Monitoring model for predicting maize grain moisture at the filling stage using NIRS and a small sample size

Xue Wang[1,2], Tiemin Ma[2,3], Tao Yang[1*], Ping Song[1], Zhengguang Chen[2], Huan Xie[2]

(1. *College of Information and Electrical Engineering*, *Shenyang Agricultural University*, *Shenyang* 110866, *China*;
2. *College of Electrical and Information*, *Heilongjiang Bayi Agricultural University*, *Daqing* 163319, *China*;
3. *School of Computer Science And Engineering*, *Northeastern University*, *Shenyang* 110819, *China*)

Abstract: The change in the maize moisture content during different growth stages is an important indicator to evaluate the growth status of maize. In particular, the moisture content during the grain-filling stage reflects the grain quality and maturity and it can also be used as an important indicator for breeding and seed selection. At present, the drying method is usually used to calculate the moisture content and the dehydration rate at the grain-filling stage, however, it requires large sample size and long test time. In order to monitor the change in the moisture content at the maize grain-filling stage using small sample set, the Bootstrap re-sampling strategy-sample set partitioning based on joint *x*-*y* distances-partial least squares (Bootstrap-SPXY-PLS) moisture content monitoring model and near-infrared spectroscopy for small sample sizes of 10, 20, and 50 were used. To improve the prediction accuracy of the model, the optimal number of factors of the model was determined and the comprehensive evaluation thresholds RVP (coefficient of determination (R^2), the root mean square error of cross-validation (RMSECV) and the root mean square error of prediction (RMSEP)) was proposed for sub-model screening. The model exhibited a good performance for predicting the moisture content of the maize grain at the filling stage for small sample set. For the sample sizes of 20 and 50, the R^2 values were greater than 0.99. The average deviations of the predicted and reference values of the model were 0.1078%, 0.057%, and 0.0918%, respectively. Therefore, the model was effective for monitoring the moisture content at the grain-filling stage for a small sample size. The method is also suitable for the quantitative analysis of different concentrations using near-infrared spectroscopy and small sample size.

Keywords: moisture content monitoring, maize, growth stage, near-infrared spectroscopy (NIRS), small sample set, model screening, optimal factor number, Bootstrap-SPXY-PLS

DOI: 10.25165/j.ijabe.20191202.4708

Citation: Wang X, Ma T M, Yang T, Song P, Chen Z G, Xie H. Monitoring model for predicting maize grain moisture at the filling stage using NIRS and a small sample size. Int J Agric & Biol Eng, 2019; 12(2): 132–140.

1 Introduction

The moisture content of maize grains is an important indicator for determining the timing of mechanical harvesting, predicting the yield, grading, and safe storage[1-3]. The development of new maize varieties suitable for mechanical harvesting is a major research focus in maize breeding. Therefore, the monitoring and testing of the moisture content at the grain-filling stage are important for crop growth and breeding, Accurate moisture content monitoring not only enables crop managers to deal with water stress caused by external factors[4] and to predict the maturity and quality of the maize ears but also serves as a tool used by crop experts for breeding and seed selection.

In maize breeding, seed and its parents are very valuable[5]. Because self-pollination of maize ears is rare, dozens or even a dozen grains in a maize mature ears. The moisture content at maize grain-filling stage is always higher than 30%; therefore, drying methods are required. During drying, only 150-200 grains in the middle of the ears can be dried for moisture content measurements[6]. Accordingly, a large sample size, equipment, and handling time are required for determining the moisture content. However, in maize breeding trials, the planting area is usually small and a small number of maize ears are available. Therefore, the sample is commonly low for moisture content measurements at maize grain-filling stage.

Near-infrared spectroscopy (NIRS) combined with chemometrics has become a popular technique for quantitative analysis[7-9], quality detection[10-12], and identification of seed varieties[13]. As a result of the changes in the actual application requirements and the continuous advancement of technology, NIRS has gradually become a monitoring technique in areas such as food, medicine, environment[14,15], materials[16,17], and crop growth[18,19]. The near-infrared devices have become smaller and real-time applications and high model accuracy are required for many monitoring and analysis methods[20]. Most of the related research has occurred in the food and medicine fields. In food research, Lopes et al.[21] applied NIRS to monitor the peroxidase bio-catalytic reaction in horseradish. Ringsted et al.[22] used NIRS to monitor the aging process of wheat bread. Genisheva et al.[23] monitored volatile compounds in wine. Other researchers used NIRS to monitor the fermentation processes of solid ethanol[24],

Received date: 2018-10-11 **Accepted date:** 2019-03-13

Biographies: Xue Wang, PhD, Lecturer, research interests: information processing in agriculture, Email: mtmwx@163.com; **Tiemin Ma,** PhD, Lecturer, research interests: network recommendation and big data analysis, Email: mtm_120@sina.com; **Ping Song,** PhD, Associate Professor, research interests: information processing in agriculture, Email: songping_1010@163.com; **Zhengguang Chen,** PhD, Professor, research interests: spectroscopy data processing, Email: ruzee@sina.com; **Huan Xie,** MS, research interests: spectroscopy data processing of spectroscopy. Email: byxh1010@sina.com.

***Corresponding author: Tao Yang,** Professor, research interests: computer technology in agriculture, College of Information and Electrical Engineering, Shenyang Agricultural University, No. 120, Dongling Road, Shenhe District, Shenyang 110866, China. Tel: +86-18804050051 , Email: 328748306@qq.com.

cider[25], and rice wine[26]. In medical research, NIRS was used to monitor drug production[27,28] and extraction processes[29]. Researchers have also used NIRS sensors for the continuous monitoring of blood glucose levels[30] and the real-time monitoring of cell biology[31]. However, few studies have been conducted on the use of NIRS for crop kernel growth monitoring.

The sample size requirements for the collection and analysis of near-infrared spectral data during crop growth are several times higher than those for conventional quantitative and qualitative analysis. The sample size is determined by the number of near-infrared monitoring times during crop growth. In general, the sample size required for the quantitative or qualitative analysis of NIRS data is generally from 100 to 200[32] and a spectrum sample consists of 3-5 grains for destructive sampling and 50 grains for non-destructive sampling. If 7 times the number of grains is required to determine the moisture in the grain-filling stage, one thousand grains are required for destructive sampling and ten thousand grains for non-destructive sampling using NIRS. The number of grains would be higher when a dry measurement method is used. Therefore, the development of a model for crop growth data monitoring using NIRS and small sample sizes is needed to meet the real-world application requirements.

In recent years, many studies have been conducted on the analysis of small sample size dataset[33,34]. Commonly, the sample size is increased using certain methods to improve the results. The bootstrap algorithm[35] is a common resampling method for small sample set in chemometrics and its reliability has been demonstrated by many researchers[36-38]. We have previously investigated the feasibility of using the bootstrap method for the quantitative analysis of the maize moisture content at the grain-filling stage[39] and a similar study was conducted by other researchers[35]. In the NIRS data, the distribution does not have to be considered and pre-processing is not required for small sample set when using the bootstrap algorithm and 10 is the minimum critical stability number of samples for small sample set[39].

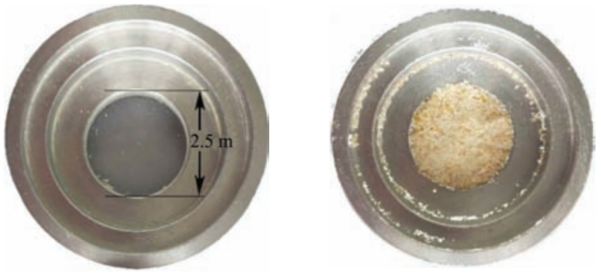
In this study, we developed a monitoring model for predicting the change in the maize moisture content at the grain-filling stage using full-spectrum NIRS and small sample sizes of 10, 20, and 50; we use the bootstrap and x-y distance sample partitioning (SPXY) and the partial least squares (PLS) regression. The model has two key steps, i.e., the selection of the optimal number of factors and the sub-model screening. The proposed model improves the efficiency and accuracy of the moisture content determination during crop growth, reduces the cost of the measurements, and provides a new method for moisture content determination during crop growth. The results do not only contribute to crop growth monitoring and seed breeding research but also provide a new method for the quantitative analysis of data with small sample sizes and different concentrations using near-infrared spectroscopy of full spectrum range.

2 Materials and methods

2.1 Materials

The samples were collected from the maize test base at the Heilongjiang Bayi Agricultural University. The base is located in the continental monsoon climate zone of the Daqing North temperate zone in China, where summers are short. The temperature was in the range of 24°C-32°C during testing. It began to rain at the end of September. The experimental planting area was about 800 m^2, the planting density was 12 plants/m^2, and the variety was “Xianyu 335”. The maize ear samples were collected every 7 d during the grain-filling stage. Each time, 10 maize ears were sampled and were quickly moved to the laboratory and stored at a low temperature to minimize the water loss after harvesting in high temperatures. Chemical testing and spectra acquisition were completed in the shortest time possible to minimize any external influences on the predictive model. The middle 200 grains of the maize ears were obtained; some grains were dried to measure the moisture content and some were used to collect the spectral data; the rest were dried naturally.

The spectrometer was a WQF-600N Fourier transform near-infrared spectrometer (FTNIR) (Beijing Rayleigh) with a wavelength range of 4000 cm^{-1} to 10 000 cm^{-1}. Each sample was scanned 32 times and the average values were obtained. For obtaining the spectral data, the grains were ground with a mill to use as few grains as possible. The sample pool was filled with the ground grain, as shown in Figure 1. More than 100 spectral signatures curves were obtained. After the abnormal samples were eliminated, 50 samples were used for the modeling set and the remaining 50 samples were used for the predicting set.



a. Sample pool before loading b. Sample pool after loading

Figure 1 Sample pool for the collection of the spectral data

A secondary drying method was used to obtain the chemical reference values. The initial drying temperature was 105°C for 2 h; then the temperature was kept constant at 85°C until the quality of the 100 grains did not change. The entire drying process typically required about 12 h. The calculation of the moisture content is shown in Equation (1):

$$WC(\%)=[(FW-DW)/\mathrm{FW}]\times 100\% \qquad (1)$$

where, WC is the percentage of the moisture content of the 100 grains, %; FW is the fresh weight of the 100 grains, g; and DW is the dry weight of the 100 grains, g.

2.2 Methods

2.2.1 Bootstrap

The bootstrap algorithm is based on the strategy of resampling. The simulated NIRS dataset possessed all the characteristic of the raw NIRS dataset and resulted in a better sample distribution. In this algorithm, the iterated dataset is added to the original dataset and the iterated weights are recalculated. The flow of the algorithm is shown in Figure 2.

2.2.2 SPXY

In the SPXY, the distance between the sample of interest and all other samples in the spectrum vector and concentration vector is calculated. The SPXY algorithm is suitable for divided samples in the NIRS quantitative analysis and is a classic sample optimization algorithm. In recent years, researchers have used it to optimize the selection of NIRS quantitative samples and have achieved good model results[40].

2.2.3 PLS and the optimal factor number

When establishing a quantitative regression model based on PLS, the number of factors directly affects the model outcome. If the number of factors is too small, some information in the spectrum may be lost and under-fitting may occur; if the number of factors is too large, noise is introduced into the model, resulting in over-fitting of the model. Both under- and over-fitting cause large prediction errors[41]. The number of PLS model factors is one of the key elements in developing a robust model for different spectral samples, different crop growth stages, and different moisture contents.

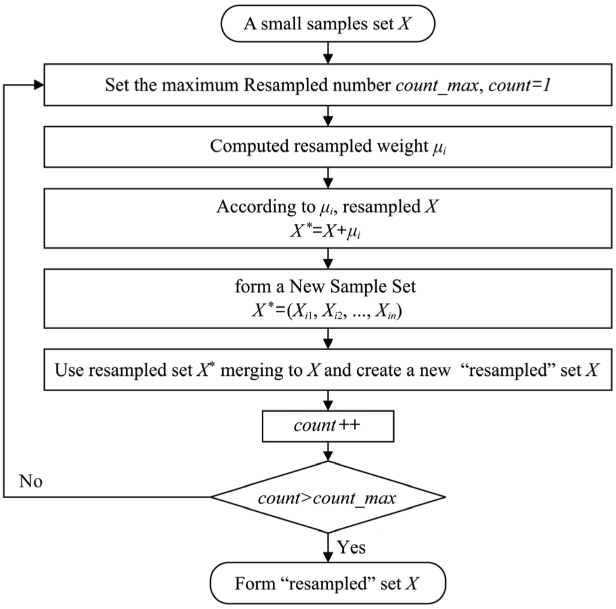

A small samples set X
Set the maximum Resampled number $count\_max$, $count=1$
Computed resampled weight μ_i
According to μ_i, resampled X
$X^*=X+\mu_i$
form a New Sample Set
$X^*=(X_{i1}, X_{i2}, ..., X_{in})$
Use resampled set X^* merging to X and create a new "resampled" set X
$count$++
No
$count>count\_max$
Yes
Form "resampled" set X

Figure 2 Flowchart of the bootstrap algorithm

In order to ensure the accuracy and stability of the mathematical model and to prevent under-fitting or over-fitting, two interaction verifications are used to determine the optimal number of factors. The first is an analysis of the interaction between the root mean square error computed from the cross-validation (RMSECV) and the number of factors in the case of different moisture contents and different number of spectral samples. The second is an analysis of the interaction between the R^2 and the number of factors in the case of different moisture contents and different number of spectral samples.

2.2.4 Sub-model screening

The performance of the sub-model is usually evaluated by the RMSECV, the root mean square error of prediction (RMSEP), and the R^2 values. The larger the R^2 value, the stronger the prediction ability of the model is; the values of the RMSECV and RMSEP are relatively small and are consistent. If the RMSECV is much larger than the RMSEP value, it means that the representative sample is poor; if the RMSECV is far less than RMSEP, it indicates poor representativeness of the modeled sample and the information cannot be fitted adequately or is over-fitted.

In order to enhance the prediction accuracy of the model, in addition to using the parameters of R^2, RMSECV, and RMSEP to screen the model, we integrated the three model evaluation parameters to create the screening threshold *RVP* of the model, where *R* represents the R^2, *V* represents the RMSECV, and *P* stands for RMSEP. The formula for calculating the threshold *RVP* is shown in Equation (2). RVP_i is the threshold of the i^{th} sub-model.

$$RVP_i = \frac{R^2}{\frac{1}{n}\sum_{i=1}^{n} R_i^2} + \frac{\min(RMSECV_i, RMSEP_i)}{\max(RMSECV_i, RMSEP_i)} \quad i\in(1,2,...,n) \qquad (2)$$

A threshold *RVP* of less than 1.99 indicates that the RMSECV has a large deviation from the RMSEP value or the R^2 value is less than the average value of the sub-model.

In the sub-model screening, we delete the sub-m*i*, which has an $RVP_i < 1.99$, the minimum R^2 value, and the maximum RSECV value and RMSEP.

2.2.5 Bootstrap-SPXY-PLS mositure monitoring model based on a small sample set

The monitoring model for determining the change in the moisture content is based on the bootstrap method, SPXY, and PLS regression using small sample sets, as shown in Figure 3. In the model, the combination of resampling and sample division is used to create the data set that meets the needs of the sample analysis and modeling. When the number of samples is between 10 and 50, the prediction accuracy and robustness of the model can be guaranteed and the accuracy of the model can be improved. The bootstrap resampling algorithm repeatedly simulates a small sample data set and uses sample merging to ensure the differences between the new data set samples, thereby creating a resampled set. The SPXY algorithm performs optimization screening of the sample set using a re-extraction strategy to form a model set of multiple subsets.

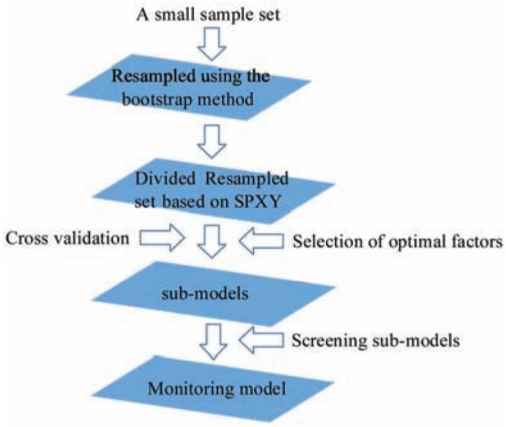

A small sample set
Resampled using the bootstrap method
Divided Resampled set based on SPXY
Cross validation
Selection of optimal factors
sub-models
Screening sub-models
Monitoring model

Figure 3 Bootstrap-SPXY-PLS moisture content monitoring model based on a small sample set

Bootstrap resampling and sample optimization of the SPXY are performed to form a set of modeling samples composed of multiple subsets. Then, one subset is randomly selected for pre-modeling based on the PLS and the RMSECV and R^2 values of each factor are recorded. By determining the mean square error of the interaction and the R^2 of the model for the different factors, the optimal number of factors for establishing the model is obtained. At this stage, the RMSECV is small and the R^2 value is large and the two are balanced. The optimal number of factors is also the optimal number of factors for the regression prediction. Based on the number of samples and the optimal modeling factor of the moisture content monitoring model in different grain-filling stages, a PLS sub-model based on the best factor number is created for the modeling subset and the sub-m*i* is determined, where i=1, 2 , 3,...,n. The means of the R^2, RMSECV, and RMSEP of the sub-m*i* are calculated and the *RVP* threshold of the sub-m*i* is calculated using Equation (2). The sub-models are screened. The sub-model among the subset of moisture content monitoring models with the minimum R^2 value, or the maximum RMSECV and RMSEP values, or the threshold *RVP* value of less than 1.99 is deleted. Finally, the selected sub-m*i* is used to perform the regression prediction using the optimal number of factors of the model. The prediction result is used as a subset of the prediction results (predicted *i*) and the prediction results of the moisture content of the different samples in the different grain-filling stages are obtained. The mean value of the set of predicted values is used as the predicted moisture content during the corresponding grain-filling stage.

3 Results and discussion

3.1 Near-infrared spectroscopy data and chemical reference value

Figure 4 shows the Fourier transform near-infrared spectra at the maize grain-filling stage after pre-processing with a Savitzky-Golay filter with a window size of 13. The seven spectral curves represent the average values of the different stages. And The water absorption spectrum ranges from 6900 cm^{-1} to 7900 cm^{-1}. The blue curve at the bottom represents data sampled on August 21st and the pink curve at the top represents data sampled on October 2nd. The spectral data represent the change in moisture of the maize kernels in the grain-filling stage.

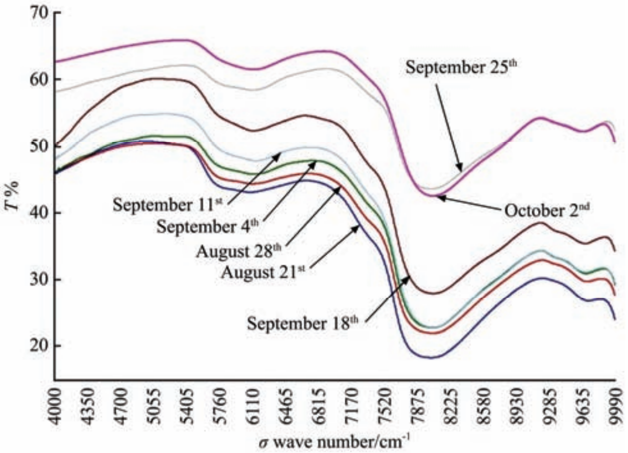


Figure 4 Fourier transform near-infrared average spectra at the maize grain-filling stage

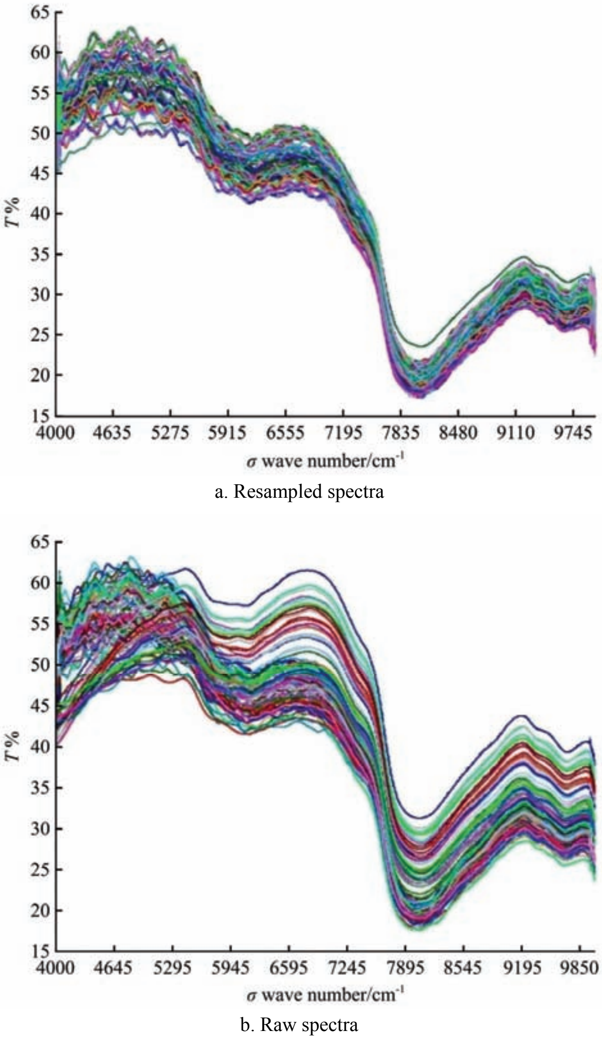
The chemical reference values of the seven batches are shown in Table 1. It is observed that the moisture content of the maize kernels decreases rapidly from August 21, 2016 to September 4; after this date, the precipitation decreases. Because of the moderately rainy weather on the day before sampling on October 2, the moisture content of the maize kernels is higher than in the samples collected on September 25.

Table 1 Chemical reference values

Sample Date	Value of chemical reference/%, w/w			
	Max	Min	Average	SD
2016/8/21	85.17	73.91	77.95	3.90
2016/8/28	74.59	68.04	71.16	1.99
2016/9/4	57.97	53.59	56.07	1.69
2016/9/11	51.25	50.34	50.81	0.35
2016/9/18	47.84	42.38	45.67	2.37
2016/9/25	28.62	26.90	27.47	0.68
2016/10/2	39.13	35.90	37.74	1.33

3.2 Analysis of Bootstrap-SPXY-PLS Moisture content Monitoring Model

The data that were resampled 500 times were considered the optimal dataset[39], i.e., *count_max*=500. The sample size was 10, 20, and 50, referred to as *X_ten*, *X_twenty,* and *X_fifty*, respectively. After resampling, the datasets are referred to as *X*_ten*, *X*_twenty* and, *X*_fifty*, respectively. The data were processed using the SPXY method to create the modeling sets. Then, sub-models were created by cross-validation. Figure 5b shows the spectra of one sub-model of the data obtained on September 11. The characteristics of the two spectra were identical but the resampled spectra represent the optimal dataset with regard to the *T* values. However, the noise was retained.



a. Resampled spectra

b. Raw spectra

Figure 5 Comparison of raw spectra and resampled spectra obtained on September 11

The results of the screened and unscreened moisture content monitoring models for the different sample sizes at the grain-filling stage are shown in Table 2. The R^2, RMSECV, and RMSEP-mean values indicate better performance of the screened model and the values of r_p are larger. However, when the number of samples is 10, the r_p value is lower. For the sample sizes of 20 and 50, the R^2 values of the models for all grain-filling stages are larger than 0.98 and the average relative increase is 0.27% and 0.24% (September 11) and 0.13% and 0.11% (September 25). The improvements are more apparent for a sample size of 10. The R^2 value of the optimized model of each grain-filling stage is 0.5% higher than that of the unscreened model. The lowest R^2 value of the model is 0.9397 or higher and the relative deviation of the mean value of RMSECV and RMSEP has decreased by 0.04%.

These results demonstrate that the model has good predictive ability and the predictive ability improves after screening. It can be seen from the r_p value that the number of samples does affect the stability of the model prediction.

Table 2 Prediction results of the Bootstrap-SPXY-PLS moisture content monitoring model

Sample Date	Sample size	Unscreened model			Screened model				
		R^2	RMSECV/%	RMSEP-mean/%	R^2	RMSEC/%	RMSECV/%	RMSEP-mean/%	r_p-mean
21/8/2016	10	0.9974	3.158	3.16	0.9983	0.116	3.134	3.134	0.9211
	20	0.9979	3.115	3.117	0.9981	0.125	3.077	3.078	0.9311
	50	0.9896	3.061	3.062	0.9932	0.232	3.046	3.047	0.9967
28/8/2016	10	0.9953	1.415	1.415	0.9959	0.084	1.393	1.394	0.9367
	20	0.9611	1.451	1.46	0.9846	0.247	1.438	1.438	0.9495
	50	0.9937	1.428	1.427	0.9951	0.092	1.408	1.408	0.9558
4/9/2016	10	0.9961	1.302	1.302	0.9967	0.069	1.294	1.294	0.9731
	20	0.9970	1.274	1.275	0.9970	0.065	1.268	1.269	0.9714
	50	0.9959	1.312	1.315	0.9968	0.067	1.292	1.296	0.9725
11/9/2016	10	0.9928	0.381	0.381	0.9953	0.003	0.378	0.379	0.9751
	20	0.9883	0.393	0.393	0.9904	0.003	0.386	0.386	0.9745
	50	0.9936	0.400	0.403	0.9957	0.027	0.390	0.399	0.9685
18/9/2016	10	0.9293	4.659	5.018	0.9397	0.431	4.738	4.738	0.9268
	20	0.9860	1.954	1.956	0.9939	0.130	1.929	1.935	0.9096
	50	0.9950	1.898	1.899	0.9961	0.118	1.889	1.89	0.9413
25/9/2016	10	0.9600	0.525	0.531	0.9628	0.096	0.525	0.525	0.9186
	20	0.9800	0.571	0.548	0.9833	0.064	0.564	0.566	0.9143
	50	0.9810	0.586	0.571	0.9873	0.059	0.572	0.573	0.9131
10/2/2016	10	0.9102	1.485	2.026	0.9479	0.241	1.415	1.417	0.9977
	20	0.9927	1.108	1.108	0.9948	0.067	1.077	1.076	0.9987
	50	0.9887	1.151	1.151	0.9961	0.060	1.128	1.127	0.9555

3.3 Selection of the optimum number of factors for different sample sizes

The number of factors is one of the key parameters to ensure the robustness of the Bootstrap-SPXY-PLS model because using the optimum number of factors improves the predictive ability of the model. After optimizing the sample set formation, we first select the subsets of the different sample sizes and the different grain-filling stages to create the corresponding pre-model as the best analysis model. By plotting the RMSECV and R^2 trend graphs of the different models, the number of different samples and the optimal number of factors for the different grain-filling stage are obtained.

Figure 6 shows the RMSECV and R^2 values of the model for the different grain-filling stages, different sample sizes, and different factor numbers. Figure 6a shows the trend of the RMSECV and R^2 values based on the Bootstrap-SPXY-PLS pre-model sampled on September 11 for the sample sizes of 10, 20, and 50.

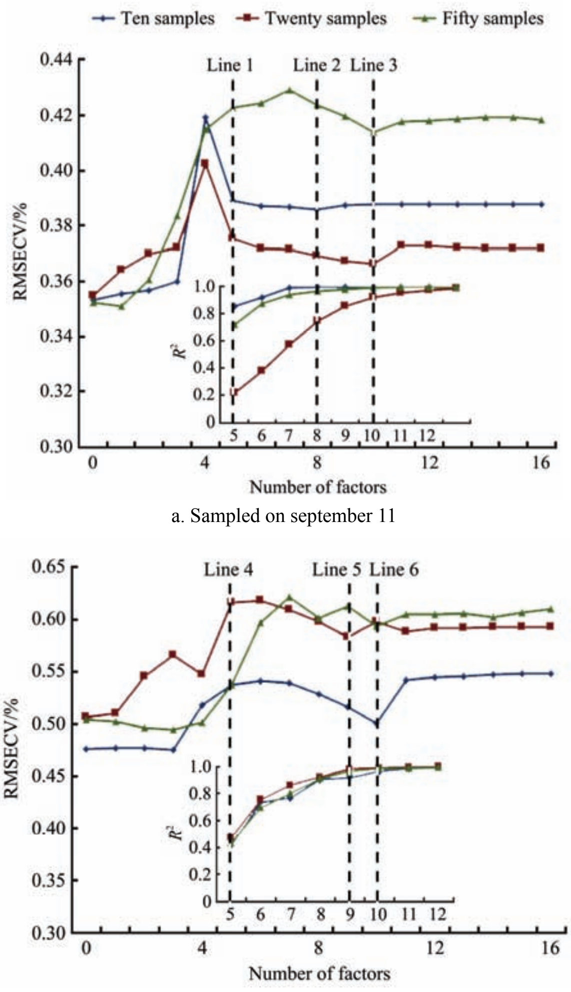
The overall trend of the RMSECV value can be divided into three parts. The first part ranges from factor = 0 to the factor of the maximum RMSECV value, which is 5 for the sample sizes of 10 and 20 and 7 for the sample size of 50. However, the R^2 value is lower when the factor number is less than 5 or 7. When the sample size is 20, the R^2 value is only 0.2 and the maximum is only about 0.8 when the sample size is 10 (line 1 in Figure 6a). The second part of the RMSECV value gradually decreases and the R^2 value gradually increases. When the sample size is 10, the RMSECV falls to the minimum value when the factor is equal to 8 and the R^2 value is 0.96 (line 2 in Figure 6a). When the sample size is 20 and 50, the RMSECV values of the factor of 10 are the minimum values in this part and the corresponding R^2 values are 0.91 and 0.99, respectively (line 3 Figure 6a). In the third part, the RMSECV value rises again, resulting in a poor prediction performance of the model, although the corresponding R^2 value increases. Therefore, based on the RMSECV and R^2 results, the factor at which the lowest RMSECV value was obtained in the second part is the optimal factor for the model.

Figure 6b shows the RMSECV and R^2 values at a moisture content of 27% sampled on September 25. Due to the decrease in the moisture content, the trend of the RMSECV value is different from that shown in Figure 6a. As the number of factors increases, the RMSECV curve increases stepwise. It is also possible to divide the curve into three parts. The first part consists of the factor numbers less than or equal to 5 at sample sizes of 10, 20, and 50, respectively. The second part consists of the factor numbers less than or equal to 10. The third part consists of the factor number larger than 11. By comparing the R^2 trend, a similar conclusion can be drawn that the minimum value of the second part of the RMSECV curve corresponds to the optimum factor number. The factor numbers of the sample sizes of 10, 20, and 50 are 10, 9, and 10, respectively and the corresponding R^2 values are 0.963, 0.983, and 0.987, respectively.

It can be concluded from the data shown in Figure 6 that for different moisture contents, the RMSECV value of the model stabilizes after a certain change for the small sample sizes of 10, 20, and 50 samples.

If a larger factor number is not suitable for a model, the R^2 results of the pre-modeling and the range of the best factor number based on the RMSECV value can be used to obtain the optimum number of factors for the different grain-filling stages and different sample sizes.

The R^2 and RMSECV values shown in Figure 6 indicate that the different moisture contents and different sample sizes have an influence on the optimum number of factors in the model. Therefore, the predictive ability of the model was improved by determining the optimal number of factors prior to creating the sub-model as shown in Table 3.



a. Sampled on september 11

b. Sampled on september 25

Figure 6 RMSECV and R^2 values versus factor number of the pre-model at the grain-filling stage for different sample sizes

Table 3 Optimum number of factors for moisture content monitoring model at grain-filling stage for different sample sizes

Sample size	Sample date (2016)						
	8/21	8/28	9/4	9/11	9/18	9/25	10/2
10	5	10	5	8	10	10	10
20	5	10	5	10	5	9	10
50	5	5	5	10	5	10	10

3.4 Screening of the moisture content monitoring sub-models at the grain-filling stage for different sample sizes

Tables 4 and 5 respectively show the screening results of the sub-models for the September 11 and September 25 sampling dates for different sample sizes.

As can be seen in Table 4, the minimum R^2 values of the sub-models are 0.9787 for sub-m8 for a sample size of 10, 0.9943 for sub-m9 for a sample size of 20 samples, and 0.9886 for sub-m5 for a sample size of 50. The maximum RMSECV values of the sub-models are 0.4142% for sub-m3 (10 samples), 0.4283% for sub-m9 (20 samples), and 0.4232% for sub-m6 (50 samples). The sub-models with a threshold RVP of less than 1.99 are sub-m3 and sub-m8 (10 samples), sub-m1 and sub-m3 (20 samples), and sub-m7 (50 samples). Therefore, for the sub-model sampled on September 11, we removed sub-m3 and sub-m8 (10 samples); sub-m1, sub-m3, sub-m6 and sub-m9 (20 samples); and sub-m5, sub-m7, and sub-m6 (50 samples).

Similar results can be seen in Table 5; for the sub-model sampled on September 25, we removed sub-m2 and sub-m10 (10 samples), sub-m3 and sub-m6 (20 samples), and sub-m6 and sub-m7 (50 samples).

All sub-models were evaluated and screened using the R^2, RMSECV, RMSEP-mean, and the threshold *RVP*. The sub-models with a minimum R^2, the maximum RMSECV and RMSEP-mean, and a threshold *RVP* of less than 1.99 were removed for the seven grain-filling stages. The sub-model screening results are shown in Table 6.

3.5 Comparison of the prediction values and reference values

The prediction results and the moisture content changes during the grain-filling stage are shown in Figure 7. At a sample size of 50, the average deviation from the reference value of the 7 predicted moisture contents at the grain-filling stage is 0.0918% and the maximum deviation is 0.3095% on September 18. At a sample size of 20, the average deviation is 0.057% and the maximum deviation of 0.1805% is predicted for October 2. At a sample size of 10, the average deviation is 0.1078% and the maximum deviation occurs on September 18th with a moisture content of 0.3036%. It is evident that the Bootstrap-SPXY-PLS optimization model for analyzing the NIRS data results in a small deviation between the prediction and the reference value of the moisture content during the grain-filling stage. The reference and prediction curves nearly coincide. However, for the drying method, we only tested the samples in the laboratory and did not predict the parameters of other samples. The NIRS method provides rapid prediction and non-destructive testing can be conducted, which is beneficial for maize breeding research.

Table 4 Evaluation of the moisture content monitoring sub-models for different sample sizes sampled on September 11

Sub-model	10 samples				20 samples				50 samples			
	R^2/%	RMSECV /%	RMSEP-mean/%	RVP	R^2/%	RMSECV /%	RMSEP-mean/%	RVP	R^2/%	RMSECV /%	RMSEP-mean/%	RVP
Sub-m1	0.9943	0.3628	0.36	1.994	0.9968	0.3539	0.35	1.989	0.9965	0.386	0.39	1.992
Sub-m2	0.9986	0.3867	0.39	1.997	0.9993	0.3727	0.37	1.996	0.9975	0.4183	0.42	1.999
Sub-m3	0.9865	0.4142	0.41	1.983	0.9985	0.365	0.36	1.988	0.9960	0.3838	0.38	1.992
Sub-m4	0.9988	0.3626	0.36	1.999-m38	0.9957	0.4007	0.4	1.997	0.9893	0.4185	0.42	1.992
Sub-m5	0.9896	0.3803	0.38	1.996	0.9952	0.4093	0.41	1.997	0.9886	0.4206	0.42	1.993
Sub-m6	0.9988	0.3769	0.38	1.998	0.9955	0.4078	0.41	1.994	0.9964	0.4232	0.42	1.995
Sub-m7	0.9942	0.3858	0.39	1.991	0.9949	0.418	0.42	1.994	0.9923	0.3742	0.37	1.987
Sub-m8	0.9787	0.3705	0.37	1.984	0.9974	0.3787	0.38	1.998	0.9967	0.4012	0.4	1.999
Sub-m9	0.9904	0.3894	0.39	1.996	0.9943	0.4283	0.43	1.994	0.9947	0.397	0.4	1.993
Sub-m10	0.9987	0.3802	0.38	2.005	0.9948	0.3986	0.4	1.995	0.9938	0.4118	0.41	1.995

Table 5 Evaluation of the moisture content monitoring sub-models for different sample sizes sampled on September 25

Sub-model	10 samples				20 samples				50 samples			
	R^2/%	RMSECV /%	RMSEP-mean/%	RVP	R^2/%	RMSECV /%	RMSEP-mean/%	RVP	R^2/%	RMSECV /%	RMSEP-mean/%	RVP
Sub-m1	0.9940	0.5444	0.54	1.993	0.9954	0.5576	0.56	1.997	0.9927	0.5887	0.59	1.999
Sub-m2	0.9884	0.5318	0.53	1.992	0.9946	0.5315	0.53	1.998	0.9923	0.5905	0.59	2.000
Sub-m3	0.9943	0.5438	0.54	1.995	0.9953	0.5775	0.58	1.997	0.9891	0.5614	0.56	1.995
Sub-m4	0.9885	0.5382	0.54	1.992	0.9938	0.5675	0.57	1.996	0.9888	0.582	0.58	1.994
Sub-m5	0.9960	0.5173	0.52	1.998	0.9944	0.5439	0.54	1.993	0.9949	0.5629	0.56	1.998
Sub-m6	0.9917	0.4964	0.5	1.991	0.9826	0.5231	0.52	1.983	0.9925	0.6025	0.6	1.997
Sub-m7	0.9951	0.4897	0.49	2.001	0.9925	0.5604	0.56	1.998	0.9781	0.5675	0.57	1.982
Sub-m8	0.9948	0.5271	0.53	1.996	0.9945	0.5294	0.53	1.999	0.9956	0.5065	0.51	1.997
Sub-m9	0.9932	0.549	0.55	1.998	0.9947	0.5494	0.55	2.000	0.9934	0.5566	0.56	1.996
Sub-m10	0.9908	0.5671	0.57	1.993	0.9964	0.542	0.54	1.999	0.9919	0.586	0.59	1.994

Table 6 Sub-model screening results

Sample date	8/21/2016	8/28/2016	9/4/2016	9/11/2016	9/18/2016	9/25/2016	10/2/2016
10 samples	1, 3-9	1, 3-7, 9, 10	2, 3, 5-10	1, 2, 4-7, 9, 10	1, 3, 6, 8, 10	1, 3-9	1-9
20 samples	1-6, 8, 10	1-3, 6-10	1, 2, 5-10	2,4,5, 7-8,10	1, 3-7, 9, 10	1-2,4,5, 7-10	1, 5-10
50 samples	1-5, 7-10	1, 2, 4-8, 10	1-4, 6-9	1-4, 8-10	1-5, 8-10	1-5, 8-10	1, 2, 4, 5, 7, 8, 10

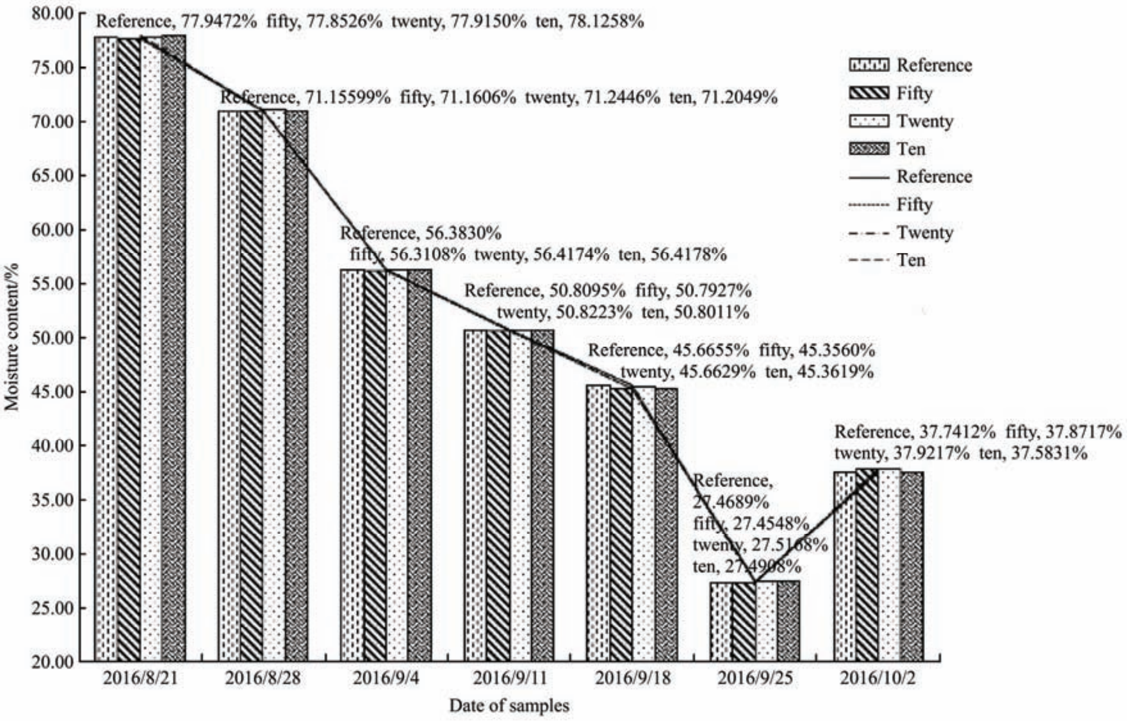


Figure 7 Comparison of the reference values and prediction values for different sample sizes

4 Conclusions

(1) We conducted a quantitative analysis of the moisture content of maize kernels at multiple stages during the grain-filling stage and created a PLS regression model based on the Bootstrap and SPXY optimization method for the sample sizes of 10, 20, and 50. The results indicated that the screened model had better performance. The screened model was developed for predicting the maize moisture content during the grain-filling stage based on the optimal number of factors. The results demonstrated that the predicted and reference values of the maize moisture content during the grain-filling stage were similar for the sample sizes of 10, 20, and 50. The average deviations between the predicted values and reference values were 0.1078%, 0.057%, and 0.0918%, respectively.

(2) By determining the R^2 and RMSECV values of the PLS sub-models, the optimal number of factors of the model for different grain-filling stages and different sample sizes were obtained. The results proved that the model evaluation index provided the optimal number of factors and, therefore, better prediction results.

(3) The comprehensive evaluation parameter *RVP* for the sub-model screening was proposed. The *RVP*, R^2, RMSECV, and RMSEP-mean values were used as comprehensive screening parameters for the sub-model. The experimental results showed that the R^2 values after screening showed an average relative increase of 0.5%, 0.27%, and 0.24% for the sample sizes of 10, 20, and 50.

The results of the regression prediction using the Bootstrap-SPXY-PLS optimization model on small sample set indicated that the moisture content can be predicted and monitored at the maize grain-filling stage. The model is based on destructive sampling but uses only one-tenth of the number of maize grains used for the model based on non-destructive sampling. Using a low number of maize grains at the filling stage is very important for crop breeding because the maize grains have not reached maturity. This is an advantage of the model used in this study based on destructive sampling. There is a certain amount of water loss during grinding. Therefore, in the future, we will conduct in-depth research on data processing algorithms using a small sample size. We will also investigate the use of multi-spectral data and methods suitable for collecting spectral information on maize ears directly in the field. The results of this study provide a new method for moisture content monitoring during crop growth stages using NIRS.

Acknowledgments

This work was financially supported by the grant from the International Cooperation and Exchange of the National Natural Science Foundation of China (No. 31811540396), China, the National Natural Science Foundation of China (No. 31701318), China, and the Training Project of Heilongjiang Bayi Agricultural University, China (No. XZR2016-09).

[References]

[1] Cai Z W, Wang K R, Guo Y Q, Xie R Z, Li L L, Ming B. Current status of maize mechanical grain harvesting and its relationship with grain moisture content. China Journal of Scientia Agricultura Sinica, 2017; 50(11): 2036–2043. (in Chinese)

[2] Li L L, Xie R Z, Lei X P, Wang K R, Hou P, Zhang F L. Analysis of Influential Factors on Mechanical Grain Harvest Quality of Summer Maize. China Journal of Scientia Agricultura Sinica, 2017; 50(11): 2044–2051. (in Chinese)

[3] Liu F H , Wang K R, Li J, Wang X M, Sun Y L, Chen Y S. Factors affecting corn mechanically harvesting grain quality. China Journal of Crops, 2013; 4: 116–119. (in Chinese)

[4] Ghorchiani M, Etesami H, Alikhani H A. Improvement of growth and yield of maize under water stress by co-inoculating an arbuscular mycorrhizal fungus and a plant growth promoting rhizobacterium together with phosphate fertilizers. Agriculture, Ecosystems & Environment, 2018; 25(8): 59–70.

[5] Wang W Y. Research on Problems and Countermeasures of Maize Breeding in the New Period. China Journal of Seed Science & Technology 2019; 2: 30–31. (in Chinese)

[6] Liu S Q, Zhong X M, Li F H, Zhu M, Wang H W, Lu X L. Comparisons of grain filling and dehydration rates in 4representative maize varieties in northeast provinces. China Journal of Seed, 2015; 34(12): 69–72. (in Chinese)

[7] Ercioglu E, Velioglu H M, Boyaci I H. Determination of terpenoid contents of aromatic plants using NIRS. Talanta, 2018; 178: 716–721.

[8] Jiang H, Lu J. Using an optimal CC-PLSR-RBFNN model and NIR spectroscopy for the starch content determination in corn. Spectrochim Acta A Mol Biomol Spectrosc, 2018; 196: 131–140.

[9] Perissinato A G, Garcia J S, Trevisan M G. Determination of β-galactosidase in tablets by infrared spectroscopy. Chemical Papers, 2016; 71(1): 171–176.

[10] Rodionova O Y, Balyklova K S, Titova A V, Pomerantsev A L. Application of NIR spectroscopy and chemometrics for revealing of the 'high quality fakes' among the medicines. Forensic Chemistry, 2018; 8: 82–89.

[11] Stefano M, Giuseppe P, Luigi R, Roberta R. High-throughput prediction of AKB48 in emerging illicit products by NIR spectroscopy and chemometrics. Microchemical Journal, 2017; 134: 277–283 .

[12] Revilla I, Vivar-Quintana A M, González-Martín I, Escuredo O, Seijo C. The potential of near infrared spectroscopy for determining the phenolic, antioxidant, color and bactericide characteristics of raw propolis. Microchemical Journal, 2017; 134: 211–217.

[13] Cui Y, Xu L, An D, Liu Z, Gu J, Li S. Identification of maize seed varieties based on near infrared reflectance spectroscopy and chemometrics. International Journal of Agricultural and Biological Engineering, 2018; 11(2): 177–183.

[14] Stockl A, Lichti F. Near-infrared spectroscopy (NIRS) for a real time monitoring of the biogas process. Bioresour Technol, 2018; 247: 1249–1252.

[15] Druckenmuller K, Gunther K, Elbers G. Near-infrared spectroscopy (NIRS) as a tool to monitor exhaust air from poultry operations. Sci Total Environ, 2018; 630: 536–543.

[16] Shan C, Wang B, Hu B, Liu W, Tang Y. Smart yolk-shell type luminescent nanocomposites based on rare-earth complex for NIR–NIR monitor of drug release in chemotherapy. Journal of Photochemistry and Photobiology A: Chemistry, 2018; 355: 233–241.

[17] Shinzawa H, Mizukado J. Near-infrared (NIR) monitoring of Nylon 6 during quenching studied by projection two-dimensional (2D) correlation spectroscopy. Journal of Molecular Structure, 2016; 1124: 188–191.

[18] Salgó A, Gergely S. Analysis of wheat grain development using NIR spectroscopy. Journal of Cereal Science, 2012; 56(1): 31–38.

[19] Qian W, Hong S, Minzan L, Wei Y. Development and application of crop monitoring system for detecting chlorophyll content of tomato seedlings. Int J Agric & Biol Eng, 2014;7(2): 138–145.

[20] Dai X, Hang S, Wen L, Yao S, Gang W. On-line UV-NIR spectroscopy as a process analytical technology (PAT) tool for on-line and real-time monitoring of the extraction process of Coptis Rhizome. Rsc Advances, 2016; 6(12): 10078–10085.

[21] Lopes L C, Brandão I V, Sánchez O C, Franceschi E, Borges G, Dariva C. Horseradish peroxidase biocatalytic reaction monitoring using Near-Infrared (NIR) Spectroscopy. Process Biochemistry, 2018. DOI: 10.1016/j.procbio.2018.05.024.

[22] Ringsted T, Siesler H W, Engelsen S B. Monitoring the staling of wheat bread using 2D MIR-NIR correlation spectroscopy. Journal of Cereal Science, 2017; 75: 92–99.

[23] Genisheva Z, Quintelas C, Mesquita D P, Ferreira E C, Oliveira J M, Amaral A L. New PLS analysis approach to wine volatile compounds characterization by near infrared spectroscopy (NIR). Food Chem., 2018; 246: 172–178.

[24] Jiang H, Mei C, Li K, Huang Y, Chen Q. Monitoring alcohol concentration and residual glucose in solid state fermentation of ethanol using FT-NIR spectroscopy and L1-PLS regression. Spectrochim Acta A Mol Biomol Spectrosc, 2018; 204: 73–80.

[25] Villar A, Vadillo J, Santos J I, Gorritxategi E, Mabe J, Arnaiz A, et al. Cider fermentation process monitoring by Vis-NIR sensor system and chemometrics. Food Chem., 2017; 221: 100–106.

[26] Kim D Y, Cho B K. Rapid monitoring of the fermentation process for Korean traditional rice wine 'Makgeolli' using FT-NIR spectroscopy. Infrared Physics & Technology, 2015; 73: 95–102.

[27] Li W, Han H, Cheng Z, Zhang Y, Liu S, Qu H. A feasibility research on the monitoring of traditional Chinese medicine production process using NIR-based multivariate process trajectories. Sensors and Actuators B: Chemical, 2016; 231: 313–323.

[28] Verstraeten M, Van Hauwermeiren D, Hellings M, Hermans E, Geens J, Vervaet C. Model-based NIR spectroscopy implementation for in-line assay monitoring during a pharmaceutical suspension manufacturing process. Int J Pharm, 2018; 546: 247–254.

[29] Nee K, Bryan S A, Levitskaia T G, Kuo J W J, Nilsson M. Combinations of NIR, Raman spectroscopy and physicochemical measurements for improved monitoring of solvent extraction processes using hierarchical multivariate analysis models. Analytica Chimica Acta, 2018; 1006: 10–21.

[30] Goodarzi M, Sharma S, Ramon H, Saeys W. Multivariate calibration of NIR spectroscopic sensors for continuous glucose monitoring. TrAC Trends in Analytical Chemistry, 2015; 67: 147–158.

[31] Li M, Ebel B, Chauchard F, Guédon E, Marc A. Parallel comparison of in situ Raman and NIR spectroscopies to simultaneously measure multiple variables toward real-time monitoring of CHO cell bioreactor cultures. Biochemical Engineering Journal, 2018,2018(137): 205–213.

[32] Zhu L W, Ma W G, Hu J, Zheng J H,Tian Y X, Guan Y J. Advances of NIR Spectroscopy Technology Applied in Seed Quality Detection. Spectroscopy and Spectral Analysis, 2015; 35(2): 346–349. (in Chinese)

[33] Wang M, Kong L, Li Z, Zhang L J. Covariance estimators for generalized estimating equations (GEE) in longitudinal analysis with small samples.

Statistics in medicine, 2016; 35(10): 1706–1721.

[34] Jennison, Christopher, Turnbull, W. Adaptive sample size modification in clinical trials: start small then ask for more? Statistics in medicine, 2015; 34(29): 3793–3810.

[35] Dwivedi A K, Mallawaarachchi I, Alvarado L A. Analysis of small sample size studies using nonparametric bootstrap test with pooled resampling method. Statistic Medicine, 2017; 36(14): 2187–2205.

[36] Krebsbach C M. Bootstrapping with small samples in structural equation modeling: goodness of fit and confidence intervals: University of Rhode Island, Rhode Island, 2014.

[37] Amalnerkar E, Lee T H, Lim W. Bootstrap Guided Information Criterion for Reliability Analysis Using Small Sample Size Information. World Congress of Structural and Multidisciplinary Optimisation, 2017; pp.326–333.

[38] Wang Y, Zhou W, Dong D, Wang Z. Estimation of random vibration signals with small samples using bootstrap maximum entropy method. Measurement, 2017; 105(7): 45–55.

[39] Wang X, Ma T, Yang T, Song P, Xie Q J, Chen Z G. Moisture quantitative analysis with small sample set of maize grain in filling stage based on near infrared spectroscopy. Transactions of the CSAE, 2018; 34(13): 203–310. (in Chinese)

[40] Li J B, Guo Z M, Huang W Q. Near-infrared spectra combining with CARS and SPA algorithms to screen the variables and samples for quantitatively determining the soluble solids content in strawberry. Spectroscopy & Spectral Analysis, 2015; 35(2): 372–378. (in Chinese)

[41] Xie Y, Li F, Fan X J, Hu S J, Xiao X, Wang J F. Components Analysis of Biochar Based on Near Infrared Spectroscopy Technology. Chinese Journal of Analytical Chemistry, 2018; 46(4): 609–615. (in Chinese)