Non-destructive method of small sample sets for the maize moisture content measurement during filling based on NIRS

Tiemin Ma[1], Guangyue Zhang[1], Xue Wang[1,2,3], Shujuan Yi[3,4*], Changyuan Wang[2*]

(1. *College of Information and Electrical Engineering, Heilongjiang Bayi Agricultural University, Daqing 163319, Heilongjiang, China*;
2. *Daqing Center of Inspection and Testing for Agricultural Products and Processed Products Ministry of Agriculture, Daqing 163319, Heilongjiang, China*;
3. *Heilongjiang Province Research Center of Ecological Rice Seedling Raising Device and Whole Course Mechanized Engineering Technology, Daqing 163319, Heilongjiang, China*;
4. *College of Engineering, Heilongjiang Bayi Agricultural University, Daqing 163319, Heilongjiang, China*)

Abstract: In maize breeding, limitations on sampling quantity and associated costs for measuring maize grain moisture during filling are imposed by factors like the planting area of new varieties, maize plant density, effective experimental spikes, and other conditions. However, the conventional method of detecting moisture content in maize grains is slow, damages seeds, and necessitates many sample sets, particularly for high moisture content determination. Thus, a strong demand exists for a non-destructive quantitative analysis model of maize moisture content using a small sample set during grain filling. The Bayes-Merged-Bootstrap (BMB) sample optimization method, which built upon the Bayes-Bootstrap sampling method and the concept of merging, was proposed. A critical concern in dealing with small samples is the relationship between data distribution, minimum sample value, and sample size, which has been thoroughly analyzed. Compared to the Bayes-Bootstrap sample selection method, the BMB method offers distinct advantages in the optimized selection of small samples for non-destructive detection. The quantitative analysis model for maize grain moisture content was established based on the support vector machine regression. Results demonstrate that when the optimal resampling size is 1000 times or more than the original sample size using the BMB method, the model exhibits strong predictive capabilities, with a determination coefficient (R^2)>0.989 and a relative prediction determination (RPD)>2.47. The results of the 3 varieties experiment demonstrate the generality of the model. Therefore, it can be applied effectively in practical maize breeding and determining grain moisture content during maize machine harvesting.

Keywords: near-infrared spectroscopy, moisture content quantitative analysis, small samples optimized, maize grain during the filling stage

DOI: 10.25165/j.ijabe.20241704.8738

Citation: Ma T M, Zhang G Y, Wang X, Yi S J, Wang C Y. Non-destructive method of small sample sets for the maize moisture content measurement during filling based on NIRS. Int J Agric & Biol Eng, 2024; 17(4): 236–244.

1 Introduction

The significance of smart agriculture lies in utilizing information technology to enhance the intelligence of every link in agricultural production. An essential aspect of smart agriculture involves establishing an agricultural big data processing system and implementing intelligent field crop monitoring. The key components of smart agriculture include non-destructive detection[1], rapid detection[2], and even in-situ detection[3]. Therefore, near-infrared spectral (NIRS) detection technology is crucial for realizing water monitoring in intelligent field crop monitoring and management[4,5].

With the advancement of NIRS technology, increasing attention has been directed towards the real-time reliability of detection technology[6]. The model's accuracy is directly linked to its real-time reliability. However, the model's accuracy relies more on the quality and quantity of the data, which conflicts with the sample size available for breeding purposes.

Small sample sizes can introduce a number of statistical problems and the model's accuracy cannot be guaranteed with Insufficient data[7]. The core objective of small-sample research is to enhance information processing by augmenting analytical training samples from the small-sample dataset. Cao et al.[8] proposed a joint probability distribution modeling method based on multidimensional Gaussian copula for the small sample case, with the best fitting ability for coal mine geotechnical strength parameters in the positive and negative correlation cases. Ma et al.[9] employed a transductive support vector machine (TSVM) for assessing forest fire susceptibility with small samples, a prediction accuracy of 0.9583 was achieved with a sample size of 28. Li et al.[10] proposed an integrated Transformer meta-learning (ETML) method for fault identification of bearings with few samples, with a test accuracy of 96.1%. Aydi et al.[11] proposed an approach based on the ordinary least squares and the multilayer perceptron (MLP) neural network, the proposed methods produced a good estimate

Received date: 2023-12-14 **Accepted date:** 2024-06-26

Biographies: Tiemin Ma, PhD, Lecturer, research interest: non-destructive quality detection, recommendation system, machine learning, Email: mtm_120@sina.com; **Guangyue Zhang**, MS candidate, research interest: near-infrared spectroscopy, model creating, Email: jiangdijidai@foxmail.com; **Xue Wang**, PhD, Associate Professor, research interest: non-destructive quality detection, near-infrared spectroscopy, data processing, Email: mtmwx@163.com.

***Corresponding author: Shujuan Yi**, PhD, Professor, research interest: agricultural mechanization engineering. College of Engineering, HeilongjiangBayi Agriculture University, Daqing 163319, Heilongjiang, China. Tel: +86-13836961877, Email: yishujuan_2005@126.com; **Changyuan Wang**, PhD, Professor, research interest: Grain oil and plant protein. College of Food, Heilongjiang Bayi Agriculture University, Daqing 163319, Heilongjiang, China. Tel: +86-13836961283, Email: byndwcy@163.com.

even for small sample sizes and are faster than maximum likelihood estimator (MLE), and bayesian least general entropy (BLGE). Bayesian and Bootstrap methods were commonly employed to handle small sample data, addressing equipment evaluation and failure prediction issues, resulting in robust parameter estimation and predictive performance[12-14]. Xu et al.[15] proposed a novel Bayesian stochastic approximation method to enhance the efficiency of sequential designs with limited sample sizes. Luo et al.[16] proposed a small sample sentiment analysis model based on causal analysis theory and naïve Bayes, which reduced the sample size requirements of traditional machine learning while achieving strong classification results. Zhang et al.[17] developed a non-parametric Bootstrap (NBP) estimation method to construct a reliability model under small-sample conditions, which improves the failure time prediction accuracy by 58.3%-91.1% compared with the original model. Heikkinen et al.[18] raised a Bayesian stable isotope mixing model that allows the application of different and specific TDFs for each isotope and each trophic step, which can be useful even with small sample sizes. Qi et al.[19] proposed a double-convergence predictive analysis model based on the combination of Bayesian theory and the deep learning algorithm Cascade-PSPNET for the small-sample problem in tomato yield statistics, and the point estimate of the rate of change of tomato yield from the test sample was the same as the expected value. Bootstrap confidence intervals and approximate confidence intervals were also calculated. Canepa[20] employed nonparametric bootstrap techniques to approximate a Bartlett-type correction for addressing fat-tailed data and relaxing assumptions in small sample scenarios. Numerous studies have affirmed the reliability of Bayes estimation and Bootstrap methods in handling small samples, underscoring the critical relationship between distribution, minimum sample values, and sample size[21-23].

Researchers have recently applied small data analysis techniques to process spectral data in NIRS analysis. Liu et al.[24] combined near-infrared spectroscopy with a discriminative non-negative representation classifier (DNRC) model, which showed better performance in Diarrhetic shellfish poisoning (DSP) toxin detection, with a classification accuracy of 99.44% for a smaller sample dataset. Zheng et al.[25] introduced the minimum sample size for load spectrum measurement and its statistical extrapolation based on GPD parameter estimation and *t*-distribution. James et al.[26] proposed the Bootstrap-based method to equally and adequately represent the confidence intervals for points close to or far away from the latent space to match the performance of well-established methods for spectroscopy data. Wang et al.[27] proposed the Bootstrap-SPXY sample selection method and developed a destructive analysis model for moisture content during the maize grain filling period, building upon the previous research findings. Although the study yielded positive results, it was limited to the destructive analysis of samples.

During the filling period, it is impractical to collect multiple samples, especially in maize breeding and seed production, particularly for parental seeds. And, the consequence of insufficient samples is reduced real-time detection model reliability and accuracy. Therefore, the objective of this study is to propose a non-destructive moisture content analysis with small sample sets during the maize grain filling stages based on NIRS. It is based on the concept of merging and involves calculating cumulative ratios using a modified Bayes Bootstrap (BB) sampling approach. This study introduced the sample optimization method Bayes-Merged-Bootstrap (BMB) to ensure that the imperfect original sample data aligns with the Bayes sample parameter estimation strategy. By optimizing the range of eigenvalues of the samples, the method enhances the quality of the modeled samples. This rapidly develops a non-destructive moisture content testing model for maize kernels during filling. The model's reliability is assessed through a multi-variety sample analysis.

2 Materials and methods

2.1 Materials

The moisture content detection process for maize breeding met the required steps and criteria. It involved four stages: corn removal, threshing, spectral acquisition, and chemical value determination. Samples were obtained from the maize testing base at Heilongjiang Bayi Agricultural University, including three common varieties.

A Bruke Tango-R Fourier transform near-infrared spectrometer with a wavelength range of 4000-10 000 cm^{-1} was used for spectral analysis. Each sample pool contained 50-80 grains. During detection, a cover accessory was placed over the sample pool to shield it from natural light, and the rotary table made one revolution per measurement. Moisture content was calculated using the weight method[28].

Sample-optimized method and modeling are implemented using RStudio 4.3.2 and Matlab 2022, respectively.

2.2 Methods

2.2.1 Bayes Bootstrap Sampling Method

The Bayes Bootstrap (BB) sampling method, also known as the random weighting method, is a Monte Carlo data simulation method[29]. It estimates distribution parameters and intervals after applying random weights to each trial sample[30] Unlike traditional non-parametric Bootstrap algorithms, BB does not rely on understanding the overall data distribution without hypothetical assumptions. It yields small root-mean-square errors, making it suitable for small sample analysis, and it can be extended to larger datasets[31,32].

$$F_n(x)=\begin{cases}0, & x<x_{(i)}\\ \dfrac{i}{n}, & x_{(i)}\le x<x_{(i+1)}\\ 1, & x\ge x_{(n)}\end{cases} \tag{1}$$

In general, using the BB method to resample, the sample data set is sorted from small to large. The sorted data set is indicated as $X=(x_{(1)}, x_{(2)}, x_{(3)}, \ldots, x_{(n)})$, $x_{(1)}<x_{(2)}<\ldots<x_{(n)}$. Under the condition of the sampling hypothesis test, the empirical cumulative ratio F of $x_{(i)}$ can be obtained, as shown in Equation (1). Where, $F_n(x)$ is the empirical accumulation function of the original sample, n is the number of original samples, x is the random variable in the accumulation function, and $x_{(i)}$ is the i-th original sample.

Furthermore, data set X is randomly sampled according to the cumulative ratio formula. However, when resampling, the random number generates a uniform random number in the interval [0, 1]. The overall sampling factor and the sample generating function are constructed based on this random number interval.

In BB resampling, new samples are generated randomly within the range of the original sample values. Although this expands the dataset, it confines the generated samples to the value range of the existing data. However, the expanded data sample does not have the practical information of the centralized data to generate the resampling samples. This has two possible implications: 1) it may weaken or even lose effective data information from the central data; 2) it does not effectively extend the samples beyond the original range. This study aimed to maximize the data range in

actual detection, which the BB algorithm could not achieve alone. Consequently, the merged ideas were added to the BB method in the later study to improve the reliability of the resampled data.

2.2.2 Sample-optimized method based on merging and the Bayes-Bootstrap

According to the previous BB algorithm analysis, the resampling data was mostly unreliable and had invalid or incomplete information. The modified cumulative ratio formula was considered since the resampling principle is mainly based on the cumulative ratio formula of the original sample in the BB algorithm. The improved cumulative ratio formula is shown in Equation (2).

$$F_n(x)=\begin{cases}\dfrac{i}{n}+\dfrac{\left(x-x_i^*\right)}{n\left(x_{i+1}^*-x_i^*\right)}, & x_i^*\le x<x_{i+1}^*,(i=0,1,\ldots,n-k-1)\\ 1-\dfrac{m}{n}\exp\left[-\dfrac{x-x_{n-m}^*}{\beta}\right], & x_{n-m}^*\le x,\beta=\dfrac{1}{m}\left[\dfrac{x_{n-m}^*}{2}+\sum_{i=n-m-1}^{n}x_i-x_{n-m}^*\right]\end{cases} \tag{2}$$

where, n is the number of original samples, m is the number of removed samples from the original samples, and x_i^* is the i-th new sample. The subtraction of the m samples from the original data resulted in the new set of samples containing n-m samples, with the empirical distribution mean replacing the sample mean. Usually, m is less than or equal to 5.

The idea of merging the Bootstrap algorithm, namely the BMB algorithm, has been further added to enhance the effectiveness of resampling samples and reduce the sample generation randomness. The steps to implement the algorithm are as follows:

Step 1: The original sample set was entered and recorded as the set N to calculate the cumulative ratio according to Equation (1). The repumping scale was set as Sample_count.

Step 2: The number of samples in this set was set to be 2 times the number of original samples, i.e., $n^*=2n$, the i-sampling was completed, and the initial value of i was 1.

Step 3: The mean empirical distribution of the sample set formed by Step 2 was calculated. A uniform random number η was generated in the interval [0, 1], and the sequence of n random numbers, U is expressed as U_0, U_1, U_2, …, U_{n-1}, U_n, where U_0=0, U_n=1. The sequence U follows the Dirichlet distribution rule, if $V_i=U_i-U_{i-1}$ (i=1, 2, …, n), then $V_1+V_2+V_3+\ldots+V_{n-1}=1$.

Step 4: The general Bayes sampling factor α was calculated, if $\eta>1-\dfrac{m}{n^*}$, $x_F=x_{(n^*-m)}-\beta ln\left[(1-\eta)\dfrac{n^*}{m}\right]$, otherwise $x_F=x_{(i)}+(\alpha-i+1)(x_{(i+1)}-x_{(i)})$. Where α is the sampling factor, $\alpha=(n-1)\eta$, $i=[\alpha]+1$.

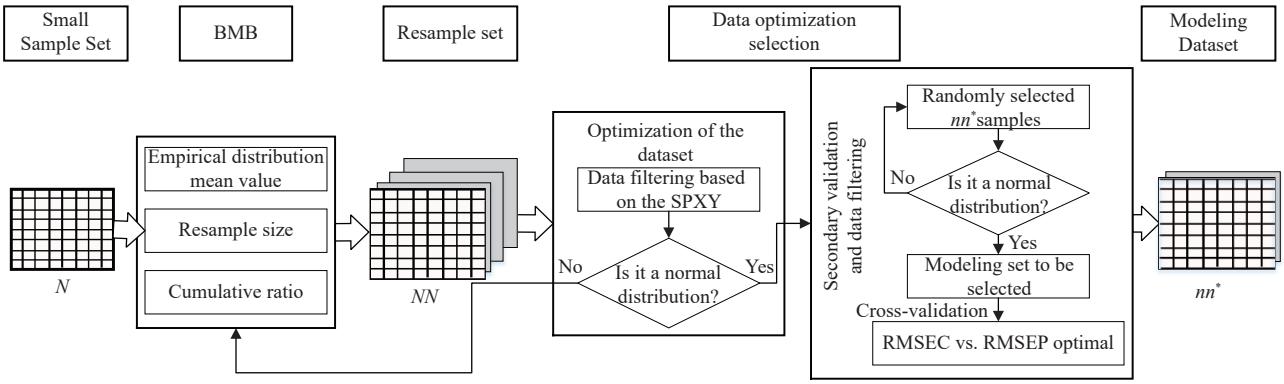
Step 5: The formula for the cumulative ratio $F_{(n)}(x)$ was updated according to Equation (2), and the i-time extracted sample was combined with the original sample to form the set of samples NN.

Step 6: The sample set was checked to see if it met the resampling scale. Step 2 must be reverted to if not, and the sample set N was updated to NN. Otherwise, it ends.

The previous sample was combined with the original sample to enhance model variability and robustness. The resampling sample size is typically denoted as 'n*', and a sample set>10 000 samples is considered statistically significant.

The sample set size after repumping has exceeded the requirements of the number of samples for spectral modeling. If used directly for modeling, the sample set would significantly impact model time complexity and stability. Therefore, sample optimization is necessary. The SPXY sample selection method was considered for optimizing resampled samples, leveraging concentration and spectral feature correlations. After optimization, the sample set size is denoted as 'nn'. These datasets will undergo individual validation for normal distribution. The initial sample optimization step is completed, if one dataset meets the criteria.

The samples resulting from the initial optimization step will then be randomly grouped, each group containing 'nn*' samples, and subjected to cross-validation testing. If normal distribution criteria are met, sample optimization is finalized, and the modelling sample set is established. Otherwise, a subset will be reselected. Figure 1 illustrates the BMB algorithm-based sample optimization method.



Note: BMB: Bayes-Merged-Bootstrap. Same below.

Figure 1 Process of the sample set optimization method based on merging and Bayes-Bootstrap sampling

2.3 Model prediction and evaluation index

To assess the model's predictive power, R^2 and RPD were employed. The root-mean-square error of cross-validation (RMSECV) and root-mean-square error of prediction (RMSEP) were selected as additional evaluation indices since this experiment primarily involves modeling data. Close values for these two indices indicate the effectiveness of the modeling dataset generated using the proposed method in this article. If RMSECV significantly exceeds RMSEP, it suggests poor representativeness of the validation sample. Conversely, if RMSEP greatly exceeds RMSECV, it indicates inadequate or overly tight information fitting in the modeling sample.

3 Results and discussion

3.1 Near-infrared spectroscopy data and chemical reference value

The grains underwent spectral data collection without any treatment, preserving the characteristics of the original components.

This approach qualifies as non-destructive testing. Figure 2 displays the average absorption spectral curves of the average Fourier NIR of the three maize varieties (‘Demeiya’, ‘Xianyu 335’, and ‘Zhengdan 958’) at seven stages of the filling period. By observing the spectral curve in the Figure, the absorption peak of the spectral wavelength around 6900 cm^{-1} varies in different grouting stages, with the difference in maize grain water content corresponding to the difference and change of the spectral curve. Higher moisture content results in an increased absorbance value within this band.

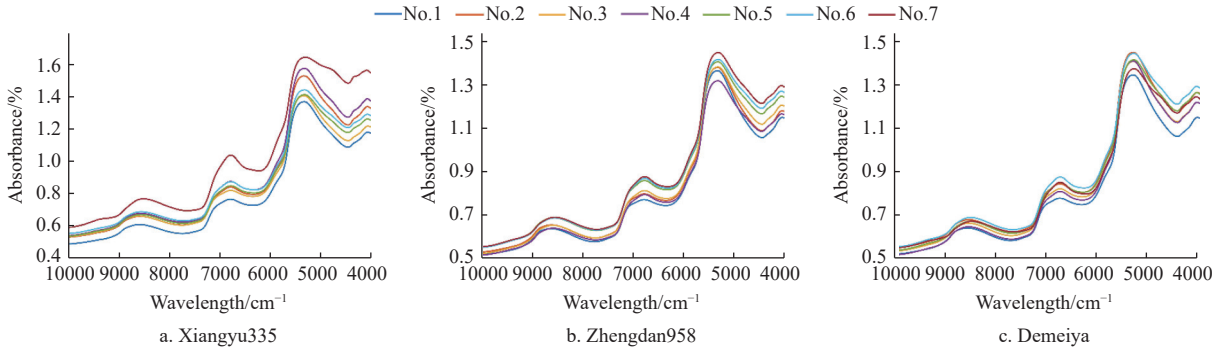


Figure 2 Average Fourier NIR absorption spectra of the three maize varieties during filling stage

Figure 3 shows the actual measurements and error analysis for the 3 varieties. Each breed yielded 100 effective spectral samples at seven stages of the filling period, with 50 samples designated for small-scale data modeling trials and 50 for prediction sets. Given the scarcity of breeding samples and limited collection time, it is assumed that constraints on sample availability resulted in a few spectra obtained during the experiment. To demonstrate the method’s ability to handle different sample numbers, the number of constructed samples is divided into three subsets, denoted as the set X, and X_fifty=50, X_twenty=20, X_ten=10.

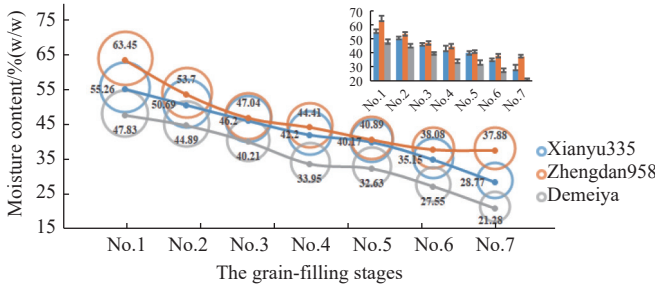
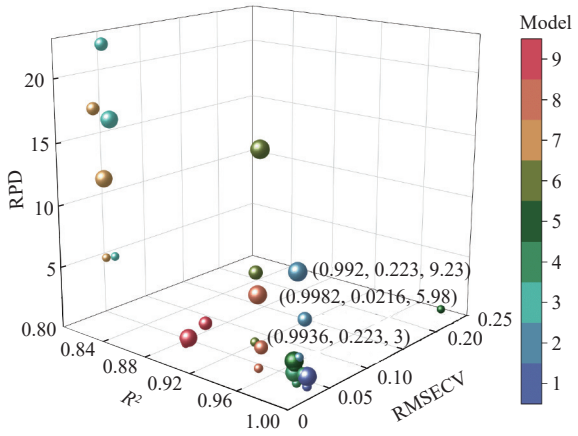


Figure 3 Practical measured values of three varieties during the filling stage

3.2 Analysis of non-destructive maize moisture content measurement modeling based on BMB

In order to create an effective moisture content detection model based on BMB, a set of optimized and raw sample data is selected as the modeling sample set, and nine moisture content detection models are created respectively, including the BMB-PLS model and BMB-nu-SVR and BMB-Epsilon-SVR model with four kernel functions. The performance and usability of the above model are evaluated through R^2, RMSECV, and RPD values. The values of these models of re-sample sizes 50, 20, or 10 are shown in Figure 4. Due to the unsatisfactory performance of the models based on the raw samples, it is not shown in the figure. It can be seen that the models in the first region where R^2 and RMSECV coordinate axes intersect are relatively better. Among these models, model-1 is based on PLS, Model-2 is based on nu-SVR using an RBF kernel function, and Model-6 is based on Epsilon-SVR using an RBF kernel function. Comparing the RPD value, model-2 is the best. So, the most effective non-destructive maize moisture content measurement with small sample sets model is the BMB optimization selection method and nu-SVR with an RBF kernel function, penalty coefficient C=5, Nu=0.5, and Gama= 0.000 685 871 1.



Note: Model-1 (BMB-nu-SVR-Linear), Model-2 (BMB-nu-SVR-RBF), Model-3 (BMB-nu-SVR-Polynomial), Model-4 (BMB-nu-SVR-Sigmoid), Model-5 (BMB-Epsilon-SVR-Linear), Model-6 (BMB-Epsilon-SVR-RBF), Model-7 (BMB-Epsilon-SVR-Polynomial), Model-8 (BMB-Epsilon-SVR-Sigmoid), Model-9 (BMB-PLS). Big, medium, and small spheres represent the original sample sizes of 50, 20, and 10, respectively.

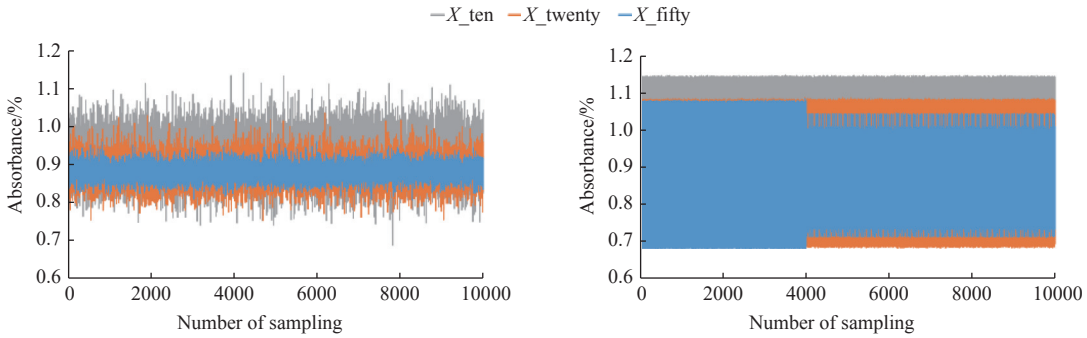
Figure 4 Evaluation comparison of different models

3.3 Analysis and evaluation of Sample Resampling

Resampling was conducted on the three original sample sets, namely X_fifty, X_twenty, and X_ten, using the BB and BMB algorithms. A resampling size of 10 000 was applied for both algorithms, and the eigenvalues were derived from the interval θ, mean m, and standard deviation μ of the resampled samples. Spectral band selection was set at 6900 cm^{-1}, specifically focusing on the ‘Demeiya’ band. The resampling analysis results are presented in Figures 5 and 6.

Figure 5 illustrates the sampling process of the BMB and BB algorithms. The BMB algorithm consistently maintains resampling samples around the sampling experience threshold. As the number of samples decreases from 50 to 20 and further to 10, the sampling interval expands, likely due to the incorporation of the sampling experience threshold. Additionally, the BB algorithm exhibits larger sampling intervals than the BMB algorithm, but the variation among the resampled data points is smaller than that of the BMB.

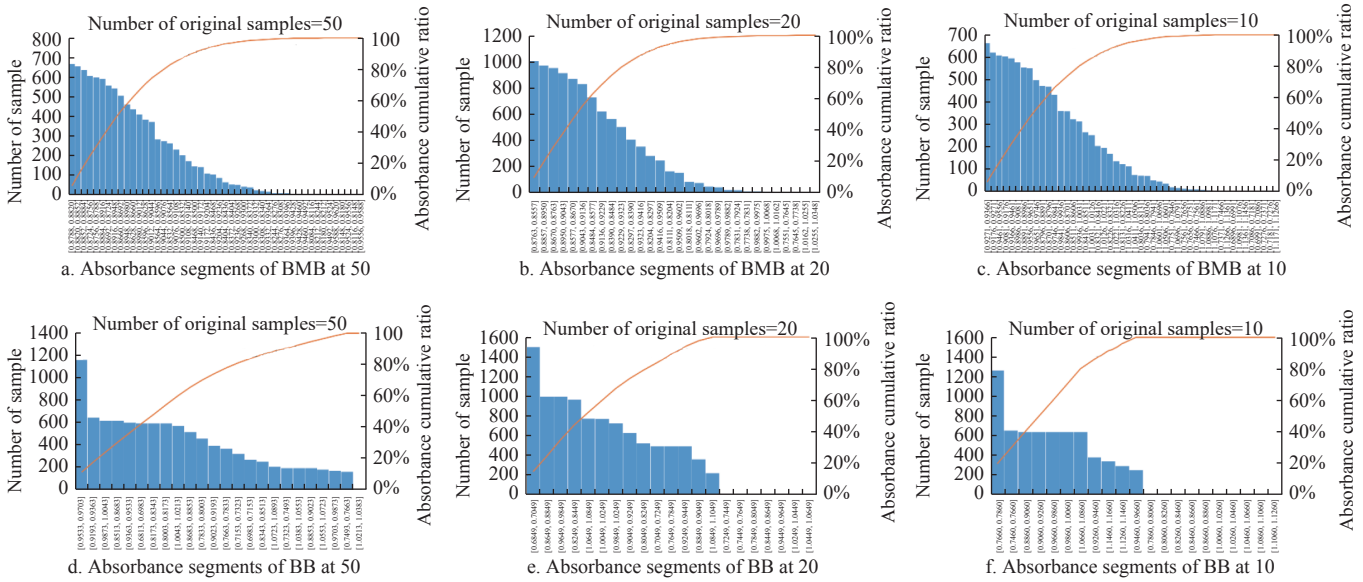
Notably, with an original sample size of 50 and a resampling size> 4000, the sample interval experiences a sharp change. Overall, the BMB algorithm demonstrates superior sampling performance across different original sample sets compared with the BB algorithm.



a. Sampling analysis diagram of BMB algorithm b. Sampling analysis diagram of BB algorithm

Note: BB: Bayes Bootstrap. Same below.

Figure 5 Comparison diagram of sampling



a. Absorbance segments of BMB at 50 b. Absorbance segments of BMB at 20 c. Absorbance segments of BMB at 10

d. Absorbance segments of BB at 50 e. Absorbance segments of BB at 20 f. Absorbance segments of BB at 10

Figure 6 Comparison diagram of the cumulative sampling ratio

The results of the two algorithms were analyzed using a cumulative ratio plot shown in Figure 6. A gradual, steady decrease in the number of resamples by the BMB algorithm was observed in segments with different absorption rates, indicating improved sample coverage within the resampled value interval. In contrast, the BB algorithm exhibited a phased and unbalanced change in the sampling accumulation ratio. For the original sample size of 50, there were 1165 samples within the segmented absorbance values [0.9533, 0.9703], twice the number of samples within the second-highest segmented absorbance values [0.9193, 0.9363]. The cumulative sample ratio demonstrated that a decrease in the original samples directly impacted the resampling coverage in the absorption rate values. This significant, trend resulted in low sample coverage within each absorbance value segment When the original sample size was reduced to twenty, seven absorbance values out of twenty-one had no corresponding resampled samples, resulting in a coverage ratio of 14:21. For an original sample size of 10, the coverage ratio dropped to 11:21. However, with an original sample number of 50, the coverage ratio improved to 23:24.

Table 1 lists the resample eigenvalues. It is observed that when the raw sample size is 50, 20, or 10, the mean (m) of the two algorithms is nearly identical, differing by only 0.0001. However, the standard deviation of the BB algorithm exceeds that of the BMB algorithm, with the highest difference being 8.61%. Analyzing the sample intervals, the resampling value range of the BB algorithm remains constant. In contrast, when the raw sample size is 50, the BMB algorithm's sample interval is only 22% of that of the BB algorithm. With a raw sample size of 20, the resampling range expands. Further, when the raw sample size is 10, the BMB algorithm's resampling range extends to 59.3% of that of the BB algorithm.

In summary, the BMB algorithm exhibits distinct advantages over the BB algorithm in resampling sample effects. First, it concerns the resampling set's interval size. While the BMB algorithm's interval is smaller than the original sample range at 50, 20, and 10, it maintains significant sampling randomness within this interval. In scenarios with a smaller sample size, an increased number of resamplings occur on the same scale. This highlights the significance of resampling in experiments, with the largest interval range at 10. Second, it pertains to the coverage ratio of the absorbance segment values within the resampled volume. The BMB algorithm consistently demonstrates a significantly higher coverage ratio at 50, 20, and 10 than the BB algorithm in absorbance segment values. This indicates that the BMB algorithm can complement bands not present in the original sample, thereby enhancing its

validity. The empirical threshold of the BMB algorithm adapts to the sample size changes during operation. As the original sample size decreases, the number of times the threshold changes increases. Consequently, the resampling set's interval range expands as the original sample size decreases, further emphasizing the merging concept within the BMB algorithm.

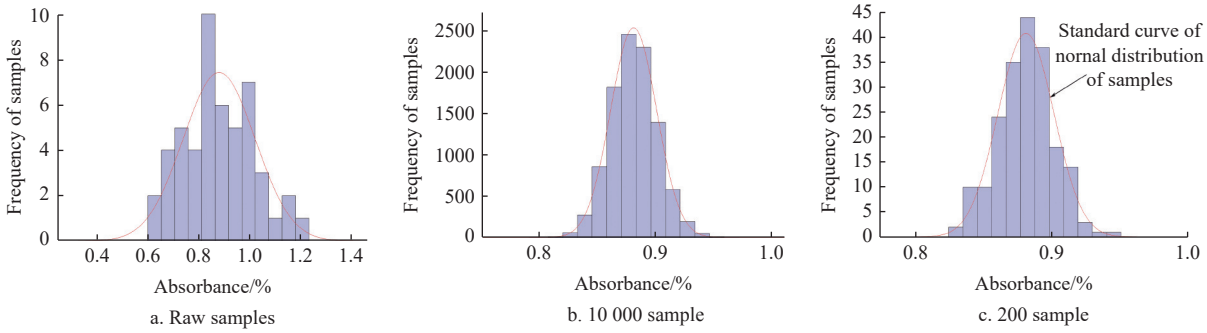
Table 1 The re-sample eigenvalues

Algorithm	Interval (θ)			Mean (m)			standard deviation (μ)/%		
	50	20	10	50	20	10	50	20	10
BMB	(0.8084, 0.9598)	(0.7551, 1.0348)	(0.6896, 1.1379)	0.8819	0.8844	0.9256	1.98	3.66	5.85
BB	(0.6813, 1.0864)	(0.6849, 1.0913)	(0.7561, 1.1508)	0.8818	0.8844	0.9257	10.46	12.27	12.35

Note: BMB: Bayes-Merged-Bootstrap; BB: Bayes Bootstrap.

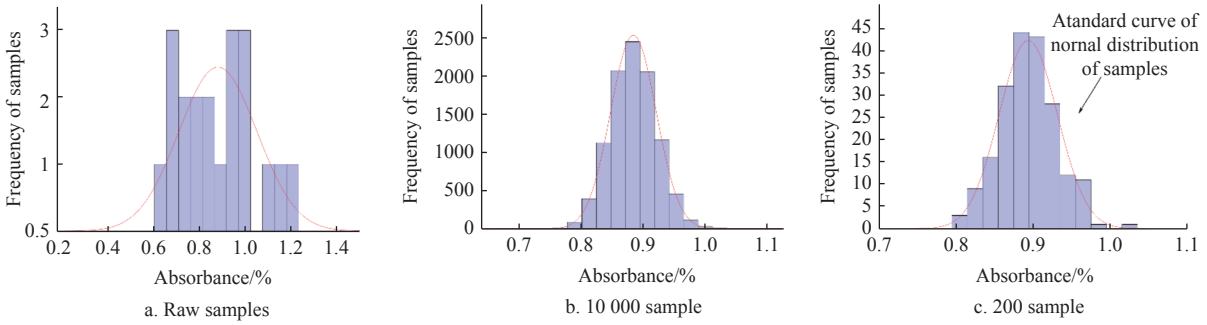
3.4 Analysis of sample size in resampling

Samples in the 6900 cm^{-1} band sensitive to water molecules, were selected for the study. As shown in Figures 7-9, a histogram of the set was plotted before and after optimized selection for comparison and the effect of sampling size on the modeled set was analyzed.



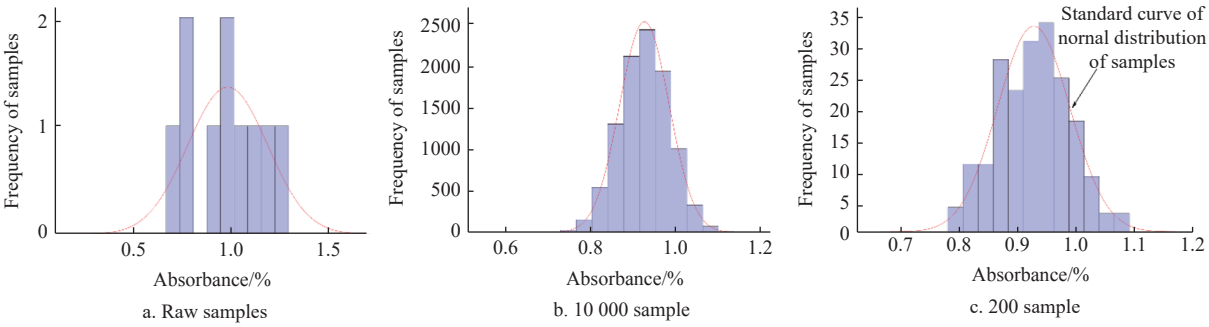
a. Raw samples b. 10 000 sample c. 200 sample

Figure 7 Comparison of samples distributions before and after optimization selection of X_fifty



a. Raw samples b. 10 000 sample c. 200 sample

Figure 8 Comparison of samples distributions before and after optimization selection of X_twenty



a. Raw samples b. 10 000 sample c. 200 sample

Figure 9 Comparison of samples distributions before and after optimization selection of X_ten

In comparing distribution frequencies before and after optimized selection from the original sample set X_ fifty, X_twenty, and X_ten, it is observed that the absorption rate values of the three original small sample sets exhibit an inclusive relationship within the range [0.6, 1.25]. However, they do not conform to a normal distribution.

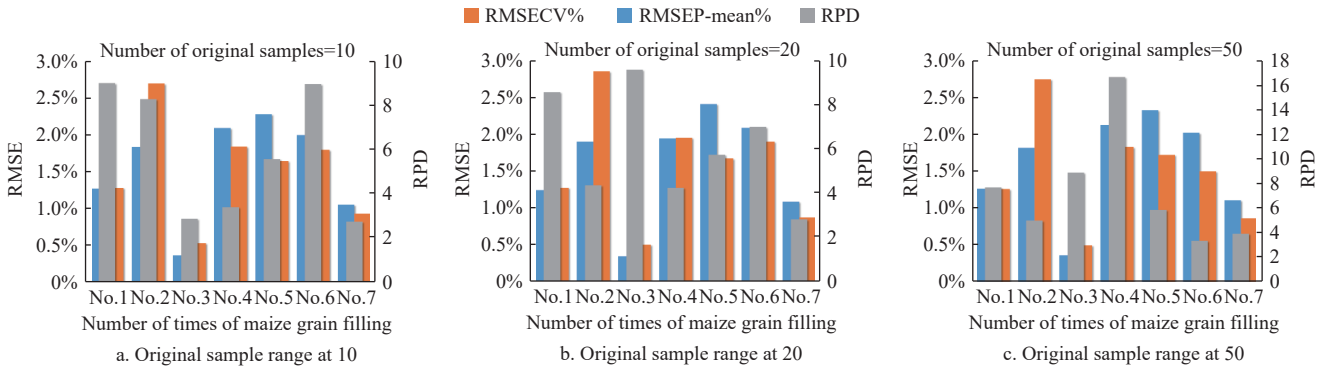
The BMB algorithm is configured with 10 000 iterations to create the resampling sets X*_fifty, X*_twenty, and X*_ten, resulting in modeling subsets denoted as X'_fifty_Subset, X'_twenty_Subset, and X'_ten_Subset. The sample size was set at 200. It is evident that all these resampled and optimized sets, X_fifty, X_twenty, and X_ten, follow normal distributions, exhibiting significantly improved distribution characteristics compared to the original sample dataset. The frequency peaks and absorption rate intervals differ slightly among the three sets. X'_fifty_Subset has a maximum frequency value of 0.88-0.92, with

an absorption rate range of [0.83, 0.95]. *X′_twenty_Subset* exhibits a frequency value of 0.85-0.95 and an absorption rate range of [0.79, 1.03]. Finally, *X′*_ten _Subset's maximum frequency value ranges from 0.9 to 1.0, with an absorption rate range of [0.78, 1.09]. After comparing these distributions to the original sample, it can be concluded that when the original sample size is 10 and the resampling size is 10 000, the absorption rate value interval and peak location most closely resemble those of the original sample. This correlation arises from the equal size of resampling samples; as the original sample set size increases, the number of combined resampling samples decreases. Thus, the number of resampling samples is linked to the original sample size. When the resampling sample size is a thousand times or greater than the original sample size, the loss of absorption rate value is minimized.

3.5 Analysis of the model effect during filling periods

Following the BMB-based small sample processing method, sample sets *X*_fifty, *X*_twenty, and *X*_ten were resampled using the BMB algorithm at a scale of 10 000, resulting in the formation of *X**_fifty, *X**_twenty, and *X**_ten. Subsequently, employing the SPXY selection method, the sample sizes were reduced to 2000, creating *X′*_fifty, *X′*_twenty, and *X′*_ten. These sets were then randomized, with each cycle forming a corresponding subset of 200 samples. If the current subset showed normal and optimally distributed distribution, it would be recorded as the subset to be modeled. After cross-validation, one of the subsets was selected as the modeling set of the respective sample set and was denoted as *X′*_fifty_Subset, *X′*_twenty_Subset, and *X′*_ten_Subset.

Spectral data samples from seven different moisture content periods during the "Demeiya" grouting process were analyzed using the BMB-SVR for the full spectrum model. Figure 10 presents the maize grain moisture content model evaluation results across various small sample sizes during the filling stages. The determination coefficient R^2 can reach 0.99 in different filling stages and for varying small sample sizes. The lowest determination coefficient of the model is 0.9896, observed during the second sampling stage with 50 samples. The RMSECV value ranged from 1.0% to 2.5%. By comparing the blue and orange columns in Figure 10, it is evident that the RMSECV and RMSP values are closely aligned, differing by <1%. Additionally, all RPD values for the model exceeded 2.4, indicating that the predictive ability of the model meets the general accuracy standards required for practical applications.

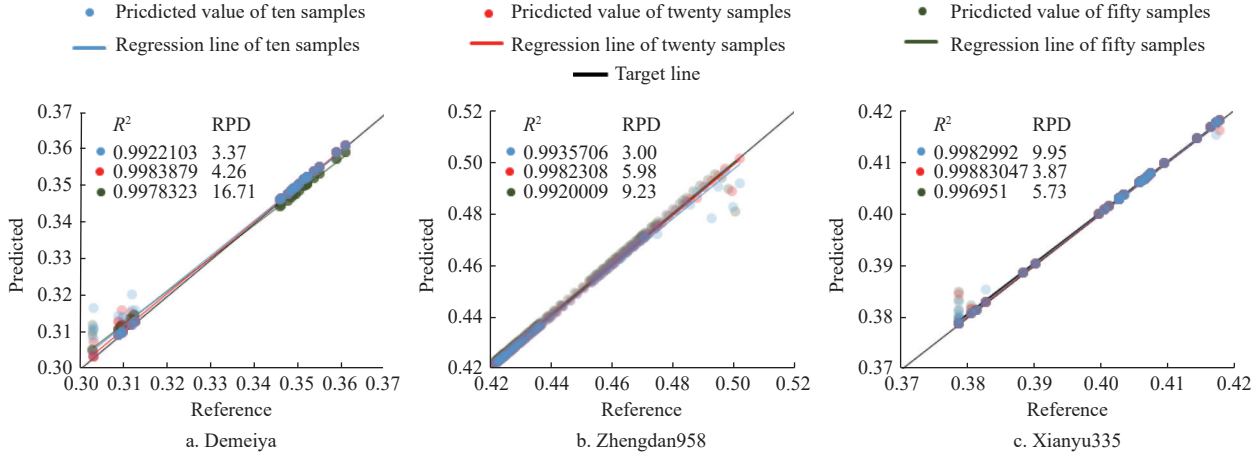


a. Original sample range at 10 b. Original sample range at 20 c. Original sample range at 50

Figure 10 Evaluation of the BMB-SPXY-SVR model in different sample sizes in grain filling

3.6 Analysis of model generality based on different varieties

To assess the model's generality, spectral datasets from the fourth sampling stage were collected for all three varieties. The prediction results are presented in Figure 11, where different colors correspond to various sample sizes. Blue, red, and green represent the original sample sizes of 10, 20, and 50, respectively. Lines depict the regression trends, and dots indicate predicted values. The black line represents the reference value trend. It is evident that the regression lines of models constructed with different sample sizes closely align with the target line during the current filling period of the three varieties. The models consistently achieved R^2>0.99, with RPD>3.0 for all three varieties. Notably, when sample sizes are 10 and 50, the regression line trends deviate slightly from the target line. The most favorable outcome occurs when the sample size is 20, where the regression line almost perfectly coincides with the target line. It is shown that when the small sample number is greater



a. Demeiya b. Zhengdan958 c. Xianyu335

Figure 11 Prediction and analysis diagram of the full-spectrum model sampled during the fourth grain filing on different small sample sizes of three varieties

than 10, the model built by the method proposed in this paper can meet practical applications across multiple varieties. Furthermore, considering the RPD values 'Zhengdan 958' and 'Demeiya', it is observed that these values increase with larger sample sizes, reaching above 4 with a sample size of 20. Although 'Xianyu' exhibits different characteristics compared to the other two varieties, it still achieves an RPD value of 3.87, with a sample size of 20. Therefore, a sample size of at least 20 is recommended for small samples.

4 Conclusions

1) The proposed Bayes-Merged-Bootstrap (BMB) small sample optimization method replaces the sample mean with the empirical distribution mean to extend the resampling data interval. This enhances the completeness of data information obtained from resampling, addressing issues of incomplete distribution and missing data in small samples.

2) The BMB algorithm significantly outperforms the Bayes Bootstrap (BB) algorithm in sample resampling effectiveness. The coverage ratio of absorbance segmentation values in the optimized dataset, based on the BMB algorithm, surpasses that of the BB algorithm when the original small-sample data intervals are comparatively smaller. Furthermore, the optimized dataset supplements the absent bands in the original small sample set. The BMB small-sample optimization method establishes the foundation for further modeling experiments.

3) The BMB algorithm predefines the number of resampling samples when resampling the sample. Smaller small sample sets undergo comparatively more frequent resampling. Hence, the key factors affecting small sample sampling effectiveness include the original small sample size and resampling scale. Using the water-molecule-sensitive band as an example, this paper showed that the sample distribution works optimally when the resampled sample size is at least a thousand times larger than the original sample.

4) Finally, a non-destructive quantitative analysis model for maize grain moisture content during the filling period was constructed based on the BMB small sample optimization method and SVR. Three common maize varieties in Northeast China were used as examples. The models achieved R^2>0.989, and RPD>2.47. These results demonstrate the model's robustness across various sample varieties, small sample sizes, and moisture content levels. It meets the maize grain moisture content detection requirements in breeding and production management for different maize varieties.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (General Program) (Grant No. 52275246), Natural Science Foundation of Heilongjiang Province(No.LH2022C061), Heilongjiang Province Postdoctoral Fund (Grant No. LBH-Z19217), Heilongjiang Bayi Agricultural University Three Horizontal and Three Vertical Support Plans (Grant No. ZRCQC201907), and Heilongjiang Bayi Agricultural University Adult Talent Research Startup Fund (Grant No. XDB202004).

[References]

[1] Zhu Y D, He H J, Jiang S Q, Ma H J, Chen F S, Xu B C, et al. Mining hyperspectral data for non-destructive and rapid prediction of nitrite content in ham sausages. Int J Agric & Biol Eng, 2021; 14(2): 182–187.

[2] Song P, Kim G, Song P, Yang T, Yue X, Gu Y. Rapid and non-destructive detection method for water status and water distribution of rice seeds with different vigor. Int J Agric & Biol Eng, 2021; 14(2): 231–238.

[3] Zhao X, Xing L Y, Shen S F, Liu J M, Zhang D X. Non-destructive 3D geometric modeling of maize root-stubble in-situ via X-ray computed tomography. Int J Agric & Biol Eng, 2020; 13(3): 174–179.

[4] Zhang M Q, Zhao C, Shao Q J, Yang Z D, Zhang X F, Xu X F, et al. Determination of water content in corn stover silage using near-infrared spectroscopy. Int J Agric & Biol Eng, 2019; 12(6): 143–148.

[5] de Medeiros D T, Ramalho F M G, Batista F G, Mascarenhas A R P, Chaix G, Hein P R G. Water desorption monitoring of cellulose pulps by NIR spectroscopy. Industrial Crops and Products, 2023; 192: 115989.

[6] Su K, Maghirang E, Tan J W, Yoon J Y, Armstrong P, Kachroo P, et al. NIR spectroscopy for rapid measurement of moisture and cannabinoid contents of industrial hemp (*Cannabis sativa*). Industrial Crops and Products, 2022; 184: 115007.

[7] Bernhard ., Grani C. Striking a balance in Fabry disease research: Mitigating the statistical dilemma arising from small sample size and modest event frequency in rare disorders. International Journal of Cardiology, 2023; 384: 52–53.

[8] Cao J Z, Wang T, Sheng M, Huang Y, Mo P Q, Zhou G Q. Assessment of multi-dimensional joint probability distribution for uncertain mechanical strength parameters under small sample test data. Probabilistic Engineering Mechanics, 2023; 74: 103511.

[9] Ma T W, Wang G, Guo R, Chen L, Ma J F. Forest fire susceptibility assessment under small sample scenario: A semi-supervised learning approach using transductive support vector machine. Journal of Environmental Management, 2024; 359: 120966.

[10] Li X Z, Su H, Xiang L, Yao Q T, Hu A J. Transformer-based meta learning method for bearing fault identification under multiple small sample conditions. Mechanical Systems & Signal Processing, 2024; 208: 110967.

[11] Aydi W, Alduais F S. Estimating Weibull parameters using least squares and multilayer perceptron vs. Bayes estimation. Computers, Materials & Continua, 2022; 71(2): 4033–4050.

[12] Kim I, Kang H-Y, Khang Y H. Comparison of Bayesian spatio-temporal models for small-area life expectancy: A simulation study. American Journal of Epidemiology, 2023; 192(8): 1396–1405.

[13] Peng Z Q, Sun Z Y, Chen J, Ping Z, Dong K Y, Li J, et al. A fault diagnosis approach for electromechanical actuators with simulating model under small experimental data sample condition. Actuators, 2022; 11(3): 66.

[14] Pawel S, Consonni G, Held L. Bayesian approaches to designing replication studies. arXiv e-prints, 2022.

[15] Xu J, Mu R J, Xiong C. A Bayesian stochastic approximation method. Journal of Statistical Planning and Inference, 2021; 211: 391–401.

[16] Luo Y J, Yang X H, Ouyang C P, Wan Y, He S X. Merging naive Bayes and causal rules for text sentiment analysis. Journal of Physics: Conference Series, 2021; 1757: 012034.

[17] Zhang W, Jiang J H. Bootstrap-based resampling methods for software reliability measurement under small sample condition. Journal of Circuits, Systems and Computers, 2024; 33(9): 2450161.

[18] Heikkinen R, Hmlinen H, Kiljunen M, Kärkkäinen S, Schilder J, Jone R I. A Bayesian stable isotope mixing model for coping with multiple isotopes, multiple trophic steps and small sample sizes. Methods in Ecology and Evolution, 2022; 13(11): 2586–2602.

[19] Qi C C, Liu W Z. Tomato production prediction based on deep learning algorithm-Cascade-PSPNET and Bayes. International Journal of Pattern Recognition and Artificial Intelligence, 2023(14): 37.

[20] Canepa A. Small sample adjustment for hypotheses testing on cointegrating vectors. Journal of Time Series Econometrics, 2022; 14(1): 51–85.

[21] Jha M K, Tripathi Y M, Dey S. Multicomponent stress-strength reliability estimation based on unit generalized Rayleigh distribution. International Journal of Quality & Reliability Management, 2021; 38(10): 2048–2079.

[22] Kline B. Bayes factors based on *p*-values and sets of priors with restricted strength. The American Statistician, 2022; 76(3): 203–213.

[23] Yang X Y, Xie L Y, Zhao B F, Kong X W, Wu N X. An iterative method for parameter estimation of the three-parameter Weibull distribution based on a small sample size with a fixed shape parameter. International Journal of Structural Stability and Dynamics, 2022; 22(12): 2250125.

[24] Liu Y, Liu Z Y, Qiao F, Xu L L, Xu Z. Identification of *Perna viridis* contaminated with diarrhetic shellfish poisoning toxins in vitro using NIRS and a discriminative non-negative representation-based classifier. Spectrochimica Acta Part A: Molecular and biomolecular Spectroscopy,

2023; 294: 122514.

[25] Zheng G F, Wang Q Y, Cai C. Criterion to determine the minimum sample size for load spectrum measurement and statistical extrapolation. Measurement, 2021; 178: 109387.

[26] Odgers J, Kappatou C, Misener R, García Muñoz S, Filipp S. Probabilistic predictions for partial least squares using bootstrap. AIChE Journal, 2023; 69(7): e18071.

[27] Wang X, Ma T, Yang T, Song P, Xie Q, 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): 203–210. (in Chinese)

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

[29] Goldstein H, Carpenter J, Kenward M G. Bayesian models for weighted data with missing values: a bootstrap approach. Journal of the Royal Statistical Society Series C: Applied Statistics, 2018; 67(4): 1071–1081.

[30] Kleebmek T, Thongmual N. Estimating missing data with Bayes Bootstrap regression imputation. Burapha Science Journal, 2021; 26(2): 816–826. (in Thai language)

[31] Luo X H, Wei C D, Wang Y R. Comparison of reliability parameters estimation methods for exponential distribution under small samples. Statistics & Decision, 2018; 493(1): 14–17. (in Chinese)

[32] Salazar J J, Pyrcz M J. Geostatistical significance of differences for spatial subsurface phenomenon. Journal of Petroleum Science and Engineering, 2021; 203: 108694.