Effect of Calibration Set Selection on Quantitatively Determining Test Weight of Maize by Near-Infrared Spectroscopy

Lianping Jia, Peng Jiao, Junning Zhang, Zhen Zeng, Xunpeng Jiang

Effect of Calibration Set Selection on Quantitatively Determining Test Weight of Maize by Near-Infrared Spectroscopy

Lianping Jia[1], Peng Jiao[1], Junning Zhang[2], Zhen Zeng[3], and Xunpeng Jiang[1(✉)]

[1] COFCO Feed Co., Ltd., Beijing 100088, China
{jialianping,jiaopeng,jiangxunpeng}@cofco.com
[2] Chinese Academy of Agricultural Mechanization Sciences, Beijing 100083, China
zjn990210@163.com
[3] Patent Examination of Cooperation Beijing Center of the Patent office,SIPO, Beijing 100081, China
360830219@qq.com

Abstract. To study the effect of calibration set on quantitatively determining test weight of maize by near-infrared spectroscopy, 584 maize samples were collected and scanned for near-infrared spectral data. Test weight was measured following the standard GB 1353-2009, resulting the sample test weight of 693~732 g•L^{-1}. Two calibration models were respectively built using partial least squares regression, based on two different calibration sets. Test weight of two calibration sets distribute differently, with normal and homogeneous distributions. Both quantitative models were selected by root mean square error of cross validation (RMSECV), and evaluated by validation set. Results show the RMSECV of the model based on normal distribution calibration set is 4.28 g•L^{-1}, the RMSECV of the model based on homogeneous distribution calibration set is 2.99 g•L^{-1}, the predication of two models have significant difference for the samples with high or low test weight.

Keywords: Test weight, Near-infrared, Calibration set selection

1 Introduction

Maize is a major energy ingredient for livestock feed in China, with more than 50% of maize used for feed [1]. The characteristic of maize varies greatly due to geography, species and other factors. To evaluate the quality of maize for feed, many standards adopt test weight as a certification and ranking indicator [2]. Test weight is defined as a measurement of bulk density or the weight of a unit volume of grain (g•L^{-1}). Generally, maize of low-level test weight is of lower feeding value than that of normal test weight [3]. The authoritative approach of determining test weight of maize involves measuring the weight of corn cereal per standard volume. Measurements normally be taken around 5 minutes with an accuracy of ±9 g•L^{-1} [4]. The maize utilized in Chinese feed industry is not only from local province, but also transported from other provinces and countries. The laboratory-based method would not be feasible for achieving this uncertain variation of different country maize to safeguard product quality stabilization at high frequency and a large volume.

Near-infrared spectroscopy (NIRS) is a non-destructive, non-pollution, fast and easily applicable technique, and it has already been widely applied to the routine analysis in the feed industry [5]. The NIRS normally determines the chemical properties of the feed, such as moisture, crude protein, crude fiber, crude fat, amino acid, vitamin, and so on [6,7]. It also shows a promising potential on predicting the physical properties of feed ingredients, such as test weight. Pomeranz [8] pointed out the high correlation

between maize test weight and 1680 nm near-infrared absorption peak. Siska and Hurburgh [9] studied the relationship between maize test weight and NIR transmit spectra. Li [10] built a NIR determination model of maize weight test with RMSECV of 8.68 g•L^{-1}. The precision and accuracy of a prediction equation based on NIR is contingent on the construction of a representative calibration set. Some methods, including random method, content grads method, Kennard-Stone algorithm (KS), and Sample set Partitioning based on joint X-Y distance (SPXY), have been developed to select samples for calibration set. Most research have focused on selecting nearest neighbor samples and getting rid of abnormal spectrum. Seldom research considers the influence of the distribution of prediction indicator.

The objective of this study is to compare the difference of NIR prediction models, which are built on two (normal and homogeneous) distributions of maize test weight of calibration set respectively.

2 Materials and Methods

2.1 Maize Samples

In total, 586 maize samples were obtained from different places of origination, including Heilongjiang, Jilin, Inner Mongolia, Hebei and Shandong province, et al. Moisture content of the sample was controlled under 14% [11]. All the samples were divided into two parallels by the quartering method. One was used to analysis test weight, the other was used to obtain NIR spectrum. Test weight determination was adept in the laboratory according to Chinese standard method of GB 1353-2009 Appendix A. Samples for NIR acquiring were milled by Retsch ZM200, and passed 1.0 sieve. All samples were stored in dry and room temperature (25 ℃).

2.2 NIR Spectrum Acquisition

A Fourier transform near-infrared spectrometer (Bruker Matrix-I, Germany） was adapted to collect the diffuse reflectance spectrum under a gold plated integrating sphere with PbS detector. Light source of high energy air cooling pre-collimation near infrared light was employed. The scanning parameters are follows: spectral wavenumber range of 10000～4000 cm^{-1}, spectral resolution of 16 cm^{-1}, scanning times for each spectrum of 64, sample pool mode of rotary. Considering of the spectral perturbation from packing density, each sample was scanned three times by three different technical staffs. Average spectra of three spectra from same sample was given to develop the calibration equations. To avoid instrumental bias in scanning process, all samples were analyzed in random sequence. Each sample was scanned 3 times and average spectrum was used.

2.3 Calibration and Validation Sets

584 samples were selected, and used for model res resulting a wide range of test weight of 693~732 g•L^{-1}. These samples were divided into a calibration set (437 samples) and a validation set (remaining 147 samples) by the method of Kenard-Stone algorithm [12] based on Euclidean distance.

Samples for 2 sub-sample sets, which were of normal distribution or homogeneous distribution, were selected from the calibration set of 437 samples. 320 samples following homogeneous distribution were also selected from same calibration set to compose homogeneous distribution calibration sample set, shorted as Set H; 320 samples following normal distribution were selected from calibration set to compose normal distribution calibration sample set, shorted as Set N.

2.4 NIR Data Analysis

There are 3 pretreatment methods selected to pretreat spectral data of maize sample, including first derivative (1D), second derivative (2D) and standard normal variate (SNV). This mathematical treatment will help diminishing baseline offset originating from maize sample particle size and packing density.

Partial least squares (PLS) discriminant analysis, as one of most acceptable approaches, was selected as the calibration method. Cross validation, such as leave one out, was operated to evaluate the best number of factors in the optimal equation. Calibration modeling statistics reckoned included the root mean square error of calibration (RMSEC), the coefficient of multi-determination in calibration (R^2_C), the root mean square error of cross validation (RMSECV) and the coefficient of determination in cross-validation (R^2_{CV}). RMSECV, which is one of key statistic parameters, will be adopted to select the optimal equation. Model of homogeneous distribution (shorted for model H) and Model of normal distribution (shorten as model N) were built based on Set H and Set N, respectively. The validation set verify the capacity of calibration equation to avoid over fitting or under fitting. Calibrations were developed using OPUS software version 7.5.

3 Results and Discussion

3.1 Statistical Analysis of Maize Test Weight

Table 1 shows the descriptive statics of test weight of maize in different sample sets. Prior to GB 1353-2009, selected 584 maize samples (693~732 g•L^{-1}) cover three grades from grade 1 to grade 3. There is no significant difference between Set N and Set H in 5 statics parameters of number of samples, minimum value, maximum value, mean value and median value. The standard derivations of two set are 12 g•L^{-1} and 10 g•L^{-1}, respectively. 147 samples in the validation set belong to the same range of test weight as the calibration set. However, test weight of set N and set H followed different distribution, as shown in Figure 1. Set H is homogeneous distribution, Set N is normal distribution under 95% confidence interval (P=0.11>0.05).

Table 1. Descriptive statics of test weight of maize in different sample sets

	Number	Min (g·L^{-1})	Max (g·L^{-1})	Mean (g·L^{-1})	Median (g·L^{-1})	SD (g·L^{-1})
Selected samples	584	693	732	713	714	11
Calibration set	437	693	732	713	715	11
Set H	320	693	732	713	712	12
Set N	320	693	732	713	712	10
Validation set	147	693	732	711	711	11

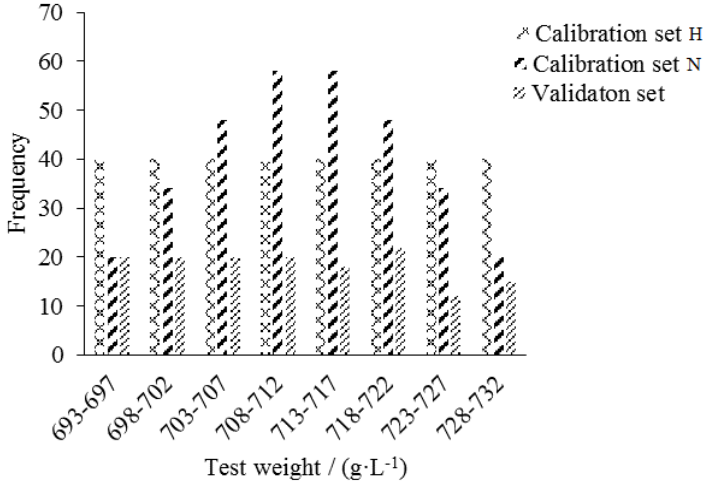

Fig. 1. Histogram of test weight of maize in different sample sets

3.2 NIR Spectra of Maize

Figure 2 is the 584 NIR spectra of maize samples. There are some regions characteristic of nutrition component absorption. The bands at 5787~5690 cm^{-1} and 4340~4261 cm^{-1} are related to the content in fat. Water, protein and starch absorption have high correlations with bands at 5200~5100 cm^{-1}, 4867~4400 cm^{-1} and 4400~4300 cm^{-1}, respectively [13-15]. Other important band to remark is C–H first overtones at 5882 cm^{-1}.

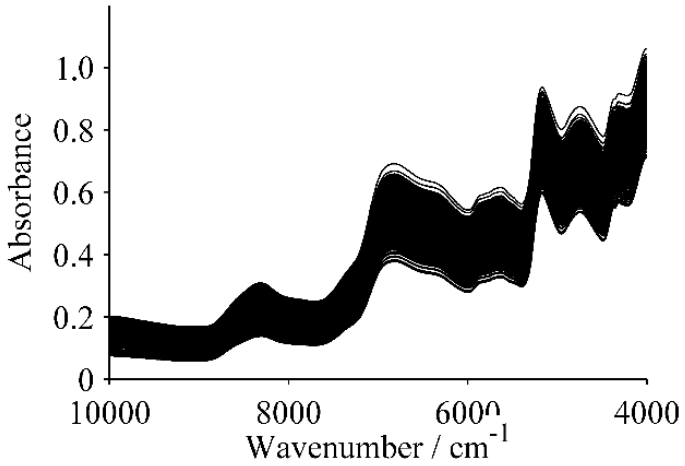

Fig. 2. Near-infrared spectra of maize

3.3 Model of Homogeneous Distribution

Model of homogeneous distribution is developed on Set H using PLS. Table 2 shows effects of using different pretreatments and spectral region on Model H. Comparing the various data pretreatments investigated, the best spectral range is 9404~5446 cm^{-1} and 4606~4243 cm^{-1} with a SNV preprocessing. The parameters of model are following: RMSECV of 4.28 g•L^{-1}, R^2_{CV} of 0.84 and RPD of 2.49. RPD value reveals this model is feasible for quantitative analysis [16]. SNV is more suitable for preprocessing than other algorithms, it is believed that particle size, surface scattering and optical path change have certain effect on maize near infrared spectra [17].

Table 2. Effects of using different pretreatments and spectral region on Model H

Spectral range	pretreatments	R^2_C	RMSEC	R^2_{CV}	RMSECV	RPD
9404~5446 cm^{-1}, 4606~4243 cm^{-1}	1D(9)*	0.90	3.48	0.77	5.15	2.07
	1D(13)	0.86	4.15	0.77	5.12	2.09
	2D(13)**	0.79	5.00	0.66	6.24	1.71
	1D(9)+SNV***	0.91	3.37	0.81	4.66	2.29
	SNV	0.90	3.46	0.84	4.28	2.49
10000~4000 cm^{-1}	1D(9)	0.87	3.9	0.77	5.10	2.09
	1D(13)	0.88	3.9	0.78	5.01	2.13
	2D(13)	0.86	4.16	0.67	6.11	1.75
	1D(9)+SNV	0.90	3.51	0.80	4.80	2.22
	SNV	0.89	3.75	0.79	4.88	2.19

Note: *1D(9): First derivative with 9 data points.window width.
**2D(13):Second derivative with 13 data points.window width.

***1D(9)+SNV: firstly treated by first derivative with 9 data points.window width, and then treated by SNV.

3.4 Model of Normal Distribution

Model of normal distribution is developed on Set N using PLS. Table 3 shows effects of using different pretreatments and spectral region on Model N. Comparing the various data pretreatments investigated, the best spectral range is 9404~5446 cm^{-1} and 4606~4243 cm^{-1} with a SNV preprocessing. The parameters of model are following: RMSECV of 2.99 $g•L^{-1}$, R^2_{CV} of 0.84 and RPD of 2.51. RPD value reveals this model is feasible for quantitative analysis. RMSECV of model B is a little lower than that of Model A, which means the better prediction.

Table 3. Effects of using different pretreatments and spectral region on Model N

Spectral range	pretreatments	R^2_C	RMSEC	R^2_{CV}	RMSECV	RPD
9404~5446 cm^{-1}, 4606~4243 cm^{-1}	1D(9)*	0.82	3.27	0.69	4.15	1.81
	1D(13)	0.83	3.17	0.69	4.21	1.78
	2D(13)**	0.84	3.09	0.65	4.44	1.70
	1D(9)+SNV***	0.82	3.27	0.70	4.14	1.82
	SNV	0.95	1.79	0.84	2.99	2.51
10000~4000 cm^{-1}	1D(9)	0.89	2.55	0.72	3.98	1.89
	1D(13)	0.91	2.41	0.75	3.77	2.00
	2D(13)	0.95	1.83	0.68	4.28	1.76
	1D(9)+SNV	0.93	2.00	0.74	3.82	1.97
	SNV	0.91	2.42	0.78	3.50	2.15

Note: *1D(9): First derivative with 9 data points.window width.
**2D(13):Second derivative with 13 data points.window width.
***1D(9)+SNV: firstly treated by first derivative with 9 data points.window width, and then treated by SNV.

3.5 Validation Set

Figure 3 show the absolute derivation of validation set predicted from two models. Samples with test weight range in 693~697 $g•L^{-1}$, the absolute prediction error of model H is lower than that of model N. Samples with test weight range in 693~727 $g•L^{-1}$, the absolute prediction error of model H is lower than that of model N. Samples with test weight range in 728~732 $g•L^{-1}$, the absolute prediction error of model H is higher than that of model N. Model N have a more accurate quantify prediction for maize sample with low or high test weight. However, prediction error of model H is almost same in the whole range of test weight. It could be seen that the test weight distribution of calibration set influence the NIR equation model and its properties.

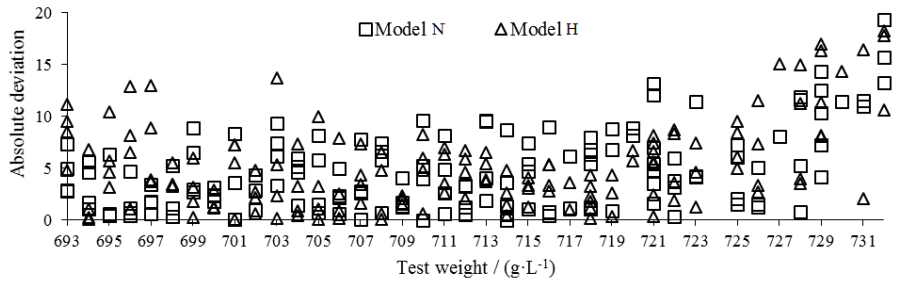

Fig. 3. Prediction results obtained by different models

4 Conclusion

The aim of the work reported here was to study effect of calibration set selection on

quantitatively determining test weight of maize by near-infrared spectroscopy. Results show the RMSECV of the model based on normal distribution calibration set is 4.28 $g \cdot L^{-1}$, the RMSECV of the model based on homogeneous distribution calibration set is 2.99 $g \cdot L^{-1}$, the predication of two models have significant difference for the samples with high or low test weight.

Acknowledgements

This work was supported financially by the National key research and development program (Project No. 2016YFD0700204) and S&T Nova Program of Beijing (Project No. Z1511000003150116).

References

1. China Feed Industry Association:China Feed Industry Yearbook.(2015) (in Chinese)
2. Wang B.J., Yang X.L., Qi B.Z.:Comparison of quality safety standards of maize in china with other foreign country. Review of China Agricultural Science and Technology. 4(5),10-14(2002) (in Chinese)
3. Johnston L.J.:Use of low-test-weight corn in swine diets and the lysine/protein relationship in corn. Journal of Swine Health & Production. 3(4),161-164(1995)
4. Zhang X.:Discussion on measuring result error of corn bulk density. Grain Distribution Technology. 2,27-28(2005) (in Chinese)
5. Perez-Marin D., Fearn T., Guerrero J.E., Garrido-Varo A.:A methodology based on NIR-microscopy for the detection of animal protein by-products. Talanta. 80 (1), 48-53(2009)
6. Chen L.J., Yang Z.L., Han L.J.:A review on the use of near-infrared spectroscopy for analyzing feed protein materials. Applied Spectroscopy Reviews. 48(7),509-522(2013)
7. Chu X.L.，Lu W.Z.:Research and application progress of near infrared spectroscopy analytical technology in china in the past five years. Sepctroscopy and Spectral Analysis. 34(10),2595-2605(2014) (in Chinese)
8. Pomeranz Y., Hall G.E., Czuchajowska Z., Lai F.S.:Test weight, hardness, and breakage susceptibility of yellow dent corn hybrids. Cereal Chemistry. 63(4),349-351(1986)
9. Siska J., Hurburgh C.T.:Corn density measurement by near-infrared transmittance. Transactions of the ASAE. 38(6),1821-1824(1995)
10. Li J.T.:Study on the rapid evaluation of nutrient values of corn and wheat by near-infrared reflectance spectroscopy.Beijing: China Agricultural University. (2014) (in Chinese)
11. Cai H.Z., Bi W.Q., Chu J.Z., Li H.H.:Study on the relationship between maize unit weight and the water content. Journal of Zhengzhou Institute of Technology. 22(3),70-72(2001).(in Chinese)
12. Kenard R.W., Stone L.:Computer aided design of experiments. Technometrics. 11,137-148(1969)
13. Pérez-Marín D.C., Garrido-Varo A., Guerrero-Ginel J.E., Gómez-Cabrera A.:Near-infrared reflectance spectroscopy (NIRS) for the mandatory labelling of compound feedingstuffs: chemical composition and open-declaration. Animal Feed Science and Technology. 116(3-4),333-349(2004)
14. Roza-Delgado B., Soldado A., Garrido-Varo A.M., Pérez-Marín D., Haba M.J., Guerrero-Ginel J.E.:Application of near-infrared microscopy (NIRM) for the detection of meat and bone meals in animal feeds: A tool for food and feed safety. Food Chemistry. 105(3),1164-1170(2007)
15. Yang Z.L., Han L.J., Li Q.F., Fan X.:Use of discriminant analysis on NIRS to detect meat-and-bone meal content in ruminant concentrates. Journal of Animal and Feed Sciences. 16,442-447(2007)
16. Malley D.F., Mcclure C., Martin P.D.:Compositional Analysis of Cattle Manure During Composting Using a Field‐Portable Near‐Infrared Spectrometer. Communications in Soil Science and Plant Analysis. 36(4),455-475 (2005)
17. Chu X.L., Yuan H.F., Lu W.Z.:Progress and application of spectral data pretreatment and wavelength selection methods in NIR analytical technique. Progress in Chemistry. (4),528-542(2004) (in Chinese)

18. Czajkowski, K., Fitzgerald, S., Foster, I., Kesselman, C.: Grid Information Services for Distributed Resource Sharing. In: 10th IEEE International Symposium on High Perfor- mance Distributed Computing, pp. 181–184. IEEE Press, New York (2001)
19. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: The Physiology of the Grid: an Open Grid Services Architecture for Distributed Systems Integration. Technical report, Global Grid Forum (2002)
20.
21.
22. .....