and on the adopted instrumental setup. In particular, the choice of the optimal combination of all these variables to produce low noise spectra is not so straightforward and can not be known *a priori* in a precise way.

In the end, even if the application of the Doehlert design required some time, it significantly helped in the selection of the measuring conditions with each instrumental setup, for the subsequent acquisition of a spectral dataset of about 15.000 NIR spectra of bread wheat under the forms of grit, wholemeal and white flour. This kind of application is quite novel in the FT-

NIR spectroscopy field, where chemometrics is heavily involved, but usually only in the phase of spectra elaboration, when the time and effort requested to correct or recover rough spectra are often excessive with respect to the time and effort spent for the spectra acquisition phase.

Aknowledgements

Regione Lombardia (Italy) is gratefully acknowledged for financial support (research project AUTOSTOC). Dr. Maria Corbellini (CRA – S. Angelo Lodigiano) and Dr. Paolo Belloni (Bruker Oprics Italia) are also gratefully acknowledged for technical support.

References

[1] Cocchi M, Corbellini M, Foca G, Lucisano M, Pagani MA, Tassi L, Ulrici A (2005) Anal. Chim. Acta 544: 100-107.

[2] Foca G, Ulrici A, Corbellini M, Pagani MA, Lucisano M, Franchini GC, Tassi L (2007) J. Sci. Food Agr. 87(5): 839-846.

[3] Foca G, Cocchi M, Li Vigni M, Caramanico R, Corbellini M, Ulrici A (2009) Chemom. Intell. Lab. Syst. 99: 91-100.

[4] Furukawa Y (2002) In: Near Infrared Spectroscopy – Principles, Instruments, Applications, Siesler HW, Ozaki Y, Kawata S, Eise HM (eds) Wiley-VCH, Weinheim, p. 92.

[5] Borasio E, Proceedings of the GranoItalia Symposium, Bologna, pp. 59-61 (1997).

[6] Ferreira SLC, dos Santos WNL, Quintella CM, Neto BB, Bosque Sendra JM (2004) Talanta 63: 1061-1067.

[7] Brown SD (2006) In: Practical guide to chemometrics – 2nd edition, Gemperline P (ed) CRC Press, Boca Raton, p. 403

[8] Savitzky A, Golay MJE (1964) Anal. Chem. 36: 1627-1639.

[9] Mallet Y, de Vel O and Coomans D (2000) In: Wavelets in chemistry, Walczak B (ed), Elsevier, Amsterdam, p. 74.

[10] Mallat S, A Wavelet Tour of Signal Processing - 2nd Ed., Academic Press, New York, US (1999).

[11] Barclay VJ, Bonner RF, Hamilton IP (1997) Anal. Chem. 69: 78-90.

[12] Misiti M, Misiti Y, Oppenheim G, Poggi JM (2004) Wavelet Toolbox, The MathWorks, Natick, US.

Figures captions

Fig. 1. Q vs. T^2 plot from the PCA model based on the experimental parameters measured on the bread wheat samples. The selected samples are represented with black squares.

Fig. 2. R and S level values used in the Doehlert design for the different sample form/instrumental setup combinations. The numbers reported in the grey circles correspond to the coordinates of Doehlert points in the experimental matrix.

Fig. 3. Examples of some variance spectra, where the high and the low frequency contributions (sharp and coarse variations, respectively) can be observed.

Fig. 4. Example of an output figure from *noise_estimate*, which was obtained by applying to the variance spectrum the FWT processing with a sym8 wavelet at decomposition levels 3 (a) and 6 (b).

Fig. 5. 3D scatter plot of noise values of the three replicate experimental measurements acquired for each one of the seven experimental conditions of the Doehlert design, obtained for grit analysed by IS.

Fig. 6. 3D response surfaces for grit analysed by IS by laboratory 1 computed on average (a) and maximum (b) noise values.

Fig. 7. 2D response surfaces obtained on grit analysed by IS for all the laboratory/sample form/instrumental setup combinations.

R (cm^{-1})	TU			IS			FO	
	Grit	Whole meal	White flour	Grit	Whole meal	White flour	Whole meal	White flour
2			(107)/(*5*)	(107)/(*6*)	(107)/(*6*)	(107)/(*6*)	(53)/(*5*)	(53)/(*5*)
4	(53)/(*4*)	(53)/(*4*)	(53)/(*4*)	(53)/(*5*)	(53)/(*5*)	(53)/(*5*)	(27)/(*4*)	(27)/(*4*)
8	(27)/(*3*)	(27)/(*3*)	(27)/(*3*)	(27)/(*4*)	(27)/(*4*)	(27)/(*4*)	(13)/(*3*)	(13)/(*3*)
16	(13)/(*2*)	(13)/(*2*)						

Table 1. (Number of points for SG filtering)/(*maximum FWT decomposition level*), as they vary with the R level.

		Optimal R (cm^{-1})	Optimal S (# scans)
Transmission Unit	**Grit**	16	155
	Wholemeal	8	200
	White flour	4	200
Integrating Sphere	**Grit**	8	155
	Wholemeal	8	155
	White flour	8	155
Fiber Optic	**Wholemeal**	4	200
	White flour	8	155

Table 2. Optimal R and S levels for spectral acquisition obtained by using the experimental design.

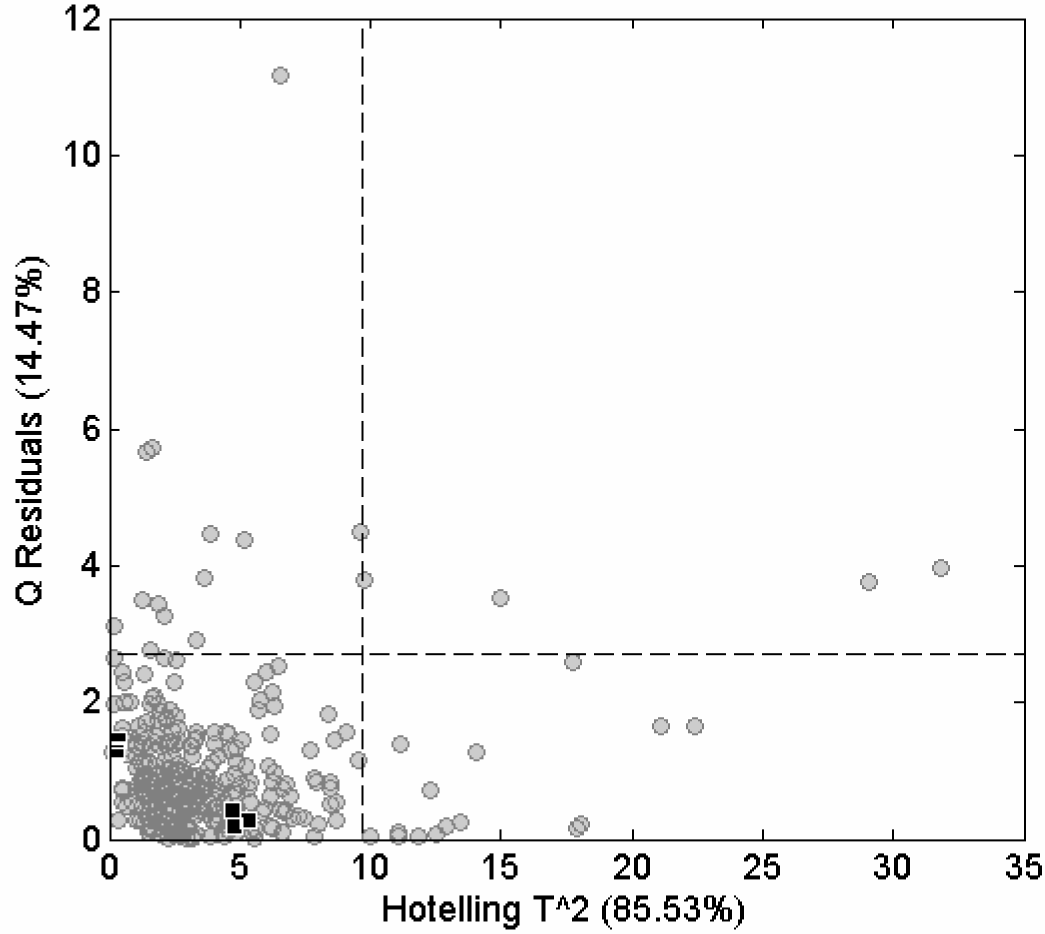

12
10
8
6
4
2
0
0
5
10
15
20
25
30
35
Q Residuals (14.47%)
Hotelling T^2 (85.53%)

Figure 1

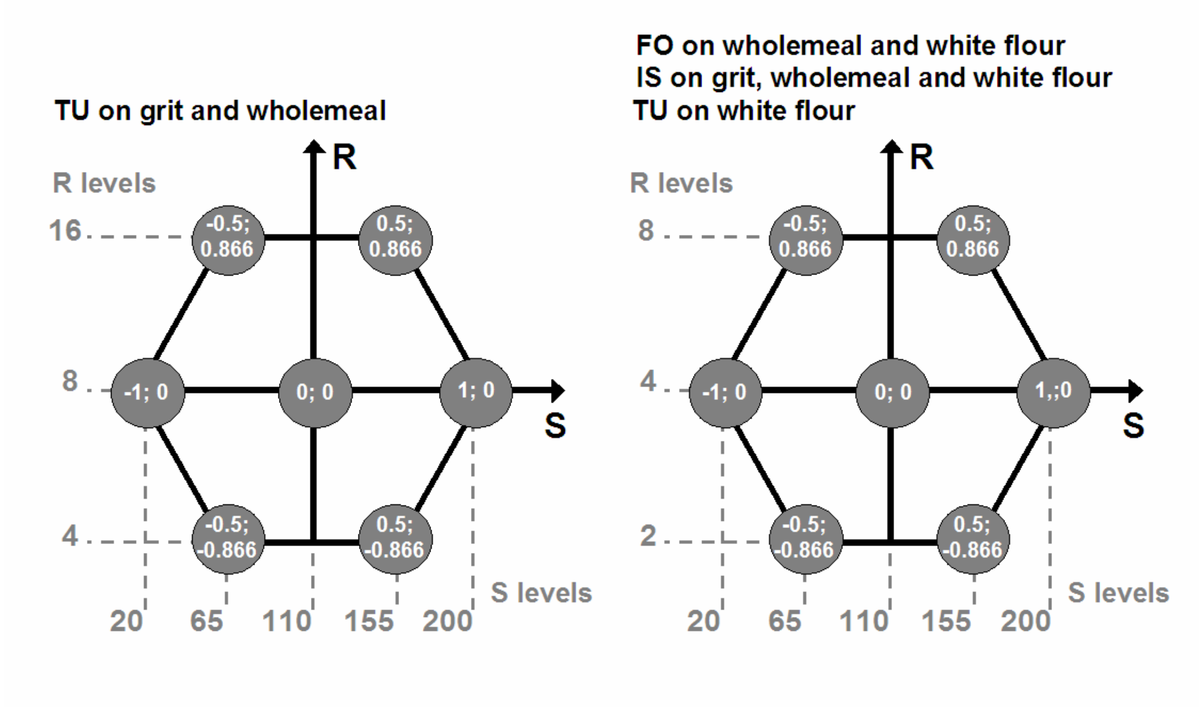
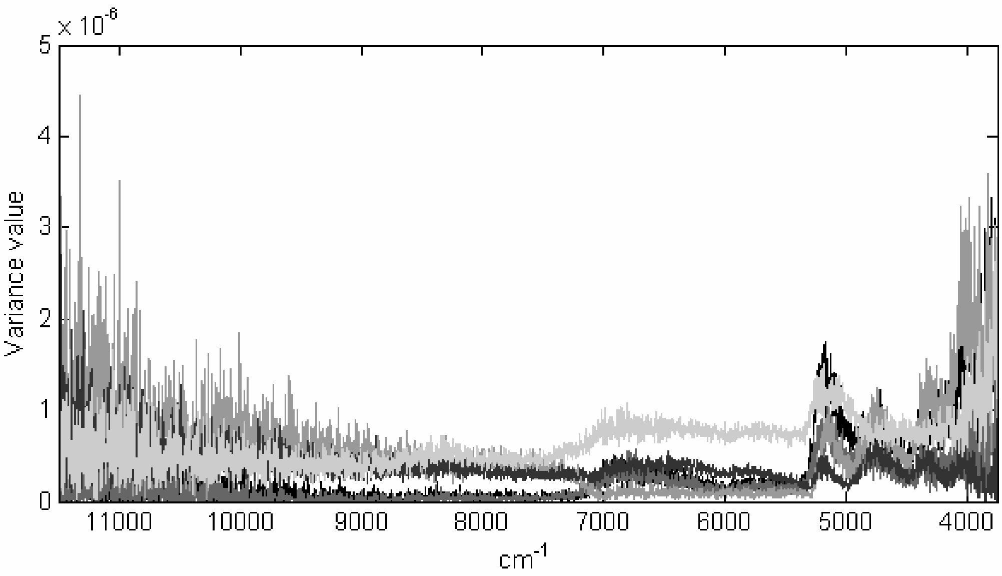

TU on grit and wholemeal
R
R levels
16
8
4
-0.5; 0.866
0.5; 0.866
-1; 0
0; 0
1; 0
-0.5; -0.866
0.5; -0.866
S
S levels
20
65
110
155
200
FO on wholemeal and white flour
IS on grit, wholemeal and white flour
TU on white flour
R
R levels
8
4
2
-0.5; 0.866
0.5; 0.866
-1; 0
0; 0
1,;0
-0.5; -0.866
0.5; -0.866
S
S levels
20
65
110
155
200

Figure 2

332 

333 **Figure 3**

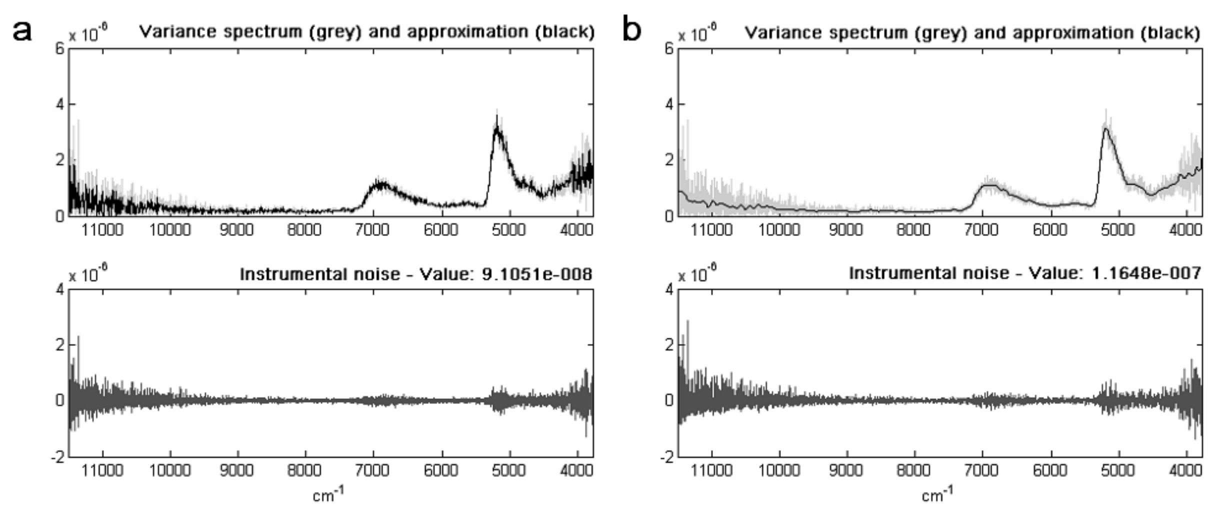

a
b
Variance spectrum (grey) and approximation (black)
Variance spectrum (grey) and approximation (black)
Instrumental noise - Value: 9.1051e-008
Instrumental noise - Value: 1.1648e-007
x 10-6
cm-1

Figure 4

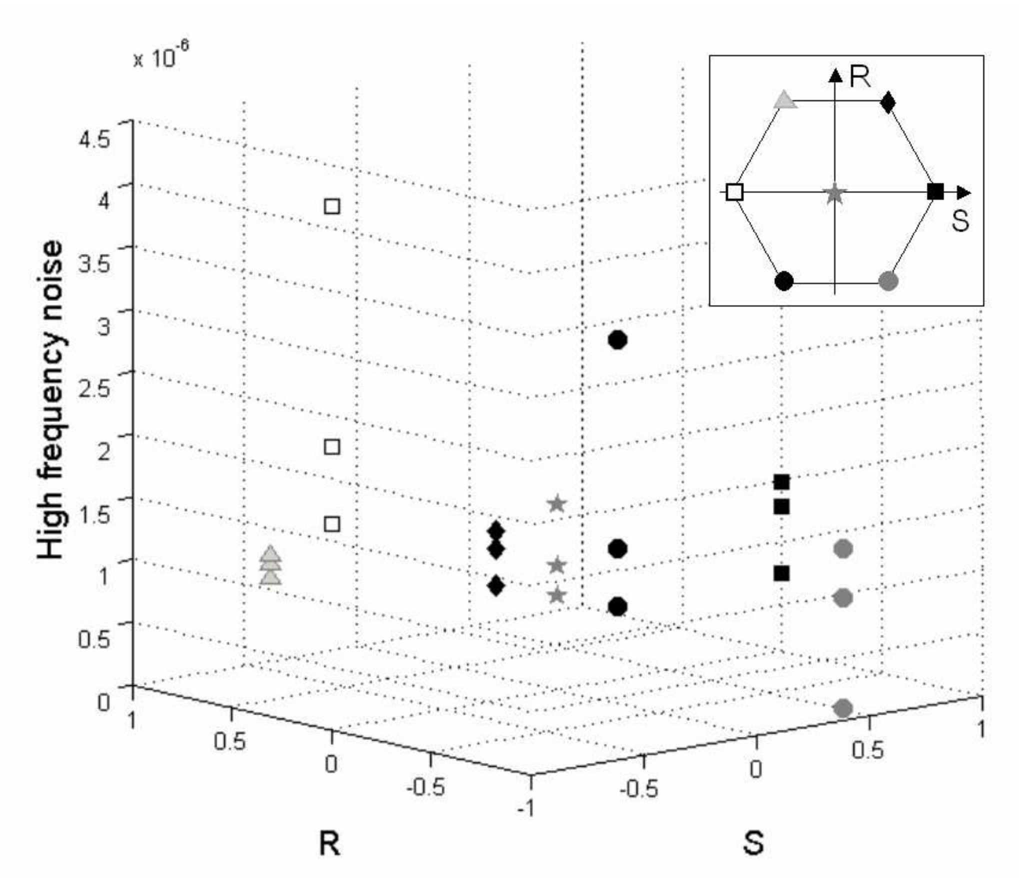

x 10^-6
High frequency noise
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
R
1
0.5
0
-0.5
-1
S
-0.5
0
0.5
1
R
S

Figure 5

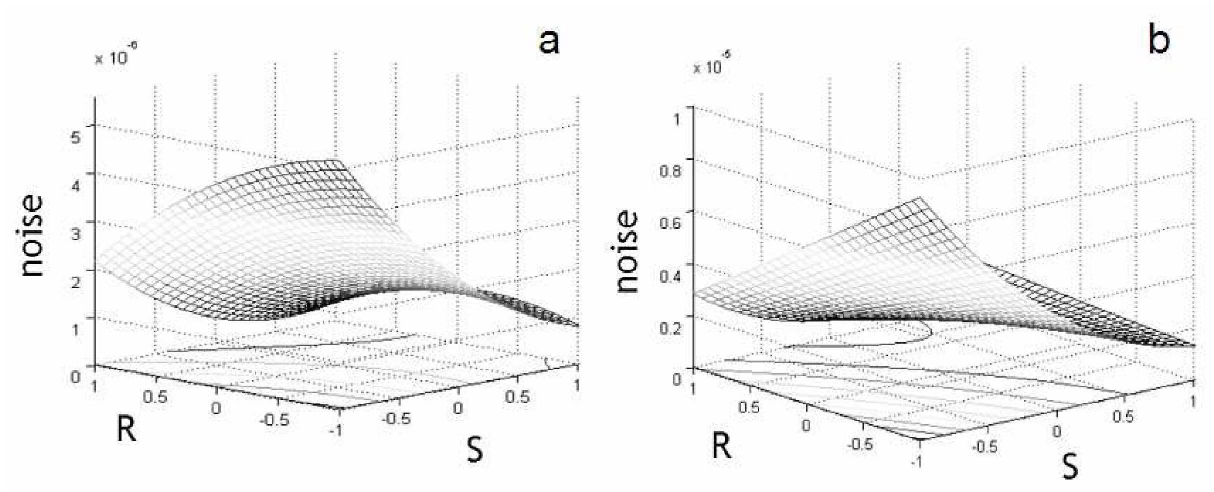


Figure 6

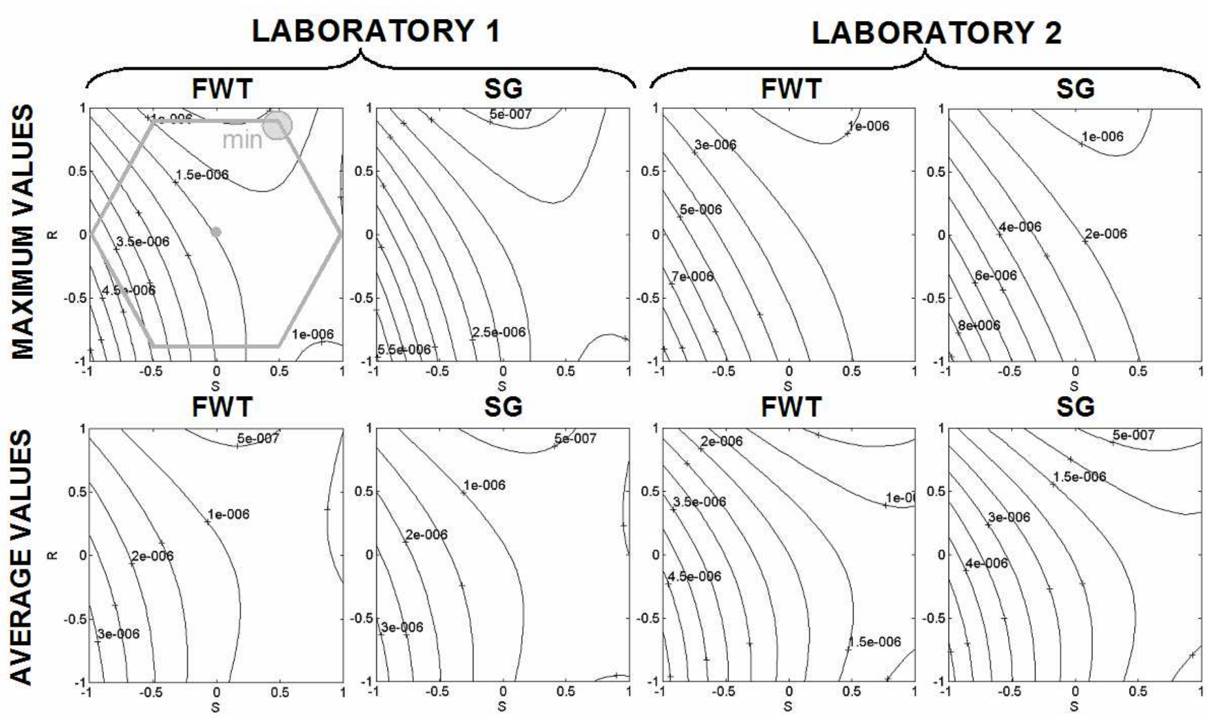

LABORATORY 1
LABORATORY 2
FWT
SG
MAXIMUM VALUES
AVERAGE VALUES
min
S
R

Figure 7