Factors influencing accuracy of NIR calibrations for the prediction of quality of Lithuania-grown rapeseed

B. Butkute

Lithuanian Institute of Agriculture. Instituto 1. Akademija. Kedainiai dist., LT-5051, Lithuania

Introduction

Near infrared (NIR) spectroscopy is a rapid, efficient, and thereby, low cost analytical technology, widely used for determination of oil and protein in rapeseed.[1–4] American Oil Chemists' Society (AOCS) has approved this method for the determination of oil, moisture and volatile matter and protein (Am 1-92). NIR can also be used to predict minor constituents, such as glucosinolates (GSL) and chlorophyll in rapeseed.[1,3,5–7] Velasco and Becker[8] reported the application of NIR calibration equations for individual GSL composition. The analysis of rapeseed in a network of NIR instruments is successfully used.[3] NIRS calibrations have shown to be sensitive to year, time of sowing, crop location and variety,[2] spectra transformation[6] or calibrating wavelength range.[6,9]

Oilseed rape is the first crop in today's Lithuanian agriculture that has found its way to the western market. The cultivation of rapeseed, especially of winter crop, is rather complex in Lithuania. Although high quality varieties are grown in Lithuania today, very changeable climatic and soil conditions determine sowing and ripening time. All this can produce fluctuations from year to year in oil, protein and glucosinolate content.[10,11] Profitability of rapeseed cultivation depends on seed oil content and the quality of rapeseed cake depends on crude protein content. The content of GSL, as antinutritive and toxic compound, is controlled in rapeseed. Rapeseed moisture content is a very important parameter for storage. In Lithuania when rapeseed is harvested in rainy weather moisture content in seed can exceed 20%. Only 00 oilseed rape varieties are currently grown in Lithuania and the effect of agroclimatic factors on their seed quality is being investigated. In order to estimate small variations of GSL content in the seed of the double low varieties we are seeking to develop an equation whose analysing accuracy would be reliable both for 00 type varieties and other imported types cultivated for technical purposes.

The aims of present study are to examine an impact of spectra math treatments, calibrating data base and wavelength range on the calibration and prediction accuracy and to develop equations practicable for the determination of glucosinolate, crude fat, crude protein and moisture contents in Lithuania-grown rapeseed.

Materials and methods

Materials

The database of NIR spectra was composed of rapeseed samples of different varieties grown in various areas of Lithuania from 1994 to 2001. The greatest number of samples were grown in the experimental fields of the Lithuanian Institute of Agriculture (LIA), variety testing stations, individual farms or agricultural partnerships.

The characterisation of the calibration sets is presented in Table 1.

Near Infrared Spectroscopy: Proceedings of the 11th International Conference.

Table 1. Calibration sets characterisation (value in air dried matter).

	Moisture %	Crude fat (CF) %	Crude protein (CP) %	Glucosinolates (GSL) μmol/g		
				I set	II set	III set
N	1125	737	619	129	134	143
Range	4.7–23.2	35.6–47.9	15.7–26.6	3.6–23.80	3.6–85.6	2.7–89.0
Mean	7.0	42.1	20.8	10.8	11.6	13.9
SD	2.66	2.22	2.06	3.15	8.83	13.64

Analysis by reference

Seed quality components were determined by the following reference methods: moisture content according to LST 1321:1993 standard requirements, i.e. the samples were dried for 40 min. at 130 ± 2°C; CP content—by Kjeldahl method; CF content—by the method ISO 659:1988 modification, when oil content in a sample is calculated by the residue of defatted by hexane matter after extraction. GSL analysis was done by gas chromatography of silyl derivatives in Poland.[12] II set was composed on the basis of I set having included the newly tested samples, in three of which GSL content exceeded 60 μmol g^{-1}. Set III is composed of set II expanded with the samples grown in 2000–2001 in Lithuania, Belarus, the Ukraine with a GSL content of 30–89 μmol g^{-1}, and below 3.6 μmol g^{-1}. Two prediction sample sets were composed: A-with 4.4–88.8 μmol g^{-1} (mean 23.5 μmol g^{-1}, $n = 54$), and B-with 4.4–13.3 μmol g^{-1} (mean 10.2 μmol g^{-1}, $n = 40$). Part of these samples were analysed for GSL in other European laboratories, regrettably by other methods: HPLC or NIR.

Scanning and calibration

Intact seed samples were scanned on a monochromator NIR Systems model 6500 (Perstorp Analytical, USA) equipped with Spinning Module using a small ring cup (Ø 4,7 cm). The number of complete scans to average was two, RMS–100 (but not more than seven scans were used). Different wavelength ranges were chosen for calibration. Mathematical treatment of spectral data was performed using the software ISI-NIRS2 Version 3.10, Intrasoft International, Port Matilda, PA, USA. Interfering effect was minimised by SNV and de-trending D (SNVD) transformation. The equations were calculated using a MPLS algorithm and cross-validation techniques.

Results

Effect of mathematical treatment on the accuracy of the equations

The following combinations of derivative, gap, smooth, smooth2 were used for moisture, CF, CP data calibration: 1,4,4,1; 2,4,4,1; 2,6,4,1; 3,6,4,1. Additionally for CP data calibration the math treatments 2,8,4,1; 3,6,6,1 were studied. Variation ranges of the statistics of the equation accuracy are presented in Table 2. The accuracy of the equations for the prediction of moisture content was slightly affected by the mathematical spectra transformation method: statistical parameters of the equations produced using 2nd derivatisation were better compared with the ones prodused using the first or third order spectral data derivatisation. The equation for whose calculation 400-2500 nm spectral data were transformed by 2,6,4,1 is considered to give the most accurate analysis for the prediction of moisture. The effect of the calibrated spectra mathematical treatment on the equations for predicting the main rapeseed quality parameter—CF content was more marked compared with the equation for moisture content. Calibration of the entire spectrum with mathematical treatment 2,6,4,1 enabled to develop a more accurate equation for the prediction of CF. Mathematical spectral data treatment when calculating calibration equations for the prediction of CP content in rapeseed by NIR did not have any significant effect on accuracy parameters of the equations in cross-

validation and prediction. In this case we can discern only some trends: slightly better calibration results, especially in the prediction, were obtained when applying the second order derivatisation of the spectral data.

Table 2. Range of variation of equation accuracy statistics as affected by mathematical treatment. The entire spectrum (400-2500 nm) is calibrated.

	SEC	*RSQ*	*SECV*	1-*VR*
Moisture	0.154 ÷ 0.161	0.994 ÷ 0.996	0.168 ÷ 0.174	0.993 ÷ 0.995
Crude fat	0.420 ÷ 0.458	0.947 ÷ 0.958	0.486 ÷ 0.541	0.925 ÷ 0.939
Crude protein	0.362 ÷ 0.410	0.948 ÷ 0.960	0.424 ÷ 0.460	0.936 ÷ 0.944
	Prediction			
	N	Range	SEP	R^2_{pred}
Moisture	52	5.69 ÷ 10.42	0.221 ÷ 0.311	0.984 ÷ 0.989
Crude fat	38	42.75 ÷ 46.23	0.475 ÷ 0.571	0.902 ÷ 0.933
Crude protein	46	18.6 ÷ 22.4	0.336 ÷ 0.387	0.922 ÷ 0.940

To investigate the effect of mathematical transformation on the accuracy of the equations for the prediction of GSL we calibrated 1100–2500 nm spectral region. Statistical parameters of calibration equations obtained by calibrating spectra I set are shown in Figure 1. The statistics of accuracy of equations developed using the raw (log/1*R*, i.e. derivative – 0) optical data was unsatisfactory. Although coefficients of correlation *RSQ*, 1-*VR* are higher and standard errors *SEC*, *SECV* became consistently lower when spectra derivatisation was higher, except for the case, when spectral data were transformed using gap 6 and smooth –4. All statistical parameters of the equation developed using math treatment 3,6,4,1 were worse than the analogous with the second order derivatised data.

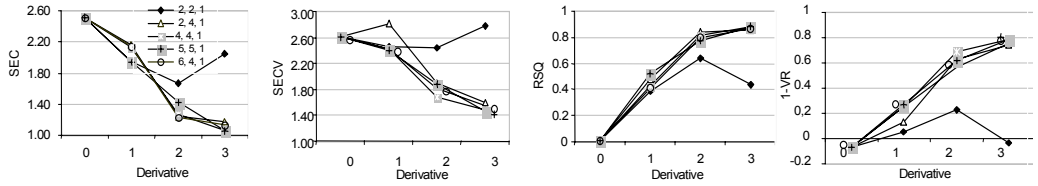

Figure 1. The equation statistics developed by using different math treatments for spectral data transformation.

The discussed spectra derivatisation affected the accuracy of equations (in calibration, cross-validation and prediction) by different intensity subject to calibrating sample set composition and GSL range.[6] It should be noted that in all investigated cases combination of all math treatment parameters – derivative, gap and smooth, is relevant.

Effect of calibrated spectral region on the statistics of the equations

Many publications indicate that the wavelength segment 400–1100 nm is "speculative waves" and little information is found in this part of spectrum about the characteristic functional groups of the identified compounds. They are rarely included into of the calibrated spectral region.[2,4,8] Correlograms of each component clearly show that in the first segment, in the visible range (400–700 nm) coefficients are close to zero (Figure 2). In the correlograms of moisture, crude protein, and fat contents we can find more or less high correlation coefficients within a large range of spectrum, while the correlogram of glucosinolates has distinct peaks in a narrow range of the spectrum—

between 1610–1630 nm. The next stage of the work is search for the spectrum range whose calibration contributes most to the accuracy of the developed equations.

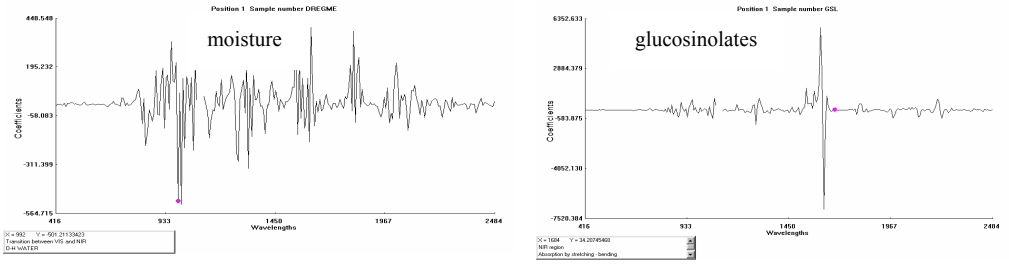

Figure 2. Plot of wavelength correlation with moisture and glucosinolate contents. Math treatment for moisture 1,4,4,1, for GSL – 2,4,4,1.

Impact of calibrated spectral region on the accuracy of calibration for prediction of moisture, CP and CF

In the present study the effect of altering the spectral range on the accuracy of equations for prediction of moisture, CP, CF was examined by producing equations with math treatment 2,4,4,1. The spectral ranges 400–2500; 750–2500; 850–2500; 1100–2500; 850–2350 nm were tested.

Table 3. The range of the accuracy of the equations for the prediction of moisture, CF and CP as affected by calibrated spectral region. Characteristics of the prediction sets are the same as in Table 2.

	SECV	1-*VR*	*SEP*	R^2_{pred}
Moisture	0.170 ÷ 0.173	0.994 ÷ 0.995	0.271 ÷ 0.278	0.985 ÷ 0.987
Crude fat	0.492 ÷ 0.538	0.926 ÷ 0.937	0.473 ÷ 0.541	0.924 ÷ 0.942
Crude protein	0.388 ÷ 0.424	0.944 ÷ 0.955	0.332 ÷ 0.364	0.929 ÷ 0.939

The number of calibrated wavelengths did not have any significant effect on the accuracy of the equations for the prediction of moisture, CF and CP in rapeseed (Table 3). Little improvement in accuracy of equations was obtained when spectral data with math treatment 2,4,4,1 were calibrated without visible spectral region: 850–2500 nm for moisture, 700–2500 nm or 1100–2500 nm for CP and 1100–2500 nm for CF.

Impact of calibrated spectral range on the accuracy of calibration for the prediction of GSL

Since European double low oilseed varieties are grown in Lithuania the greatest attention was paid to the development of an accurate equation using the database composed of 00 rapeseed samples, namely I set. Previous approaches to the prediction of GSL content in rapeseed through NIRS-6500 have mainly used the wavelength range from 1100 to 2500 nm. Lila and Furstoss[7] using filter NIR instrument noticed that specific wavelengths for GSL analysis are in the region of 1600-1680 nm. In the present study the effect of spectral range on the accuracy of prediction of total GSL content in rapeseed was examined for the equations developed by four different spectral math treatments. For that we calibrated the spectra of the following spectral regions: 400–2500, 700–2500, 8500–2500, 1100–2500 nm etc. (Figure 3). When the equations were produced with math treatment 1,4,4,1, statistics of calibration and cross-validation improved consistently with calibrating spectral interval approach to the 1500–1800 nm. Statistics of the equations developed using math treatment 2,4,4,1 and 3,4,4,1 depending on the wave interval of calibrated spectra did not change so intensely and regularly as when calibrating spectral data of the first derivatisation. All the statistical indicators of the statistics of the equations produced by using raw log 1/*R* spectral data were very low. However, when calibrating a narrow wavelength range (1500–1800, 1600–1800 or 1600–1750)

statistics in calibration and cross-validation became similar to the statistics of the equations developed by derivatised spectral data. The equations for the prediction of GSL were unreliable in all cases when the visible spectral region (400–700 nm) was included in calibration [Figure 3(c)].

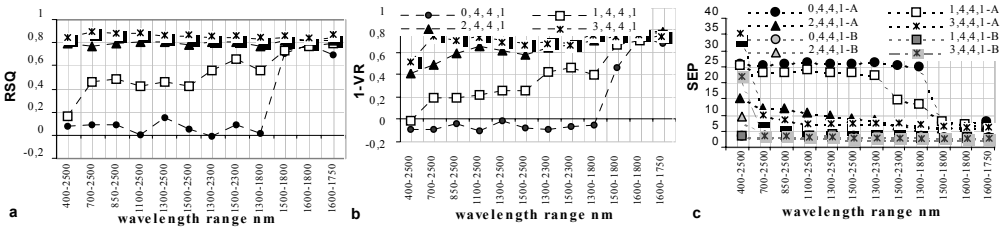

Figure 3. Effect of calibrated spectra derivatisation and region on the equations accuracy statistics in calibration (a -**RSQ), cross- validation (**b - **1-*VR*) and prediction (**c - ***SEP*). A-prediction sample set with GSL range 4.4–88,8 μmol g^{-1} and B- with GSL range 4.4–13.3 μmol g^{-1}.

The accuracy of the equations obtained by calibrating spectral data of III sample set (with different types of varieties) of the second order derivatisation was especially dependent on the wave range of the calibrated spectra. This was evident on statistical indicators in calibration [*RSQ*, Figure 4 (a)], in cross- validation [1-*VR*, Figure 4 (b)] and in prediction [*SEP*, Figure 4(c)]. The same trend was revealed in the cross-validation and prediction accuracy for the other sets of calibration (I and II), except for a few exceptions. While analysing using the equations obtained by calibrating 1500–1800 nm spectra wavelength interval, all the statistical indicators were the best for each calibrating set.

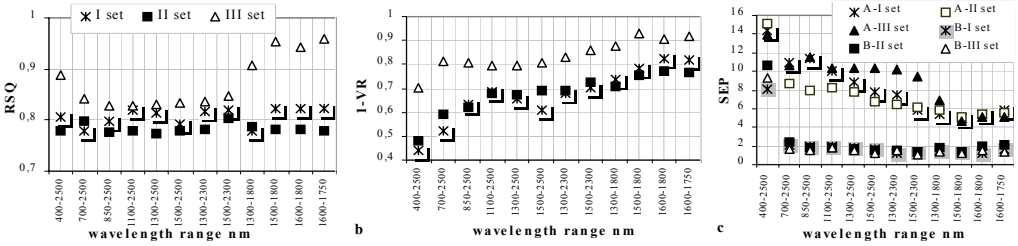

Figure 4. Impact of calibrated sample set and wavelength range on the accuracy of equations for the prediction of GSL: a - coefficient of correlation in calibration (*RSQ*), b – coefficient of correlation in cross-validation (I-*VR*), c – standard error of prediction (*SEP*) for sample set with GSL range 4.4–88.8 μmol g^{-1} (A) and with GSL range 4.4–13.3 μmol g^{-1} (B).

The *SEP* for prediction set with low content of GSL varied from 1.18 to 1.28 μmol g^{-1}, while for the other prediction set, that included the samples with GSL >30 μmol g^{-1}—from 4.67 to 5.17 μmol g^{-1}. Predicted and reference GSL value for 00 rapeseed samples differed by less than 4 μmol $g^{-1}$and for samples containing over 20 μmol $g^{-1}$of GSL by less than 8 μmol g^{-1}. The *SEP* did not exceed the error of reproducibility by HPLC method specified in the ISO standard 9167-1. When calibrating database with a large number of 00 samples and only a few samples with GSL content exceeding 60 μmol g^{-1} (II sample set) the effect of calibrated wave range was similar to that for I set calibration. Samples with a high GSL content were eliminated during the process of calibration.

Discussion

The present study has over again demonstrated that the accuracy of an NIR prediction depends on the successful completion of several factors of calibration. Development of an accurate equation for GSL determination was affected by more factors compared with the equations intended for other rapeseed quality parameters. The first reason why calibration sensitivity of the equations for GSL was very high could be that GSL content in rapeseed is very low in comparison with the contents of the other quality components. The highest content of GSL (for example, 85 μmol g^{-1}) accounts for only about 3.3% of the total seed mass. The absolute majority of the samples in our database were the seed of double low varieties, i.e. seeds with 3–20 μmol g^{-1} GSL, or 0.12–0.8% of the seed mass. The optimal spectral range for GSL calibration was 1500-1800 nm. This spectral area seems very specific and highly sensitive[5–7] and could be associated to a well-defined structure of GSL as β-thioglucoside attached to carbon atom in N-hydroximine sulphate esters. Another reason is the abundance and diversity of database[2]. The database for the calculation of the equation for the total glucosinolate content was significantly less numerous dues to limited possibilities to determine the content of these compounds by accurate reference methods. In conclusion, the NIR calibration equations developed in this study could be used to evaluate changes of quality affected by Lithuanian meteorological conditions, growing location, genotype and cultural practices in 00 rapeseed and to detect samples with a high GSL content.

Acknowledgements

The author would like to thank Dr K. Michalski from the Plant Breeding & Acclimatisation Institute (Poland) for his benevolent technical advice and for the analyses of GSL content by reference. The Lithuanian State Science and Studies Foundation supported this work.

References

1. J.K. Daun, in *Brassica Oilseeds: Production and Utilization*, Ed by D.S. Kimber and D.I. McGregor. CABI Publishing, Wallingford, UK, p.243 (1995).
2. C.F. Greenwood, J.A. Allen, A.S. Leong, T.N. Pallot, T.M. Golder and T. Golebiowski, in *Proceedings of the 10th International Rapeseed Congress,* Ed by N. Wratten and P. Salisbury. Canberra, Australia, CD-ROM (1999).
3. P. Tillmann, T.-C. Reinhardt and Ch. Paul, *J. Near Infrared Spectrosc*. **8**, p.101 (2000).
4. L. Velasco, Ch. Möllers and H.C. Becker, *Euphytica,* **106**, 79 (1999).
5. R. Biston, P. Dardenne, M. Cwikowski, M. Marlier, M. Severin and J.-P. Wathelet, *JAOCS*, **65**, p.1599 (1988).
6. B.Butkute, *Biology*. Academia, Vilnius, Lithuania, **2** (supplement), p. 70 (2000).
7. M. Lila and V. Furstoss, *Agronomie*. **6**, p.703 (1986).
8. L.Velasco and H.C. Becker, *Plant Breeding*. **117**, p. 97 (1998).
9. P. Williams in *Near Infrared Spectroscopy: Proceedings of the 11th International Conference*, Ed by A.M.C. Davies and NIR Publications, Chichester, UK, L.5.1 (2004)
10. B. Butkute, A. Masauskienė and L. Sliesaraviciene*, Scientific Papers Agric. Univ. Cracow.* Krakow, **77 (**375), p.165 (2001).
11. G. Walton, P. Si and B. Bowde, in *Proceedings of the 10th International Rapeseed Congress,* Ed by N. Wratten and P. Salisbury. Canberra, Australia, CD-ROM (1999).
12. K. Michalski, K. Kolodziej and J. Krzymanski, in *Proceedings of 9th International Rapeseed Congress,* Ed by D. Murphy. Cambridge, UK, **3**, p. 911 (1995).