The Legume Content in Multispecies Mixtures as Estimated with Near Infrared Reflectance Spectroscopy: Method Validation

F. Locher, H. Heuwinkel,* R. Gutser, and U. Schmidhalter

ABSTRACT

The legume content in multispecies mixtures can accurately be estimated by means of near infrared reflectance spectroscopy (NIRS) provided that there is a valid calibration. This study was conducted (i) to test the applicability of two narrow-based calibrations, A and B (origin of the calibration sets: one farm, five harvests), for plant material not present in the calibration set and (ii) to compare the predictive ability of both models to a third, broad-based calibration, C (different farms and harvests). The prediction accuracy of the NIRS models, A (end-points calibration) and B (calibration with incremental standards), was tested with defined legume–grass mixtures of nine test sets, which differed in origin and harvesting dates. Most of the test-set spectra were later used as additional calibration samples for Model C. Calibrations A and B, which differed in calibration design, showed varying root mean square errors of prediction for each of the nine test sets (A: 3.3–12.5%; B: 3.9–10.3%). The slope of the line from regression of NIRS predicted on true values ranged between 0.93 and 1.09 (r^2 always > 0.92). All predictions were precise but biased. Prediction accuracy was worst for samples mixed of plant material that was grown in monoculture. After a bias correction and the exclusion of the mixtures of monocultural samples from the test sets, both models showed the same error (standard error of prediction after bias correction $< 6\%$). Calibrations A, B, and C were compared by predicting two external test sets. The broad-based Calibration C offered the same prediction accuracy as Calibrations A and B but did not reduce the bias. With Model C, there was no consistent reduction in the proportion of outliers in the test sets with a Mahalanobis distance > 3 compared with the Models A and B. It is concluded that predictions are biased even if more natural variability is included in the calibration set. Therefore, based on these data, a bias correction is always necessary. After a bias correction, all calibrations offered a highly accurate tool to estimate the legume content of mixtures independent of origin and harvesting dates. Contrary to our expectations, the most simple model, A, which was derived from the least number of calibration spectra and which represented the least natural variability, proved to be as accurate and robust in the prediction of independent test sets as the broader models, B and C.

Accurate knowledge of legume content in multispecies mixtures is necessary to estimate the amount of N_2 fixed by the legume–*Rhizobium* spp. symbiosis (Boller and Nösberger, 1987). Near infrared reflectance spectroscopy is a promising tool for measurement of the legume content of mixtures (Coleman et al., 1985, 1990; Petersen et al., 1987; Pitman et al., 1991; Shaffer et al., 1990; Wachendorf et al., 1999). The crucial point in estimating the legume content with NIRS is instrument calibration. A comprehensive set of samples (calibration set) representing the entire population is needed for the calibration procedure (Shenk and Westerhaus, 1991). From the calibration set, a prediction model is derived by means of chemometric tools like partial least squares regression (PLSR), which allows the prediction of legume content from near infrared reflectance spectra of unknown samples. Prediction accuracy therefore relies heavily on the extent to which the calibration set represents the samples to be predicted (Martens and Naes, 1987; Park et al., 1998). This representativeness can be difficult to achieve with natural products, which are affected by many sources of variability. Internal validations, such as cross-validation (Naes et al., 2002) or internal test-set validations, may yield lower prediction errors than external validations (Dardenne et al., 2000) because the predicted samples may belong to the same natural population as the calibration set (="closed" population, e.g., same harvest and field). Based on their calibration design, Coleman et al. (1990) concluded that botanical composition can be best estimated by NIRS when the population is closed. It is, however, desirable to have models with broader applicability. This can only be proved if an existing model is tested with "open" populations. This important step of external validation is rarely reported. For a realistic judgment of predictive ability, a test set completely independent of those samples used for calibration should be taken (Dardenne et al., 2000; Broad et al., 2002). Test sets should still fall within the general definition for the NIRS model (e.g., calibrations for legume–grass mixtures may not be appropriate for mixtures of legumes with other dicots). The accuracy of a NIRS calibration is described by a small root mean square error of prediction (RMSEP), and the regression line between the NIRS predicted and the true values should result in the target line, i.e., a slope and a coefficient of determination (r^2) close to 1.

Quality and future applicability of a calibration can be judged by the ratio of the standard deviation of the reference values to the standard error of prediction (RPD) value (Williams, 1987). Diller (2002) suggested RPD values lower than 2 indicate unsuitable calibrations while values between 2 and 3, 3 and 5, and 5 and 10 indicate calibrations with limited, satisfactory, or good quality, respectively. Values greater than 10 are excellent, as stated by Diller (2002). Caution is required in interpreting the RPD values because they depend strongly on the distribution and number of reference values. Therefore, they cannot be regarded as the ulti-

Dep. of Plant Sci., Technische Universität München, D-85350 Freising, Germany. Received 25 Feb. 2003. *Corresponding author (hauke@wzw.tum.de).

Published in Agron. J. 97:18–25 (2005).

677 S. Segoe Rd., Madison, WI 53711 USA

Abbreviations: MD, standardized Mahalanobis distance; NIRS, near infrared reflectance spectroscopy; PLSR, partial least squares regression; RMSEC, root mean square error of calibration; RMSECV, root mean square error of cross-validation; RMSEP, root mean square error of prediction; RPD, ratio of the standard deviation of the reference values to the standard error of prediction; SEP, standard error of prediction; $SEP_{biascor}$, standard error of prediction after bias correction.

mate criterion for prediction quality. However, in combination with RMSEP, they are helpful for judging the predicted values and may prevent erroneous assumptions concerning the real quality of a calibration model.

Finally, the number of outliers in sets of independent spectra as judged from the standardized Mahalanobis distance (MD) can give hints as to the robustness of the calibration. The MD is defined as the difference between an independent spectrum and the average of all calibration spectra in the factor space. Dardenne (1996) observed that with increasing number of calibration spectra, there was a decreasing number of outliers. He interpreted this as the greater ability of the calibration set to describe the variation in the validation set.

Near infrared reflectance spectroscopy calibrations should ideally not decrease in accuracy when new spectra from other harvests or origins are predicted. However, there is a trade-off between robustness and accuracy of a NIRS calibration that cannot be solved (Diller, 2002). Natural products are not as homogenous as well-defined chemical or pharmaceutical substances. Plants consist of substances (structural and soluble carbohydrates, protein, fat, organic acids, etc.) that all contain the most NIRS-relevant -C-H, -N-H, or -O-H groups. Growing conditions, developmental stage, species, variety, and many other natural influences alter the relative proportions of organic substances in plant material (Buxton et al., 1987; Buxton and Mertens, 1991; Daccord et al., 2001; Jeangros et al., 2001). This leads to a more practical aspect concerning the calibration procedure: For the acquisition of a sufficient number of calibration standards, laborious hand sorting of mixtures is usually necessary for determination of legume concentration in mixtures. To avoid hand sorting and thereby save time, it could be more efficient to use legumes and grasses grown in monoculture as calibration standards. This would ease development of calibration sets but is only possible if the plants grown in monoculture have the same spectral characteristics as plants grown in mixture.

To achieve a model that is robust to natural variations such as changes in species composition or the nutritional status of the plants, most of the occurring variation should be included in the calibration set (Shenk and Westerhaus, 1991). Therefore, it is not surprising that the calibration procedure may need several growing seasons and a minimum number of spectra until the prediction model can be called robust (Dardenne, 1996). Even then, the model is restricted in its applicability because new variation may occur regularly. But once a broad-based calibration set is established, only little adjustments of the calibration model should be necessary by including selected new spectra in the calibration set (Dardenne, 1996). The more variable the calibration set is, the less often such an adjustment will be necessary. In any case, one must be aware of the limitations of calibration models, especially if they are applied to situations that are broader than those represented in the initial calibration set (Brereton, 2000).

Reference samples with known legume content for calibration can be gained by different protocols: (i) References may come from artificially mixed, weighted mixtures (e.g., Pitman et al., 1991) or from real samples taken adjacent to NIRS sampling points (e.g., Coleman et al., 1990; Wachendorf et al., 1999); (ii) the pure samples used to create artificial mixtures may have been grown in mixtures or pure stands. The use of samples from pure stands to create artificial herbage mixtures for calibration was successful for the determination of composite grass or legume concentration but not for the determination of single species therein (Pitman et al., 1991). Coleman et al. (1990) restricted the use of this strategy to closed populations. Introducing intermediate samples, to cover more regularly the whole range of possible legume content, is recommended to be more robust to weak nonlinearities (Naes et al., 2002). So far, our studies on samples from a "closed" population confirmed the potential to use an end-points calibration, i.e., only pure samples for calibration (Locher et al., 2005).

This study was conducted (i) to compare the predictive ability of two narrow-based calibration models, end-points calibration (Model A) vs. a calibration with intermediate samples (Model B), by predicting the legume content of samples of different geographical origins, cutting dates, species composition, and growing conditions and (ii) to check both calibrations against a broad-based calibration (Model C) with intermediate samples comprising most of the variation tested.

MATERIALS AND METHODS

Plant Material

To obtain material for NIRS calibrations and validations, aboveground biomass of diverse legume–grass mixtures was harvested in the years 1999–2002 on seven farms in different regions of Germany (Table 1). These regions differed in soil conditions (FAO classification): luvisols from periglacial sediments (Buchloe), eutric cambisols partly with loess cover (Dürnast, Scheyern, Viehhausen), gleysols from fluviatile sediments (Giessen) and luvisols from loess or loessic loam (Remlingen, Kuernach). The differences in the seeded legume-grass mixture and the varying nature of soil and climate conditions were presumed to result in different characteristics of the legume-grass mixtures, which would function as independent samples for calibration and validation purposes. Clipped samples were separated into legume and grass fractions. Weeds were generally removed, because they were negligible for total yield. Although grass and legume fractions consisted in most cases of different legume or grass species, species composition was not analyzed because only the total legume content was of interest. The dried samples (60°C $>$ 72 h) were ground in a shear mill (Brabender, Duisburg, Germany) to pass a 1.5-mm screen. From the hand-sorted, dry and ground 'pure' legume and grass samples artificial mixtures of known legume content were recombined by weighing. Usually the legume and grass batch of one sample was used to mix one intermediate sample. Only the intermediate samples from August 1999 at Scheyern were mixed from two composite legume and grass batches (Locher et al., 2005). In all cases the intermediate standards covered incrementally the possible range of legume content (0-100%) in roughly 5% increments, i.e., for each selected farm and cutting date a set of 18-21 samples (Table 2). The intermediate standards were used for calibration of Models B and C as shown in Table 2. Further on, they were used for validation the different models (Tables 3–5) as long as they were not used to calibrate them. At the research station of

Table 1. Origin in Germany of calibration and validation samples for analysis of legume content in mixed legume–grass samples. Locations are characterized by their coordinates, the mean annual temperature, and the amount of annual precipitation as described by the closest weather station. The botanical composition of the mixture is defined by the species originally sown.

Origin	Longitude	Latitude	Altitude above sea level	Temperature	Precipitation	Species sown§
			m	°C	mm	
Buchloe	10°42′	48°00′	640	7.0	970	3, 6
Dürnast†	11°43′	48°24′	470	7.5	770	1–8
Giessen	08°49′	50°19′	128	9.3	682	3, 5, 6, 8
Kuernach	10°01′	49°51′	185	8.5	600	1, 6
Remlingen	09°41′	49°48′	185	8.5	600	1–5, 8, 9
Scheyern‡	11°27′	48°23′	458	7.5	800	1–3, 5, 7, 9
Viehhausen	11°37′	48°24′	480	7.5	800	1–5, 8

† Species were sampled from monoculture plots; samples were only used for validation.
‡ Origin of the Calibration sets A and B.
§ 1 = alfalfa (*Medicago sativa* L.), 2 = white clover (*Trifolium repens* L.), 3 = red clover (*Trifolium pratense* L.), 4 = orchardgrass (*Dactylis glomerata* L.), 5 = timothy (*Phleum pratense* L.), 6 = perennial ryegrass (*Lolium perenne* L.), 7 = annual ryegrass (*Lolium multiflorum* L.), 8 = tall fescue (*Festuca arundinacea* Schreb.), 9 = oatgrass (*Arrhenatherum elatius* L.).

the Chair of Plant Nutrition, Technical University of Munich, located at Dürnast, samples were clipped in May, July, and October 2001 (Table 3) from an experiment where all the species normally grown in mixture were growing in monoculture without fertilization. Harvesting pure stands and mixing defined standards from these samples saved the time for laborious hand sorting but introduced some new sources of variability (e.g., the N limitation that the pure grass stands suffered from). Models A, B, and C were compared by two test sets consisting of samples from Remlingen and Kuernach (Table 5), which were not present in any of the three calibrations (Table 2).

Near Infrared Reflectance Spectroscopy

Log $1/R$ (R = reflectance) spectra were taken with a Fourier Transform Near Infrared Reflectance Spectrometer (FT-NIRS, Vector 22/N, Bruker, Ettlingen, Germany) coupled to an external integration module. A rotating sample cup (9 cm diam.) was used to present the samples (>10 g) to the measurement area (2.0 cm diam.). A metal stamp (822 g) was used to compress the sample to a uniform sample density and to avoid any influence of external light. Spectra from diffuse reflection were recorded by a PbS detector between 10 000 and 3500 cm^{-1} (1000–2857 nm).

Measurement conditions were tested to ensure a high signal/noise ratio (Broad et al., 2002) and a high resolution at acceptable measurement duration, which is reflected by the number of scans. For the legume/grass samples, a spectral resolution of 10 cm^{-1} and an averaged spectrum made of 30 scans were found ideal. By using these sample presentation and measurement conditions, about 44 cm^2 of sample surface was measured. Three replicated measurements of each validation sample were performed for the test sets, and the predicted values were averaged and then compared with the true values to calculate the error figures.

Calibration Procedure

Multivariate calibration was performed with PLSR (Martens and Naes, 1987; Brereton, 2000). Because of noise above 7500 cm^{-1} (1333 nm) and below 3950 cm^{-1} (2532 nm), the spectral region used for calibration by a chemometrics software (OPUS 4.0, Bruker, Ettlingen, Germany) was generally restricted to the range from 7500 to 3950 cm^{-1}. An optimization routine offered by OPUS checking various wave number regions and data pretreatments was run to determine the best calibration algorithm. Three models (A, B, and C) were set up with three different calibration sets (Table 2). Model A was an end-points calibration with "pure" legume and "pure" grass samples representing 100 and 0% legume content while Model B was set up with pure and intermediate samples to cover more regularly the whole range of possible legume content (Locher et al., 2005). With Model C, more variability was introduced to the calibration set by adding intermediate samples of different origins, harvesting dates, and species composition. This was done to enhance robustness and therefore to broaden the applicability.

Calculation of Error Figures

For the evaluation of the performance of the three models, independent test sets were predicted. The prediction results

Table 2. Description of three calibration sets that were used to set up prediction models for the legume content in multispecies mixtures by means of partial least square regression. All models covered the possible range of legume content (0–100%). Calibration standards were developed from legume–grass mixtures.

Model	Calibration set description and strategy	Origin of calibration set	Harvest	Number of samples measured as			*n*†
				Legume	Grass	Mixture	
A	end-points calibration: pure grass (consisting of several grass species) and pure legume samples (consisting of several legume species)	Scheyern	May 1999	36	36		320
		Scheyern	July 1999	36	36		
		Scheyern	Aug. 1999	36	36		
		Scheyern	Oct. 1999	35	18		
		Scheyern	May 2000	36	29		
B	mixed calibration: calibration set of Model A plus 3*21 artificially mixed standards (5% increments)	Scheyern	Aug. 1999			63	388
C	mixed calibration: calibration set of Model B plus artificially mixed standards of different origins and harvesting dates (roughly at 5% increments).	Scheyern	July 2000			21	497
		Scheyern	Sept. 2000			21	
		Giessen	July 2001			18	
		Buchloe	May 2002			20	
		Remlingen	May 2002			20	
		Viehhausen	May 2002			20	

† Number of calibration spectra used to set up the models. The number of calibration spectra does not match the theoretical number because different numbers of outliers were removed during the model development.

Table 3. Comparison of the two partial least squares regression (PLSR) Models A and B. Part I shows the model calibration and cross-validation errors. Part II shows the results of external validations before (first number) and after (second number) bias correction. Samples of different origin or harvesting date were predicted with both models. The predicted samples were not used for calibration. Part III shows the results for samples grown only in mixtures. Compared to part II, the validation test-sets with monoculture samples from Dürnast were excluded.

	Model	*n*†	RMSE‡	Slope§	r^2¶	Bias#	RPD††
			%			%	
			RMSEC/RMSECV‡‡				
Part I	A	334	2.2/2.3				
	B	387	2.3/2.5				
			RMSEP/$SEP_{biascor}$§§				
Part II	A	178	7.9/6.5	1.0/1.0	0.94/0.94	−4.5/0	4
	B	178	7.1/6.5	1.0/1.0	0.94/0.94	−2.9/0	4
			RMSEP/$SEP_{biascor}$				
Part III	A	120	5.9/5.2	0.99/0.99	0.96/0.96	−2.6/0	5
	B	120	7.3/5.7	0.99/0.99	0.96/0.96	−4.4/0	5

† Number of calibration or validation samples.
‡ RMSE, root mean square error.
§ Slope of the line from regression of NIRS predicted on true values.
¶ Coefficient of determination of the line from regression of NIRS-predicted on true values.
# Bias is the mean difference between the true and NIRS predicted legume content values as derived from validation and is naturally zero after bias correction.
†† RPD, ratio of the standard deviation of the reference values to the standard error of prediction using only the $SEP_{biascor}$ (see Eq. [5]).
‡‡ RMSEC, root mean square error as derived from calibration; RMSECV, root mean square error as derived from cross-validation.
§§ RMSEP, root mean square error as derived from external test-set validation; $SEP_{biascor}$, standard error of prediction after bias correction.

were exported to a spreadsheet (EXCEL 7.0, Microsoft Corp., Redmond, WA, USA) and then regressed against the true values.

The RMSEP (a measure for *accuracy*) and the bias (indicating *systematic errors*) were calculated according to Eq. [1] and [2] (Naes et al., 2002).

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_{P_i} - y_i)^2}{n}} \quad [1]$$

$$\text{bias} = \frac{\sum_{i=1}^{n} (\hat{y}_{P_i} - y_i)}{n} \quad [2]$$

where $\hat{y}_{P_i}$ = NIRS predicted values, y_i = true values, and n gives the number of samples. The standard error of prediction (SEP) indicates the *precision* of a model and is defined as the standard deviation of the predicted residuals (Naes et al., 2002). The SEP is therefore lower than the RMSEP because the bias is subtracted and may lead to wrong assumptions on the predictive quality of a model. If the SEP is presented, Naes et al. (2002) recommend that bias also be reported. To prevent confusion due to the different error values used in literature, we added the index "biascor" according to Diller (2002). The standard error of prediction after bias correction ($SEP_{biascor}$) was calculated according to Eq. [3].

$$\text{SEP}_{biascor} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_{P_i} - y_i - \text{bias})^2}{(n-1)}} \quad [3]$$

Equation [4] describes the relation between $SEP_{biascor}$ and RMSEP.

$$\text{RMSEP} \approx \sqrt{\text{SEP}_{biascor}^2 + \text{bias}^2} \quad [4]$$

Table 4. Comparison of external validation of two partial least squares regression (PLSR) Models A and B: In total, 179 samples of different origins or harvesting dates were predicted with Models A and B. The predicted samples were not used for calibration. The prediction accuracy was calculated separately for each sample origin or harvesting date. The validation test sets contained artificially mixed samples covering the possible range of legume content (0–100%) in roughly 5% increments. Each artificially mixed sample was prepared from the legume and grass batch of one sample harvested in field.

		RMSEP/$SEP_{biascor}$§		Slope¶		r^2#		Bias††		RPD‡‡	
		Model used for prediction									
Origin and harvesting date of the test set†	*n*‡	A	B	A	B	A	B	A	B	A	B
		%						%			
Scheyern July 2000	21	7.6/7.4	7.8/7.3	1.01	1.01	0.94	0.95	−1.5	−2.8	4	4
Scheyern Sept. 2000	21	4.6/4.2	4.8/4.2	1.00	0.98	0.98	0.98	−1.7	−2.3	7	7
Dürnast May 2001	20	8.8/6.9	8.0/7.8	1.03	1.04	0.93	0.92	−5.5	1.9	4	4
Dürnast July 2001	20	12.5/7.2	7.8/7.8	1.01	1.00	0.93	0.92	−10.2	−0.4	4	4
Dürnast Oct. 2001	18	11.4/6.3	3.9/3.7	1.09	1.04	0.96	0.98	−9.6	−1.2	5	8
Giessen July 2001	18	4.6/3.6	4.4/3.5	0.97	0.97	0.98	0.98	3.0	2.7	9	9
Buchloe May 2002	20	7.9/2.1	6.9/2.3	1.03	1.04	0.99	0.99	−7.6	−6.5	14	13
Viehhausen May 2002	20	3.3/1.8	7.5/1.8	1.01	1.02	0.99	0.99	−2.7	−7.3	15	12
Remlingen May 2002	20	5.6/3.1	10.3/3.1	0.93	0.93	0.99	0.99	−4.6	−9.8	10	10

† Origin as described in Table 1.
‡ Number of validation samples.
§ RMSEP, root mean square error of prediction; $SEP_{biascor}$, standard error of prediction after bias correction.
¶ Slope of the line from regression of NIRS-predicted on true values.
# Coefficient of determination of the line from regression of NIRS-predicted on true values.
†† Bias is the mean difference between the true and NIRS-predicted legume content values as derived from validation.
‡‡ RPD, ratio of the standard deviation of the reference values to the standard error of prediction using the $SEP_{biascor}$ (see Eq. [5]).

Table 5. Comparison of three partial least squares regression (PLSR) models. Models A and B were established with a calibration set from one farm (narrow calibration sets). Model C was established with a calibration set consisting of plant material from different farms and harvesting dates (broad-based calibration). The predicted samples were not used for the calibration of any model (i.e., external validation). Both validation test sets contained artificially mixed samples covering the possible range of legume content (0–100%) in roughly 5% increments. Each sample was prepared from the legume and grass batch of one sample harvested in field.

	Cross-validation	External validation								Percentage of test-set spectra with MD¶¶			
Model	RMSECV†	Origin of the test set‡	Harvesting date	*n*§	RMSEP¶	Slope#	r^2††	Bias‡‡	RPD§§	>3	>4	>5	>6
	%				%			%		%			
A	2.3	Remlingen	July 2002	19	4.4	0.93	0.98	2.8	7	70	39	30	16
B	2.5			19	5.1	0.95	0.98	3.9	7	56	39	19	14
C	3.7			19	3.2	0.92	0.99	−0.2	9	100	82	70	46
A	2.3	Kuernach	July 2002	20	4.7	1.01	0.99	−4.2	10	98	85	37	15
B	2.5			20	2.9	1.02	0.99	−2.4	13	87	47	12	3
C	3.7			20	5.4	0.97	0.99	−5	10	38	3	0	0

† RMSECV, root mean square error as derived from cross validation.
‡ Origin as described in Table 1.
§ Number of validation samples.
¶ RMSEP, root mean square error of prediction.
# Slope of the line from regression of NIRS predicted on true values.
†† Coefficient of determination of the line from regression of NIRS predicted on true values.
‡‡ Bias is the mean difference between the true and NIRS predicted legume content values as derived from validation.
§§ RPD, ratio of the standard deviation of the reference values to the standard error of prediction (see Eq. [5]).
¶¶ Comparison of the predicted test-set spectra to the mean spectrum of the calibration set; if their standardized Mahalanobis distance (MD) was beyond the cutoff value of 3, 4, 5, or 6, respectively, they were marked as outliers. Data indicate the proportion of test-set samples that are outliers.

Exact equality between $SEP_{biascor}$ and RMSEP is not obtained because $n - 1$ is used in the denominator of $SEP_{biascor}$ instead of n, which is used for RMSEP (Naes et al., 2002).

The slope of the line and the coefficient of determination (r^2) from regression of predicted on actual values were derived from an ordinary least squares fit as offered by EXCEL 97 (Microsoft Corp., Redmond, WA, USA).

The RPD value was calculated according to Eq. [5] (Williams, 1987).

$$\text{RPD} = \frac{s_{\text{ref}}}{\text{SEP}_{\text{biascor}}} \quad [5]$$

where s_{ref} is the standard deviation of the reference values.

For the determination of spectral outliers, the MD was used. The MD is the distance of an independent spectrum to the average of all calibration spectra in the factor space.

RESULTS AND DISCUSSION

Comparison of the Models A and B

Model A (calibration set without intermediate standards) and Model B (calibration set with intermediate standards) resulted in similar root mean square errors of calibration (RMSEC) and root mean square errors of cross-validation (RMSECV; Naes et al., 2002) (Table 3, Part I). These low error figures of about 2% legume content would be more than sufficient for prediction of legume content in N_2 fixation studies. However, the variability represented by the Calibration Sets A and B did not lead to models that were able to predict the independent standards with the same accuracy as proved for the internal validation. When the error figures were calculated across all 179 external samples tested (regardless of origin or harvesting date), the models showed a higher validation error (Table 3, Part II). If the error figures had been similar for cross-validation and external validation, it would have suggested robustness of the existing models (Dardenne, 1996). But this was not achieved and could be the result of the rather narrow calibration sets employed. The slope of the regression line between the predicted and the true values was exactly 1.0 with both models. However, systematic deviations (bias) from true values were observed, which contributed much to the RMSEP of 7 to 8% legume content. Model B showed a smaller bias. This could be interpreted as an effect of the intermediate samples included in the calibration set of Model B, which possibly increased the robustness of the prediction model to new variability. However, in any case, the RMSEP was not satisfactory. Accuracy of prediction is enhanced by subtracting the observed bias from each predicted value (Martens and Naes, 1987; Diller, 2002). This can be done if linearity is observed over the whole range as was the case for our data. Subtracting bias had two effects (Table 3, Part II): The error figures for both models decreased, and differences between the two models disappeared. After bias correction, there was no advantage of Model B with intermediate standards compared with Model A without intermediate standards. The values for RPD figures showed that after bias correction, both models, A and B, resulted in satisfactory predictions according to criteria of Diller (2002).

Influence of Origin and Harvesting Date

If the test sets (i.e., origins or harvesting dates) were analyzed separately without a bias correction, neither of the two models performed consistently better than the other (Table 4). The external validations showed for each test set a RMSEP that was two- to fivefold higher than the RMSECV (Table 3, Part I). This clearly reflects differences in the plant material that combine effects of site-specific growing conditions, different legume species, grass species, varieties, or plant age. Also, differences in the sample preparation, e.g., residual moisture, storage conditions, or sample age are known to cause systematic prediction errors (Fales and Cummins, 1982). However, this is unlikely in our case as we kept the sample preparation and storage conditions as con-

stant as possible for both calibration and validation samples. We assume that the remaining variation in residual moisture was also present in the calibration set, and models derived from samples including that variation should then be independent of minor changes. So the systematic errors were presumably caused by the samples themselves and not by their preparation. The worst predictions with Model A were for those samples that were grown in monoculture (origin: Dürnast). For this additional variability, Model B was superior, and the calibration strategy with intermediate samples resulted in more accurate predictions compared with the end-points strategy of Model A. But in general, both models resulted in less accurate predictions for those untypical test sets, apart from the test set of October 2001, which was sufficiently well predicted by Model B. Obviously, samples from Dürnast collected from pure stands were not represented by samples of both calibration sets, which were all derived from mixtures. The question arises as to why both models performed so differently with these test sets. One reason may be the N limitation of the grasses later in the season; this is not found in grasses grown in mixtures with legumes due to higher N availability (Ta and Faris, 1987). One could conclude that Model A (end-points calibration) gives higher weights to N-sensitive wavelengths because one main difference between grasses and legumes will be their N content. Thus, Model A could be partly based on an autocorrelation of N content and legume content. This kind of autocorrelation will be reduced by calibrating with intermediate mixtures that are created from plant material with similar N contents. The intermediate standards of Calibration B were mixtures of grass and legume fractions from one harvesting date (Aug. 1999) where grasses and legumes had a similar N content (3.2 and 3.4%). Automatically higher weights may be given to other distinguishing wavelength regions during the PLSR. This could be an explanation for the better performance of Model B with respect to monoculture samples. This explanation is strengthened by the results in Table 3, Part III, where monoculture samples were excluded from the validation set. Then Model A had a lower RMSEP across all remaining test sets than Model B compared with Table 3, Part II. After bias correction, both models again resulted in similar $SEP_{biascor}$. For the uncorrected prediction values, Model A was more accurate and robust than Model B when only plant material grown in mixtures was tested. But Model A seemed to be more sensitive to potentially new variation introduced by N limitation. However, it is noteworthy that the lower N content of the grass is only one possible explanation. Of course, other near infrared–relevant plant parameters are also affected by N deprivation, but their contribution to the erroneous prediction cannot be discussed here because no specific information was collected. Whatever may have caused the poor results for the monoculture standards, these findings indicate that a simplified calibration based on such samples will most likely not result in good predictions for plants grown in mixture.

The predictions for all single test sets were substantially biased, but the precision was generally high at least for standards that were derived from mixtures (Table 4). Apart from the monoculture standards, the only test set consisting of plant material grown in mixtures that was predicted with insufficient precision was that of Scheyern harvested in July 2000. The bias correction did not lower errors unlike for the other test sets. It is unclear what lowered the precision. An extraordinarily long growing period (9 wk) of this legume–grass compared with that represented in the calibration set (average growing period 6 wk) or exceptionally dry weather conditions during the growing period may serve as an explanation. All remaining test sets consisting of plant material grown in mixture could be accurately predicted after a bias correction. Therefore, collection and remixing of a representative set of samples for each new harvest or origin as described above are evidently necessary for calculation of bias. Subtracting the bias from the predicted values enhanced the accuracy of prediction. The bias correction appears to be inconvenient but reasonable, as for each origin or harvesting date, excellent precision, linearity over the whole range, slopes of the regression line close to 1.0, and high r^2 values were observed. Furthermore, by doing such reference analyses, new standards will be obtained by which an existing model can be broadened in its applicability. After several years, broadening the calibration set continuously with those standards could lead to a model that shows unbiased predictions in the future and that will then be robust to additional variability. As long as this is not achieved, a bias correction has to be recommended based on our data.

Comparison of Models A, B, and C

To test a broad and possibly more robust model, a third calibration set was established by including standards from different origins and harvesting dates (see Table 2). This Model C was expected to have a broader applicability and minor bias in the prediction of new samples compared with the narrow-based Models A and B. The higher variability included in the calibration set increased RMSECV as expected (Table 5). The best model was calculated by the same data pretreatment (first derivative and vector normalization), and the multidimensional data were reduced to 10 factors as with Models A and B. Besides, the three models were slightly different with respect to uninformative spectral regions that were given zero weight for prediction. Two test sets (origin: Remlingen and Kuernach, July 2002) were kept for the external validation of the three models (Table 5) to allow direct comparison of the three models. Analysis of variance (ANOVA) proved that the differences in the predicted values of both test sets were not significant for the three models (data not shown). Though having higher calibration errors, the accuracy of Model C was comparable to Models A and B, but there was no consistent reduction in bias observed with Model C. We conclude that Model C was still too narrow to be robust against the new variation introduced by the test sets. Of course, only two test sets were used to compare the three models, and

the general predictive ability of models should probably not be assessed from test sets that cannot represent all possible variability (Martens and Naes, 1987). But the two test sets showed that Model C, which theoretically should cover more variability than Models A and B, was not automatically superior to Model A or B. Whereas Model C gave the best predictions for the multispecies mixture of Remlingen, it was worst for the binary mixture of Kuernach with respect to the RMSEP. Therefore, to obtain accurate values, the procedure of bias correction is even necessary with Model C. In general, however, the necessity of the bias correction depends on the purpose of the NIRS estimations. The excellent linearity between true and predicted values confirms that all three models are adequate to describe the spatial variability of the legume content in the field, regardless of systematic errors. Furthermore, compared with the possible range of legume content in the field (0–100%), the difference among the models in predicting power is negligible. Theoretically, Model C should then be preferred because it represents the highest variability of the three models. But, from a practical point of view, Models A and B are to be preferred because the same results were achieved with much less effort in sampling. Furthermore, in case of the Model A, even mixing of intermediate standards as a potential source of error is avoided.

Since the introduced Models A, B, and C do represent in that sequence an increasing variability of samples, this should be the case for the spectra as well. Therefore, the likelihood that new spectra (=samples) will fit to the population determined by the calibration set should increase from Model A to Model C, and consequently the proportion of outliers, which are determined by the MD, should decrease from Model A to Model C. However, the expected general reduction in the proportion of outliers within a test set for Model C compared with Model A and B was not observed (Table 5). Model C showed the lowest RMSEP with the highest proportion of outliers in the test set of Remlingen whereas for the other test set, the RMSEP was worst with the lowest proportion of outliers. These contradictory results showed that the accuracy of the predictions was not necessarily affected by high spectral distances. We conclude that the high MDs of the test sets were generated in spectral regions that were less important for the prediction of the legume content. The robustness of the models could be enhanced by excluding these regions and by calculating the models anew. However, the selection of causal wave numbers represents a difficult task, particularly with full-spectrum multivariate techniques such as PLSR (Osborne et al., 1997) and was outside the scope of this investigation. If we consider both, accuracy of the predictions on one side and the proportion of outliers on the other, a standardized MD of 6 or even 7 seems to be justified as the cutoff value for the detection of outliers in predicting legume content, which is larger than reported (Shenk and Westerhaus, 1991).

CONCLUSIONS

In this study, NIRS was challenged by testing different models with independent samples, i.e., models were derived from a closed population and then applied to an open population. A broad-based data set showed that near infrared technology offers highly accurate and easy predictions of the legume content in multispecies mixtures even for open populations. Origin, harvest, and species composition caused biased predictions but did not affect linearity. To enhance accuracy, a bias correction for each new harvesting date seemed necessary based on our findings. Broadening the variation in the calibration set did not result in less biased predictions. Therefore, it is concluded that a useful calibration for the determination of the legume content in multispecies mixtures can be set up even within 1 yr, but reference analyses for control of the results are necessary. However, one has to decide which degree of accuracy is desired for the parameter in question and for the purpose of the final application: e.g., for the detection of areas with different legume content in the field, precise but biased predictions are sufficient, even if high accuracy would be desirable. In this regard, all of the three tested models proved to be adequate for the description of spatial variability of the legume content in the field regardless of origin, growing conditions, or species composition.

ACKNOWLEDGMENTS

We are grateful to the staff of the Research Station at Dürnast for harvesting and sample processing and to Dr. K. Möller (University of Giessen) and M. Helmert for providing samples for method validation. Many thanks to the reviewers and the associate editor for their most valuable suggestions. This research was part of the scientific activities of the Research Network Munich on Agroeosystems (FAM), which are supported by the German Federal Ministry of Education and Research (BMBF 0339370), Berlin, Germany. Rent and operating expenses of the Research Station Scheyern are paid by the Bavarian State Ministry for Education and Culture, Science and Art, Munich, Germany.

REFERENCES

Boller, B.C., and J. Nösberger. 1987. Symbiotically fixed nitrogen from field-grown white and red clover mixed with ryegrass at low levels of ^{15}N-fertilization. Plant Soil 104:219–226.

Brereton, R.G. 2000. Introduction to multivariate calibration in analytical chemistry. Analyst (Cambridge, UK) 125:2125–2154.

Broad, N., P. Graham, P. Haily, A. Hardy, S. Holland, S. Hughes, D. Lee, K. Prebble, and P. Warren. 2002. Guidelines for the development and validation of near-infrared spectroscopic methods in the pharmaceutical industry. p. 1–21. *In* J.M. Chalmers and P.R. Griffiths (ed.) Handbook of vibrational spectroscopy. Vol. 5. John Wiley & Sons, Chichester, UK.

Buxton, D.R., and D.R. Mertens. 1991. Errors in forage-quality data by near infrared spectroscopy. Crop Sci. 31:212–218.

Buxton, D.R., J.R. Russell, and W.F. Wedin. 1987. Structural neutral sugars in legume and grass stems in relation to digestibility. Crop Sci. 27:1279–1285.

Coleman, S.W., F.E. Barton II, and R.D. Meyer. 1985. The use of near-infrared reflectance spectroscopy to predict species composition of forage mixtures. Crop Sci. 25:834–837.

Coleman, S.W., S. Christiansen, and J.S. Shenk. 1990. Prediction of botanical composition using NIRS calibrations developed from botanically pure samples. Crop Sci. 30:202–207.

Daccord, R., Y. Arrigo, B. Jeangros, J. Scehovic, F.X. Schubiger, and J. Lehmann. 2001. Nutritional value of pasture plants: Content of cell wall components. (In German with English abstract.) Agrarforschung 8(4):180–185.

Dardenne, P. 1996. Stability of NIR spectroscopy equations. NIR news 7(5):8–9.
Dardenne, P., G. Sinnaeve, and V. Baeten. 2000. Multivariate calibration and chemometrics for near infrared spectroscopy: Which method? J. Near Infrared Spectrosc. 8:229–237.
Diller, M. 2002. Investigations for the development of a NIRS-method for potatoes in organic farming with special reference to the influence of the year and the potato line. (In German.) Ph.D. thesis. Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany.
Fales, S.L., and D.G. Cummins. 1982. Reducing moisture induced error associated with measuring forage quality using near infrared reflectance. Agron. J. 74:585–588.
Jeangros, B., J. Scehovic, F.X. Schubiger, J. Lehmann, R. Daccord, and Y. Arrigo. 2001. Nutritional value of pasture plants: Dry matter-, crude protein- and sugar content. (In German with English abstract.) Agrarforschung 8(2):1–8.
Locher, F., H. Heuwinkel, R. Gutser, and U. Schmidhalter. 2005. Development of near infrared reflectance spectroscopy calibrations to estimate legume content of multispecies legume–grass mixtures. Agron. J. 97:11–17 (this issue).
Martens, H., and T. Naes. 1987. Multivariate calibration. John Wiley & Sons, New York.
Naes, T., T. Isaksson, T. Fearn, and T. Davies. 2002. A userfriendly guide to multivariate calibration and classification. NIR Publ., Chichester, UK.
Osborne, S.D., R.B. Jordan, and R. Künnemayer. 1997. Method of wavelength selection for partial least squares. Analyst (Cambridge, UK) 122:1531–1537.
Park, R.S., R.E. Agnew, F.J. Gordon, and R.W.J. Steen. 1998. The use of near infrared reflectance spectroscopy (NIRS) on undried samples of grass silage to predict chemical composition and digestibility parameters. Anim. Feed Sci. Technol. 72:155–167.
Petersen, J.C., F.E. Barton II, W.R. Windham, and C.S. Hoveland. 1987. Botanical composition definition of tall fescue–white clover mixtures by near infrared reflectance spectroscopy. Crop Sci. 27: 1077–1080.
Pitman, W.D., C.K. Piacitelli, G.E. Aiken, and F.E. Barton II. 1991. Botanical composition of tropical grass–legume pastures estimated with near-infrared reflectance spectroscopy. Agron. J. 83:103–107.
Shaffer, J.A., G.A. Jung, J.S. Shenk, and S.M. Abrams. 1990. Estimation of botanical composition in alfalfa/ryegrass mixtures by near infrared reflectance spectroscopy. Agron. J. 82:669–673.
Shenk, J.S., and M.O. Westerhaus. 1991. Population definition, sample selection, and calibration procedures for near infrared reflectance spectroscopy. Crop Sci. 31:469–474.
Ta, T.C., and M.A. Faris. 1987. Effects of alfalfa proportions and clipping frequencies on timothy–alfalfa mixtures: I. Competition and yield advantages. Agron. J. 79:817–820.
Wachendorf, M., B. Ingwersen, and F. Taube. 1999. Prediction of the clover content of red clover– and white clover–grass mixtures by near-infrared reflectance spectroscopy. Grass Forage Sci. 54:87–90.
Williams, P.C. 1987. Variables affecting near-infrared transmittance spectroscopic analysis. p. 143–148. *In* P.C. Williams (ed.) Near-infrared technology in the agricultural and food industries. Am. Assoc. of Cereal Chemists, St. Paul, MN.