Prediction of the Heating Value of Biomass Fuel from IR and NIR Data

Bettina Liebmann*, **Anton Friedl, Kurt Varmuza**

Vienna University of Technology, Institute of Chemical Engineering
Getreidemarkt 9/166-2, A-1060 Vienna, Austria
www.lcm.tuwien.ac.at, www.vt.tuwien.ac.at

liebmann@mail.zserv.tuwien.ac.at, afriedl@mail.zserv.tuwien.ac.at,
kvarmuza@email.tuwien.ac.at

1. Motivation

Biomass is becoming increasingly important as a **renewable source of energy**. Apart from traditional incineration of wood, recent developments also consider the incineration of cereals.
Spectroscopy is a promising alternative to time-consuming calorimetric experiments for determining an important property of a fuel: its **heating value**.
This study aims at developing **PLS calibration models** for the prediction of higher heating values of **wood and cereal** samples, based on infrared and near-infrared data.

2. Experimental

2.1 Samples

20 wood samples: sawdust of spruce, pine and larch with varying amounts of additives (bark, rye/maize flour, starch)
15 cereal samples: wheat, rye, barley, maize and triticale flours

2.2 Instrumentation

Four different set-ups for spectroscopic measurements:

A. NIR-spectrometer with sapphire ball probe
(Brimrose Luminar 5030)

B. NIR-spectrometer with rotating sample cup
(Brimrose Luminar 5030)

C. VIS-NIR-spectrometer with Direct Contact Analyser
(Foss NIRSystems 6500)

D. IR-spectrometer with ATR-Durascope sampler
(Bruker Equinox 55)

2.3 Data

NIR absorbance data (A. - C.): 1100-2300 nm, $\triangle\lambda = 8$ nm, 1st derivative Savitzky-Golay, resulting in $p = 145$ features

IR absorbance data (D.): 4000-600 cm^{-1}, $\triangle\nu = 8$ cm^{-1}, 1st derivative Savitzky-Golay, resulting in $p = 435$ features

Heating values: The higher heating value in kJ/kg dry biomass is determined by the bomb calorimetric method [1], range: 18,143-19,125 kJ/kg

3. Chemometrics

3.1 Variable Selection

Variable selection by Genetic Algorithm (GA) using software MobyDigs [2]

- regression method: OLS
- performance criterion (fitness): R^2_{adj}, adjusted squared correlation coefficient between y and $\hat{y}$ for full cross validation [3]
- max. number of features selected in a model: 15
- number of iterations per trial: 1-1.4 million
- typical computation time: 60-120 minutes
- use variables present in the 10 best models

3.2 Repeated Double Cross Validation for prediction performance of PLS Models

Application of new software developed in R [4,5].

- The data set is randomly partitioned into $s = 4$ segments: 3 segments for calibration, 1 segment as test set.
- A PLS model is derived from the calibration set with optimum number of PLS components estimated by cross-validation.
- Application of PLS model to test set results in n/s predicted values $\hat{y}_i$.
- Systematic variation gives a $\hat{y}$ for each object.
- The whole process is repeated $k = 10$ times.
- Finally, $k \cdot n = 10 \cdot 35 = 350$ values $\hat{y}$ are available.

Performance criteria derived from prediction errors:

***SEP*$_{test}$** standard deviation of prediction errors
***TI*$_{90}$** tolerance interval of prediction errors (difference of 95 % and 5 % percentile)
***PC*$_{med}$** median of $s \cdot k$ calculated optimum numbers of PLS components

3.3 Full Cross Validation for prediction performance of PLS Models

For reference, PLS regression as implemented in software Unscrambler [6] has been performed.

***SEP*$_{CV}$** standard deviation of prediction errors by full cross validation (leave-one-out)
***PC*$_{opt}$** optimum number of PLS components given by software

4. Results

Table 1: *Prediction performance of PLS models using **all features** of NIR (A.-C.) and IR data (D.)*

data set	p	***SEP*$_{test}$**	PC_{med}	SEP_{CV}	PC_{opt}
A.	145	**141**	1	124	4
B.	145	**115**	6	102	5
C.	145	**89**	5	87	8
D.	435	**141**	3	137	2

Table 2: *Prediction performance of PLS models using **GA selected features** of NIR (A.-C.) and IR data (D.)*

data set	p	***SEP*$_{test}$**	PC_{med}	SEP_{CV}	PC_{opt}
A.	19	**96**	14	62	14
B.	21	**94**	10	56	17
C.	20	**96**	11	70	10
D.	19	**48**	13	62	7

p number of features
PC number of PLS components (*med.* median, *opt.* optimum)
SEP standard deviation of prediction errors (in kJ/kg)
SEP_{test} test sets in repeated double cross validation (10 repetitions)
SEP_{CV} leave-one-out cross validation

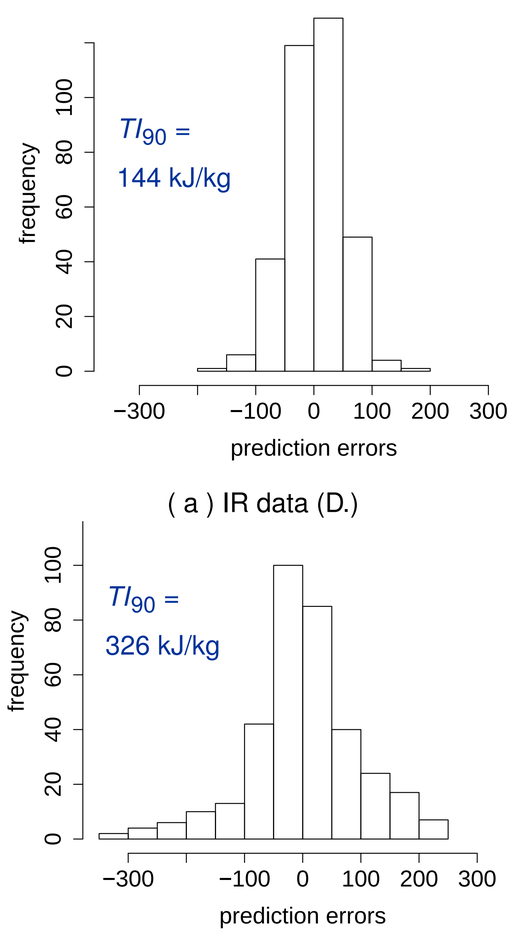

Figure 1: *Frequency distribution and tolerance interval of prediction errors of PLS models using 19 features of (a) IR data (best model) and (b) NIR data*

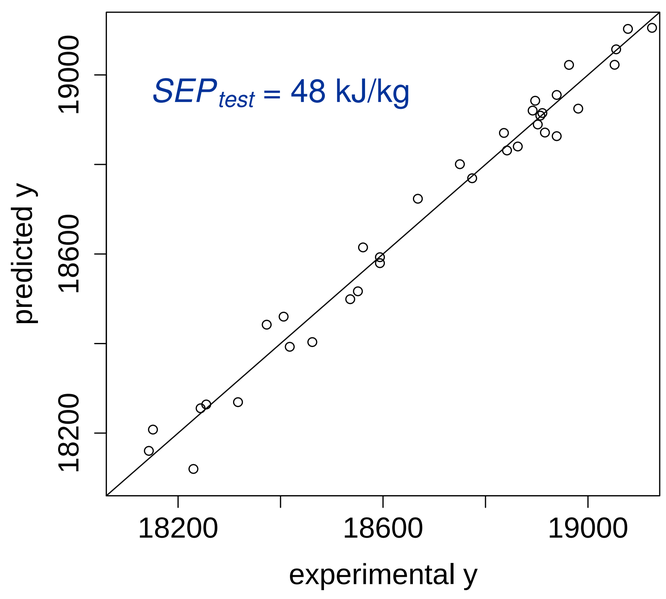

Figure 2: *Prediction of higher heating values (kJ/kg) by a PLS model using **19 IR absorbances selected by a genetic algorithm**. The predicted values are means from 10 replicates of double cross validation.*

5. Conclusions

- Easily available (Near-)Infrared spectroscopy data are very promising for the determination of heating values of biomass.
- Variable selection by Genetic Algorithm improves prediction performance in most cases, especially for IR data.
- Differences in prediction quality due to instrumental set-up (NIR data) decrease after variable selection.
- A single PLS-model can be used for the heating value of wood as well as cereal samples.
- Simple leave-one-out cross validation tends to give too optimistic results, which can be misleading for variable selection.
- Repeated double cross validation allows realistic estimation of prediction performance.
- Implementation of repeated double cross validation in software R is fast with typical computation times of 10 seconds per job.

6. Acknowledgments

We gratefully acknowledge financial support by *Hochschuljubiläumsstiftung der Stadt Wien*, project no. H-702/2005. We thank A. Kandelbauer (Competence Center Wood K-Plus, St. Veit an der Glan, Austria) for instrumental support, E. Padouvas for calorimetric measurements, and H. Schausberger (Saatbau Linz, Austria) for providing cereal samples. We thank P. Filzmoser for collaboration and development of a repeated double cross validation software in R.

7. References

1. Friedl, A., Padouvas, E., Rotter, H., Varmuza, K., Anal. Chim. Acta 544 (2005) 191-198.
2. Software MobyDigs, v 1.0. Talete srl, www.talete.mi.it, Milan, Italy, 2004.
3. Frank, I. E., Todeschini, R., The data analysis handbook. Elsevier, Amsterdam, 1994.
4. Software R, v 2.5.1. R Development Core Team, www.r-project.org, 2007.
5. Mevik, B. H., Wehrens, R.: J. Statistical Software 18 (2007), issue 2.
6. Software The Unscrambler v 9.0. Camo Process AS, www.camo.no, Oslo, Norway, 2004.

Poster Presentation: Conferentia Chemometrica, September 2-5, 2007, Budapest, Hungary