Identification of maize seed varieties based on near infrared reflectance spectroscopy and chemometrics

Yongjin Cui[1], Lanjun Xu[2], Dong An[1,3*], Zhe Liu[1], Jiancheng Gu[4], Shaoming Li[1], Xiaodong Zhang[1], Dehai Zhu[1,3]

(1. *College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China*;
2. *Beijing Agricultural Machinery Experiment Appraisal and Popularization Station, Beijing 100079, China*;
3. *Key Laboratory of Agricultural Information Acquisition Technology (Beijing), Ministry of Agriculture, Beijing 100083, China*;
4. *Beijing Kings Nower Seed S&T Co., Ltd., Beijing 100080, China*)

Abstract: False seeds can often be seen in the maize seed market, leading to a serious decline in maize yield. Those existing variety identification methods are expensive, time consuming, and destructive to seeds. The aim of this study is to develop a cheap, fast and non-destructive method which can robustly identify large amounts of maize seed varieties based on near-infrared reflectance spectroscopy (NIRS) and chemometrics. Because it is difficult to establish models for every variety in the market, this study mainly investigated the performance of models based on a large number of samples (more than 40 major varieties in the market). The reflectance spectra of maize seeds were collected by two modes (bulk kernels mode and single kernel mode). Both collection modes can be applied to identification, but only the single kernel mode can be applied to purity sorting. The spectra were pretreated with smoothing, the first derivative and vector normalization; and then principal component analysis (PCA), linear discriminant analysis (LDA) and biomimetic pattern recognition (BPR) were applied to establish identification models. The environmental factors such as producing areas and years have a significant influence on the performance of the models. Therefore, the method to improve the robustness of the models was investigated in this study. New indexes (correct acceptance degree (CAD), correct rejection degree (CRD) and correct degree (CD)) were defined to analyze the performance of the models more accurately. Finally, the models obtained a mean correct discrimination rate of over 90%, and exhibited robust properties for samples harvested from different areas and years. The results showed that NIR technology combined with chemometrics methods such as PCA, LDA, and BPR could be a suitable and alternative technique to identify the authenticity of maize seed varieties.

Keywords: maize, seed variety identification, near-infrared reflectance spectroscopy (NIRS), biomimetic pattern recognition (BPR)

DOI: 10.25165/j.ijabe.20181102.2815

Citation: Cui Y J, Xu L J, An D, Liu Z, Gu J C, Li S M, et al. Identification of maize seed varieties based on near infrared reflectance spectroscopy and chemometrics. Int J Agric & Biol Eng, 2018; 11(2): 177–183.

1 Introduction

Maize is an important food crop and industrial raw material. Different varieties lead to differences in many characters of maize, such as yield, quality, disease and insect resistance, and abiotic stress tolerance. Therefore the identification of the maize variety is an essential thing. There are many traditional and prevailing methods that can identify the varieties of maize (e.g., seed cultivar identification, grain morphology, fluorescent scanning, protein electrophoresis and DNA molecular markers)[1-3]. But all of these methods must be operated in the lab, and the result cannot be obtained very rapidly. In some occasions the identification of maize varieties needs to be done on-the-spot and the result needs to be got in a very short time. Besides, another shortcoming of the traditional and prevailing methods is the cost. For example, the cost of the DNA fingerprinting method is about ￥1000 RMB per maize kernel, the cost of the field planting method is about ￥100 RMB per maize kernel, and the cost of the protein electrophoresis is about ￥10 RMB per maize kernel. Furthermore, the traditional methods cause damage to seeds, but only the non-destructive identification method can be applied to the purity sorting of substandard seeds.

The near infrared spectral region is consistent with the absorption region of the vibrational frequency of the hydrogen containing groups (e.g. C-H, N-H, O-H) in the organic molecules, which makes the near infrared spectral region contain a wealth of material information[4]. As an optical signal, the near infrared spectrum also has the advantage of easy access, high speed, non-destruction. And the identification method based on spectra

Received date: 2016-08-26 **Accepted date:** 2017-06-20

Biographies: Yongjin Cui, Master, research interests: artificial intelligence, Email: cuiyongjin@cau.edu.cn; **Lanjun Xu,** Bachelor, Senior Agricultural Engineer, research interests: agricultural informatization, Email: 250980662@qq.com; **Zhe Liu,** PhD, Associate Professor, research interests: seed industry information technology, Email: liuzhe_23@163.com; **Jiancheng Gu,** Bachelor, Agronomist, research interests: seed engineering, Email: gu8024@163.com; **Shaoming Li,** PhD, Associate Professor, research interests: seed industry information technology, Email: lshaoming@sina.com; **Xiaodong Zhang,** PhD, Professor, research interests: seed industry informatization, Email: zhangxd@cau.edu.cn; **Dehai Zhu,** PhD, Professor, research interests: agricultural informatization, Email: zhudehai@263.net.

***Corresponding author: Dong An,** PhD, Professor, research interests: signal processing. College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China. Tel: +86-13366883406, Email: andong@cau.edu.cn.

is much cheaper because it does not need any chemical as the protein electrophoresis method does and any field as the field planting method does. And the entire spectrum acquisition process takes only a few seconds once. All of these advantages make the near infrared spectroscopy analysis technology unique in the analysis of organic matter. The near-infrared spectra of maize seeds were taken as the identification information in this study. In the application of near-infrared reflectance spectroscopy (NIRS) on crop variety identification, Zhou et al.[5] reduced the dimension of spectra of Longjing tea and the other 10 varieties by PCA, and then classified the samples into two types of Longjing tea and non Longjing by Fisher classification function; the correct rate was over 93%. Liang et al.[6] reduced the dimension of spectra of 5 rice varieties by PCA, and then classified rice varieties by neural network algorithm; the correct rate of the test set was 100%. Cao et al.[7] reduced the dimension of the spectra of 4 grape varieties by PCA, and then classified grape varieties by neural network algorithm; the correct rate was over 97%.

In the application of NIRS on crop producing area identification, Zhuang et al.[8] classified Laoshan tea and Rizhao tea by neural network algorithm, the correct rate was over 98.33%.

In all researches above, people only made the variety or the producing area to be the only variable in the experimental design. But with the deepening of research on the qualitative analysis of crops using NIRS, the problem of this kind of experimental design is gradually revealed. By analyzing the near infrared spectra of 8 varieties of maize, Han et al.[9] came to the conclusion that the producing areas have a significant influence on the spectrum, and the discrimination of the producing areas reached 95% more than the discrimination of varieties (90%). In the study of near infrared spectra of maize, Liu[10] came to the conclusion that the influence degree of producing areas was 0.9740, the influence degree of variety was 0.8111, and the spectral difference between the samples of the same variety but different producing areas makes the identification rate of the variety worse.

The producing year is also an important factor affecting the spectra of crops. Zhao[11] came to the conclusion that there is a large gap in the feature space between samples harvested from different years, which leads to miscarriage and a low recognition rate. In the identification of wheat producing areas, Zhao et al.[12] achieved a recognition rate of over 87% for samples from different areas of the same producing year, but when identifying samples from another producing year with the existing models, the recognition rate is 48.3%.

The researches above show that the producing areas and years have a significant influence on spectra, which leads to a serious decline in the variety identification rate. At present, the application of NIRS in the identification of crop varieties is limited to the laboratory research stage without any promotion in the actual production. The influence of producing areas and years is one of the key factors that restrict the practical application. Therefore, in this study producing area and producing year were both considered as variables besides varieties in order to explore feasible methods to overcome their impacts.

In addition, there are so many varieties of crops, but the previous researches above only took 4-5 varieties into consideration, which makes the researches lack of practical value. Therefore, a large number of samples (more than 40 major maize varieties in the market) were applied in this study. And the biomimetic pattern recognition (BPR) was applied to build the identification model, which is proved to be more stable and more excellent than SIMCA and PLS-DA[13-15]. To analyze the performance of the models more accurately, new indexes (correct acceptance degree (CAD), correct rejection degree (CRD) and correct degree (CD)) were defined besides the correct acceptance rate (CAR) and correct rejection rate (CRR).

In order to make the research results applicable to purity sorting, the reflectance spectra of maize seeds were collected in two modes (bulk kernels mode and single kernel mode). Both of the two modes can be used into identification, but only the single kernel mode can be used into purity sorting.

In general, the aim of this study is to develop a cheap, fast and non-destructive method which can robustly identify large amounts of maize seed varieties based on NIRS and chemometrics (PCA, LDA, BPR).

2 Materials and methods

2.1 Sample preparation and spectral data acquisition

Common NIRS bulk sample analyzers provide measurements of samples of about 250 g on average. Single seed differences cannot be identified and no discrimination of single kernel is possible[16]. The sufficient information can be obtained from different parts of different seeds by measuring spectra in bulk kernels mode, but it does not work when identifying impure or damaged seeds. It is more appropriate to measure spectra in single kernel mode in these occasions.

To investigate the feasibility of identifying maize variety based on NIRS fully, both spectra of bulk kernels and single kernel were collected. All NIR spectra were collected with Fourier transform near infrared (FT-NIR) spectrometer (MPA spectrometer, Bruker Co., Germany) in diffuse reflectance mode fitted with a tungsten lamp and a cooled PbS detector. Spectra were collected at resolution 16 cm^{-1} over the range of 12 000-4000 cm^{-1} (833-2500 nm, total of 1037 wavelength points) at the room temperature (around 25°C). All the spectra were recorded as log (1/*R*) with respect to a golden reference standard. The software OPUS 6.5 (Bruker Co., Germany) was available to modify spectrometer set-up and store the spectral data.

A circular sample cup (Φ50 mm) with a quartz window was used when collecting spectra of bulk kernels. About 200 kernels of seeds were placed in sample cup. When the sample cup rotates, the average of 64 successive scans will be stored as the spectrum. Two spectra sets (A1 and A2) were collected from bulk samples. Set A1 contains 48 varieties (50 kernels per variety, Supplementary Table 1). Set A2 contains 10 varieties with producing areas distributed in five provinces and producing years distributed in four years (Supplementary Table 2).

When collecting spectra in single kernel mode, a grain of seed was placed in the sample hole (Φ10 mm) using tweezers, with the kernel germ facing the light source and detector, and then the hole was covered by a gold-plated lid[17]. Each spectrum is the average of 20 scanned interferograms. Spectra sets B1 and B2 were obtained in single kernel mode. Set B1 contains 42 varieties (75 kernels per variety, Supplementary Table 3). Set B2 contains 2 varieties, with producing areas distributed in three provinces and producing years distributed in three years (Supplementary Table 4).

2.2 Data preprocessing and feature extraction

In this part, smoothing was performed with a moving window of 9 data points to eliminate the noise before the first-derivative transformation (differential width is 9 data points). Furthermore, vector normalization was used to eliminate the random error caused

by sample placement and instrument state[18].

Finally, principal component analysis (PCA) and linear discriminant analysis (LDA) were carried out to extract spectra feature. PCA was employed to eliminate collinearity in the spectral data, and then significant components were selected. LDA focuses on finding optimal boundaries between classes. The calculations in this work were carried out in Matlab 7.13 (Mathworks, USA).

2.3 Establishment of the variety identification models

In this paper, the data in Set A1, Set B1 and Set B2 was used to explore the feasibility of building models for large numbers of maize varieties. The BPR model of one class was established by constructing a subspace with a closed surface to cover the sample points based on their distribution.

When establishing identification model of one variety, half of the samples in one variety were randomly selected as the training set, and the remaining half of the samples in this variety were used to examine whether the model can recognize the samples from the same variety. The percentage of the samples that were correctly recognized was calculated as CAR. All samples of other varieties were used to examine whether the model can reject the samples not belonging to a particular variety. The percentage of the samples correctly rejected was calculated as CRR.

In conventional identification methods, the identification result for one sample is simply yes or no, through which the reliability and the performance of models cannot be quantified. In this study, there were some other parameters defined to solve this problem, such as CAD, CRD and CD, and the definitions are given as follows.

Given that D_1, D_2 present the distances between samples S_a, S_b and model unit, S_a represents samples that belonging to the model while S_b is not. R is the radius of the model (Figure 1).

$$\mathrm{CAD}=(R-D_1)/R \quad (1)$$

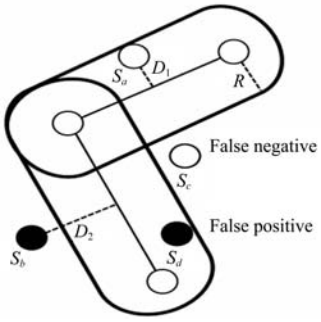


Figure 1 Schematic diagram of BPR model

If 0≤CAD≤1, S_a is correctly accepted by model, the lager the CAD, the more certain that S_a is belonging to the model. If CAD<0, S_a is wrongly rejected by model, called false negative sample (S_c in Figure 1).

$$\mathrm{CRD}=(D_2-R)/D_2 \quad (2)$$

If 0<CRD≤1, S_b is correctly rejected by model, the lager the CRD, the more certain that S_b is not belonging to the model. If CRD≤0, S_b is wrongly accepted by the model, called false positive sample (S_d in Figure 1).

$$\mathrm{CD}=(\mathrm{CAD}+\mathrm{CRD})/2 \quad (3)$$

where, CD<1, CD is a parameter that takes both CAD and CRD into account, the lager the value of CD, the better the performance of model. CD is also used to get an appropriate radius of the model. As the increase of model radius, CAD increases while CRD decreases, the radius which gets the highest CD value is chosen as the radius of the model (Figure 2).

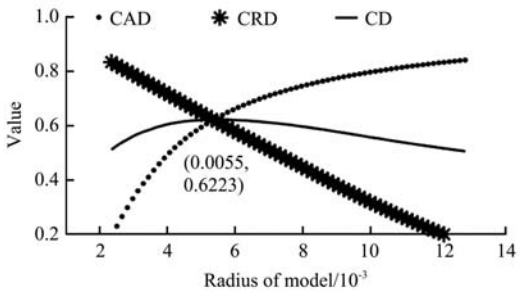


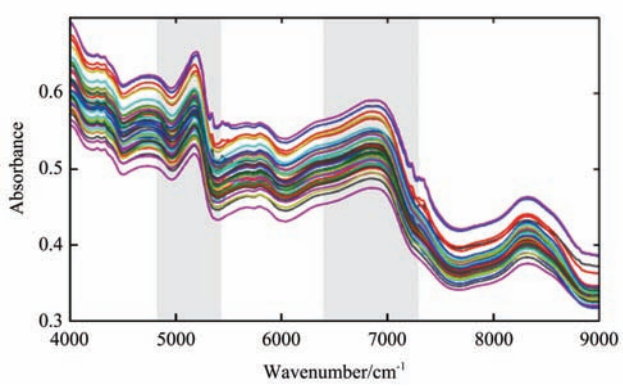
Figure 2 Relationship between model radius and CD, CAD and CRD

3 Results and discussion

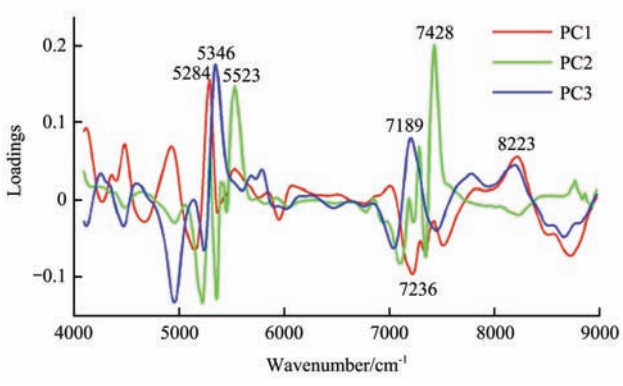
3.1 NIR spectra analysis

The spectra region 4000-9000 cm^{-1} (1111-2500 nm, total of 649 wavelength points) was primarily chosen because it is an important band for maize seeds. Additionally, in this region, signal-to-noise ratio is relatively high. The average spectra of 42 varieties in Set B1 are similar in shape and have absorption bands around 4700 cm^{-1}, 5200 cm^{-1}, 6000 cm^{-1}, 7500 cm^{-1} and 8000 cm^{-1} (Figure 3a). These bands are mainly related to O–H or C–H functional vibrations and overtones of sugars (6944 cm^{-1}), protein (4878 cm^{-1}, 5051 cm^{-1}) and starch (4440 cm^{-1}, 4762 cm^{-1}, 5000 cm^{-1}, 5263 cm^{-1}, 6329 cm^{-1}, 6494 cm^{-1}, 6545 cm^{-1}), etc[4].

To investigate the basis for the spectral discrimination between varieties, spectra data in Set B1 was processed by PCA and the PCA eigenvectors were analyzed (Figure 3b). PC1, PC2 and PC3 account for 90.3% of the NIRS data variance, the characteristic regions of eigenvector curves of the PCs are coincident with the spectra in Figure 3(a). It demonstrates that these regions contain much information about variety differences.



a



b

Figure 3 Average spectra of the 42 varieties (a) and the first three principal component curves (b) in Set B1

3.2 BPR model

There is no possibility to establish models for all varieties in market at one time. Therefore, identification models should be able to reject varieties which haven't participated in building models.

There is a great difference between spectra in Set A1 and Set B1 because they were collected in different modes. Therefore, the identification models of Set A1 and Set B1 were established separately. Models were built for 40 varieties (2000 spectra in total) in Set A1, the 50 spectra of each variety were divided equally into training set and validation set. The remaining 8 varieties were taken as the second validation set (400 spectra in total) to test the rejection ability of models. Similarly, models of the 40 varieties (3000 spectra in total) in Set B1 were established with 38 samples from each variety in the training set and 37 in the validation set. The remaining 2 varieties in Set B1 and all samples in Set B2 were taken as the second validation set (390 spectra in total).

3.2.1 Optimization of data dimensions

PCA was applied on pretreated spectra in order to reduce the dimensionality of the spectral data while retaining as much information as possible. Selecting suitable number of principal components (PCs) is difficult because too many PCs will include undesired instrument noises, and too few PCs will lose important spectral information[19].

The minimum number of principle components was selected to maintain more than 99% of the variance in training samples. Furthermore, LDA was used to reduce data dimensions. The dimension of data was chosen according to the accuracy. The first 50 PCs contain 99.9% variance in training samples of Set A1 (Figure 4a). Models based on the first 10 dimensions of data treated by LDA got the highest correct rate (Figure 4b), and the accuracy did not change as the number of dimensions increased. For models in Set B1, the first 57 PCs (Figure 4c) and the first 17 dimensions of data treated by LDA were chosen to established models. The accuracy decreased as the data dimension increased after its peak (Figure 4d). Therefore, it's efficient to optimize the data dimension by accuracy when building identification models.

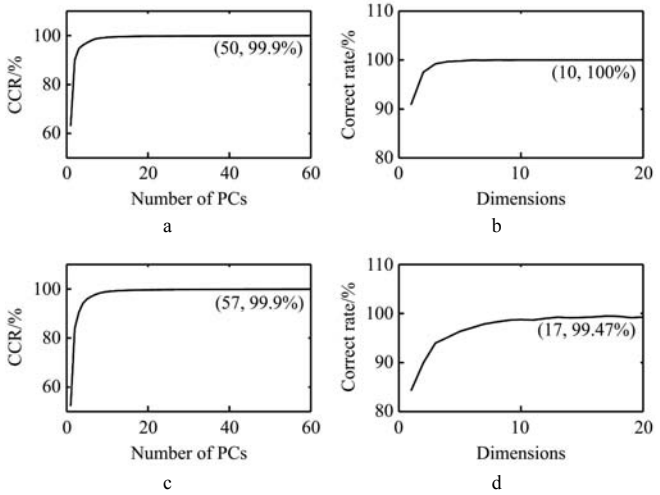


Figure 4 Cumulative contribution rate (CCR) curve of first 60 PCs in Set A1 (a) and Set B1 (c) and relationship between correct rate and data dimensions of Set A1 (b) and Set B1 (d)

3.2.2 Identification results of BPR model

A variety model should reject seed samples of other varieties as completely as possible. The average CAR and CRR of validation set in Set A1 are 99.90% and 99.99%. For models in Set B1, the average CAR and CRR of the validation set reach 94.60% and 99.89% (Table 1). As a whole, identification models based on BPR can reject more than 99% of samples belonging to other varieties successfully, and correctly recognize more than 90% of samples belonging to models.

The second validation set was used to test the rejection ability of models. The average CRR and CRD in identification results of 40 models in Set A1 are 99.21% and 0.523 (Table 2). It reveals that models in Set A1 can well reject samples of varieties which have not participated in building models. Rejection ability of the 40 models in Set B1 is comparatively lower, the average CRR and CRD are 97.39% and 0.3290 (Table 3). The results show that the range of CRD values of these false recognition samples is 0.013-0.221, which indicates that these false positive samples distribute at the edge of the model subspace. These samples can be rejected by adjusting the radius of models.

Table 1 Model performance in four data sets

		Set A1	Set A2	Set B1	B1+B2
Model numbers		40	10	40	44
Training set	CAR/%	100	99.60	99.87	99.28
	CRR/%	99.66	99.37	97.30	97.23
	CAD	0.587	0.481	0.448	0.489
	CRD	0.587	0.477	0.449	0.477
Validation set	CAR/%	99.90	98.23	94.60	90.33
	CRR/%	99.99	99.93	99.89	99.80
	CAD	0.575	0.443	0.431	0.480
	CRD	0.587	0.471	0.448	0.481

Table 2 Rejection performance of 40 models in Set A1

Model ID	Correct rejected sample		False positive sample	
	CRD	CRR/%	Number	CRD
1	0.586	99.75	1	-0.057
3	0.470	97.75	9	-0.081
11	0.512	98.75	8	-0.084
19	0.416	99.25	3	-0.055
32	0.571	99.75	1	-0.078
The other 35	0.572	100	0	-
Average	0.523	99.21	-	-

Table 3 Rejection performance of 40 models in Set B1

Model ID	Correct rejected sample		False positive sample	
	CRD	CRR/%	Numbers	CRD
1	0.375	99.5	2	-0.013
2	0.4	99.74	1	-0.221
4	0.3949	99.74	1	-0.218
7	0.197	91.28	34	-0.126
10	0.3	99.74	1	-0.071
12	0.215	90	39	-0.128
13	0.306	99.74	1	-0.025
15	0.296	96.92	12	-0.097
18	0.341	99.5	2	-0.06
20	0.379	98.97	4	-0.105
23	0.367	99.5	2	-0.046
26	0.243	89.5	41	-0.134
28	0.387	99.74	1	-0.095
32	0.363	97.44	10	-0.12
33	0.335	99.23	3	-0.117
35	0.336	99.74	1	-0.049
37	0.221	90.51	37	-0.103
40	0.301	99.74	1	-0.076
The other 22	0.494	100	0	-
Average	0.3290	97.39	-	-

In short, both models in Set A1 and Set B1 get an acceptable performance. These results demonstrate that it is feasible to establish variety identification models based on NIR and BPR. The performance of models in Set A1 is better than models in Set B1, because spectra in Set A1 were collected from bulk kernels, which can avoid the randomness of single kernel mode and get more information about samples.

The sample stage rotates when collecting spectra of bulk kernels, therefore the spectra contains different information from different parts of many seeds, and can better reflect the characteristics of a variety. The seed is static with the germ facing the light source and detector while measuring spectra of single kernel. As a consequence, only the information on the germ can be acquired in single kernel mode. The kernel shape and size also have a great influence on the spectra. Spectra of bulk kernels contain more information, and the consistency of bulk kernels spectra is higher than that of single kernel spectra, therefore variety identification models based on bulk kernels perform better.

Some researchers have proposed some strategies to overcome drawbacks of measuring spectra of single kernel. They reported that scanning the seeds on movement yields better results than static scan[20-22], probably because scattering effects are minimized during the seed movement, reducing the variability due to kernel shape and size[16]. A better result can be expected if the spectrum of a single kernel is measured in motion in the future study.

3.3 Model robustness

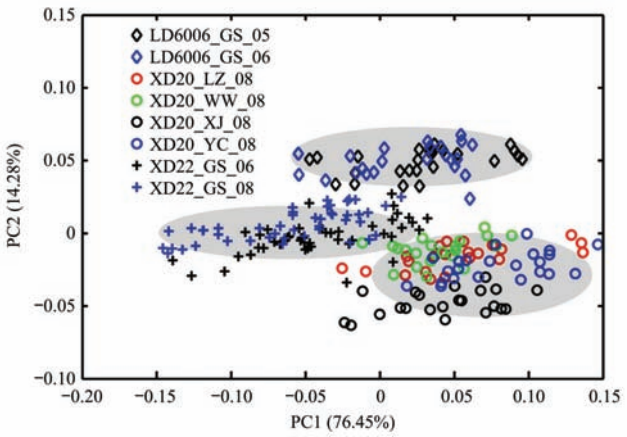
Different growth conditions such as different producing areas and years lead to differences in seed component and appearance, which have a significant influence on the spectra, even though the samples are of the same variety.

The spectral data in Set A2 and Set B2 was applied to discuss whether differences between different varieties (D-variety) are larger than differences within varieties (differences between samples of the same variety from different producing areas (D-area) or years (D-year)). D-variety, D-area and D-year were obtained by calculating the Euclidean distance between the centroids (the average coordinates) of each variety. The ratios of D-variety and D-area and D-year were calculated to quantize the potential of overcoming influences of producing areas and producing years.

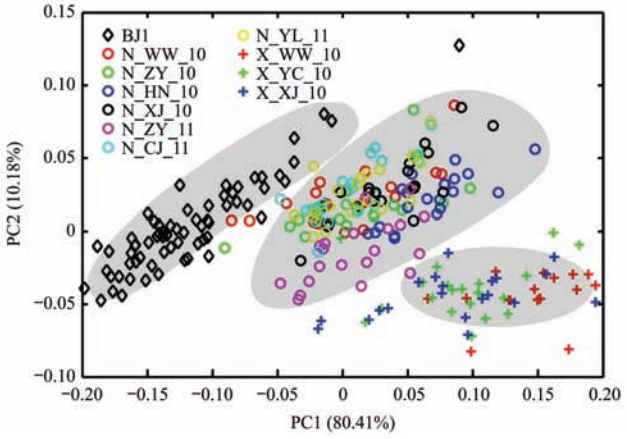
PCA was applied to the pretreated spectral data of three varieties (Ludan6006 (LD6006), Xundan20 (XD20) and Xundan22 (XD22)) in Set A2 (250 samples in total). The first two PCs cluster all the samples into three groups according to variety (Figure 5(A); 90.73% data variance described). LD6006 contains two samples which were harvested in 2005 and 2006 in Gansu province (LD6006_GS_05, LD6006_GS_06). XD22 contains two samples which were harvested in 2006 and 2008 in Gansu province (XD22_GS_08, XD22_GS_06). XD20 contains four samples which were harvested in four different areas in 2008 (XD20_LZ_08, XD20_WW_08, XD20_XJ_08, XD20_YC_08). The difference between different samples of one variety is relatively smaller than that between varieties. To investigate it more carefully, the first nine PCs (cumulative variance contribution rate 99.04%) were selected to calculate distances. The ratio between D-variety and D-year is 4.16, and the ratio between D-variety and D-area is 2.36. Since D-variety is larger than D-year and D-area, spectra collected from bulk samples can overcome the influences of producing area and producing year.

A similar conclusion can be drawn for spectral data collected in single kernel mode. The first two PCs account for 90.6% of variance of spectra in three varieties (BJ1, Nonghua101 (N), and Xundan20 (X)). BJ1, N and X were selected from Set B1 and Set B2 (275 samples in total). N contains six samples which were harvested from five regions (Wuwei (WW), Hainan (HN), Changji (CJ), Zhangye (ZY), Yili (YL)) and two years (2010 (10), 2011 (11)). X contains three kinds of samples which were harvested from Wuwei (WW), Yinchuan (YC), and Xinjiang (XJ) in 2010. All the samples cluster into three groups according to varieties (Figure 5b). Pretreated spectral data were analyzed by PCA and the first nine principle components (cumulative contribution rate 99.10%) were selected to calculate distances. The ratio between D-variety and D-year is 5.56, and the ratio between D-variety and D-area is 4.01, therefore D-variety is larger than D-year and D-area. The results show that spectra collected in single kernel mode can also overcome the influences of cultivation area and harvest year.

Differences between samples harvested from different areas and years do exist (Figure 5), therefore it is not appropriate to use only one kind of samples to establish identification models of one variety. Representative samples from different areas and years should be chosen as training samples. Taking Nonghua11 (N) and Xundan20 (X) as an example, both of them consisted of samples from three areas in 2010, and those samples which can represent the distributions of all samples were selected to establish models (Figure 6). The CAR and CRR of 10 models in Set A2 are 98.23% and 99.93%, and CAR and CRR of 44 models in Set B1 and Set B2 are 90.33% and 99.80% (Table 1). These results reveal that identification models based on BPR and NIR are robust to complicated samples from different areas and years.



a. Spectra collected in bulk-kernel mode from three varieties (Ludan6006 (LD6006), Xundan20 (XD20) and Xundan22 (XD22)) in Set A2



b. Spectra collected in single-kernel mode from three varieties (BJ1, Nonghua101 (N), and Xundan20 (X)) in Set B1 and B2)

Figure 5 Coordinates of the first two principal components of spectra from maize seeds harvested in different areas and years

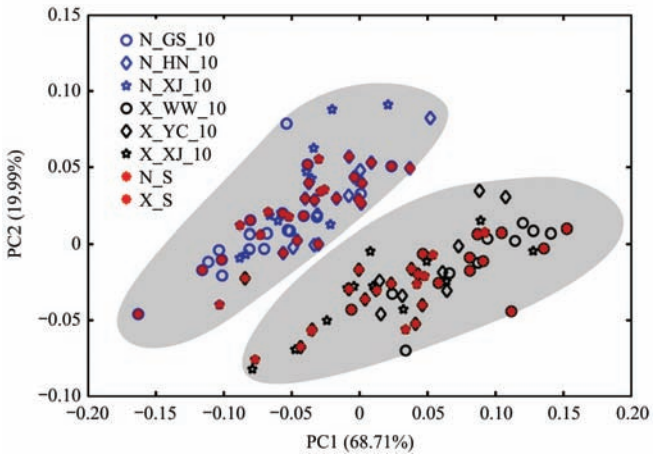


Figure 6 The most representative samples (marked by red) to establish identification models.

4 Conclusions

This paper studied the feasibility of identifying authenticity of a large number of maize seed varieties via combining NIRS with chemometrics methods. The spectra of seeds were measured in two modes, bulk kernels mode and single kernel mode. The former can obtain signals of higher quality than the latter, but cannot discriminate single kernel. The spectra of single kernel are influenced by seed shape and size. The 40 identification models based on spectra of bulk kernels obtained an average accuracy above 99%, and correctly rejected the other 8 varieties (400 samples) which have not participated in developing models with an average accuracy of 99.21%. The average identification and rejection accuracy of 40 models established based on the spectra of single kernel are 97.2% and 97.4%, respectively.

The producing areas and producing years have a significant influence on spectra of samples of the same variety, which leads to a serious decline in the variety identification rate. The spectral difference between different samples of the same variety but from different producing areas and years is relatively smaller than that between varieties. An effective method to improve the robustness of identification models is to contain spectra of seeds from different producing areas and years in the training set. Models based on spectra of bulk kernels (10 varieties in dataset A2, seeds of one variety come from different areas or years) achieved an average accuracy of 99%. And the 44 single kernel models (dataset B1 and B2, the two varieties in B2 come from different areas and years) achieved an accuracy of 95%. These results show that models based on samples from different producing areas and producing years are robust.

These promising results indicate that it is feasible to identify large amounts of maize varieties by NIRS and chemometrics methods. This cheap and non-destructive method can obtain identification result in several minutes, and has a great potential for application in authenticity and traceability of cereals.

Acknowledgements

The authors gratefully acknowledge that this work was financially supported by the National Key Scientific Instruments and Equipment Development Project (2014YQ470377), National Special Fund for Agro-scientific Research in Public Interest (Grant No. 201203052), Science and Technology Project of Beijing (Grant No. D131100000413002) and China Agricultural University Education Foundation Dabeinong Education Funds (1081-2413001).

The maize seed samples were provided by Shanxi Tunyu Seed Industry Co., Ltd., Beijing Dabeinong Technology Group Co., and Beijing Kings Nower Seed S&T CO., LTD.

Supplementary tables

Supplementary Table 1 48 varieties in Set A1 (50 samples per variety)

Serial number	Name	Serial number	Name	Serial number	Name
1	CH3018	17	JNN7	33	NH0709
2	DF26	18	JYD42102	34	NH13
3	DH605	19	L1224	35	NK718
4	DH662	20	L303	36	SS50904
5	DT6508	21	L503	37	W8105
6	DY405	22	LD527	38	XDY48
7	DY406	23	LD53	39	ZD528
8	DY605	24	LD565	40	ZB138
9	FN2146	25	LTN1	41	ZB818
10	FR721	26	LTN3	42	ZD1001
11	FY469	27	MC105	43	ZD5118
12	G0810	28	MC112	44	ZD6016
13	GD592	29	MC115	45	ZHN2
14	J257	30	N905_906	46	ZS516
15	JD618	31	ND451B	47	W09_6981
16	JK528	32	ND668	48	ZYF815

Supplementary Table 2 The samples of 10 varieties in Set A2

Variety	Area	Year	Grade	Sample number	
				Per variety	Total
Xundan20	Linze	2008	/	25	279
Xundan20	Wuwei	2008	/	25	
Xundan20	Xinjiang	2008	/	25	
Xundan20	Yinchuan	2008	/	25	
Xundan20	Linze	2009	/	50	
Xundan20	Wuwei	2009	/	50	
Xundan20	Xinjiang	2009	/	29	
Xundan20	Yinchuan	2009	/	50	
Jingyu16	Xinjiang	2009	Large	50	100
Jingyu16	Xinjiang	2009	Small	50	
Tunyu88	Xinjiang	2006	/	50	100
Tunyu88	Xinjiang	2007	/	50	
Ludan6006	Linze	2005	/	25	50
Ludan6006	Wuwei	2006	/	25	
Nongda84	Linze	2008	/	25	50
Nongda84	Wuwei	2008	/	25	
Nonghua101	Wuwei	2009	Large	50	150
Nonghua101	Wuwei	2009	Small	50	
Nonghua101	Hainan	2009	/	50	
Xundan22	Wuwei	2006	/	50	100
Xundan22	Wuwei	2008	/	50	
Zhongdan808	Neimeng	2006	/	50	50
Nonghua16	Xinjiang	2007	/	50	50
Nonghua8	Wuwei	2007	/	50	50

Supplementary Table 3 42 varieties in Set B1 (75 samples per variety).

Serial number	Name	Serial number	Name	Serial number	Name
1	BJ1	15	BJ1313	29	BJ52
2	BJ1079	16	BJ1910	30	BJ53
3	BJ1080	17	BJ193	31	BJ591
4	BJ1112	18	BJ2	32	BJ6
5	BJ1116	19	BJ205	33	BJ7
6	BJ1125	20	BJ25	34	BJC11-1
7	BJ1130	21	BJ260	35	BJC11-19
8	BJ1133	22	BJ267	36	BJC11-5
9	BJ1134	23	BJ28	37	HNC12-4
10	BJ1135	24	BJ3	38	HNC12-6
11	BJ1136	25	BJ30	39	HNC12-8
12	BJ1211	26	BJ31	40	HNC12-9
13	BJ1226	27	BJ4	41	BJ29
14	BJ1243	28	BJ5	42	BJ32

Supplementary Table 4 The samples of two varieties in Set B2

Variety	Area	Year	Sample number	
			Per variety	Total
Nonghua101	Hainan	2009	20	160
Nonghua101	Hainan	2010	20	
Nonghua101	Xinjiang	2010	20	
Nonghua101	Wuwei	2010	20	
Nonghua101	Zhangye	2010	20	
Nonghua101	Zhangye	2011	20	
Nonghua101	Changji	2011	20	
Nonghua101	Yili	2011	20	
Xundan20	Linze	2009	20	80
Xundan20	Wuwei	2010	20	
Xundan20	Yinchuan	2010	20	
Xundan20	Xinjiang	2010	20	

[References]

[1] He K Q, Cheng X X. The analysis of esterase isozyme in different maize species. Chinese Agricultural Science Bulletin, 2008; 4(24): 221. (in Chinese)

[2] Sharopova N, McMullen M D, Schultz L, Schroeder S, Sanchez-Villeda H, Gardiner J, et al. Development and mapping of SSR markers for maize. Plant Mol. Biol., 2002; 48: 463.

[3] Zhao J R, Sun S X, Wang F G. Research trends in China maize variety identification by DNA fingerprinting. Beijing: China Agricultural Science and Technology Press, 2008.

[4] Yan Y L, Zhao L L, Han D H, Yang S M. Basics and application of near infrared spectral analysis. Beijing: China Light Industry Press, 2005; 38p.

[5] Zhou J, Cheng H, Ye Y, Wang L Y, He W, Liu X, et al. Recognition for raw material cultivar of manufactured tea with fisher discriminant classification with principal components analysis. Acta Optica Sinica, 2009; 29(4): 1117–1120. (in Chinese)

[6] Liang L, Liu Z X, Yang M H, Zhang Y X, Wang C H. Discrimination of variety and authenticity for rice based on visual/near infrared reflection spectra. Journal of Infrared and Millimeter Waves, 2009; 28(5): 353–356. (in Chinese)

[7] Cao F, Wu D, He Y, Bao Y D. Variety discrimination of grapes based on visible-near reflection infrared spectroscopy. Acta Optica Sinica, 2009; 29(2): 537–540. (in Chinese)

[8] Zhuang X G, Wang L L, Wu X Y, Fang J X. Origin identification of Shandong green tea by moving window back propagation artificial neural network based on near infrared spectroscopy. Journal of Infrared and Millimeter Waves, 2016; 35(2): 200–204.

[9] Han Z Z, Wan J H, Zhang H S, Deng L M, Du H W, Yang J Z. Variety and origin identification of maize based on near infrared spectrum. Journal of the Chinese Cereals and Oils Association, 2014; 29(1): 21–24. (in Chinese)

[10] Liu X. Research of automatic maize seed purity sorting system. Beijing: China Agricultural University, 20115; pp.0–13. (in Chinese)

[11] Zhao S Y. Study on the influence of different producing places and years for the maize seeds purity sorting. Beijing: China Agricultural University, 2016; pp.17–21. (in Chinese)

[12] Zhao H Y, Guo B L, Wei Y M, Zhang B. Near infrared reflectance spectroscopy for determination of the geographical origin of wheat. Food Chemistry, 2013; 138: 1902–1907.

[13] Wang S J, Wang B N. Analysis and theory of high-dimension space geometry for artificial neural networks. Acta Electronica Sinica, 2002; 30: 1417. (in Chinese)

[14] Wang S J, Liu Y Y, Lai J L, Liu X X. Biomimetic pattern recognition and multi-weighted neuron. Beijing: National Defence Industry Press, 2013; 156p.

[15] Guo T T. Study on the cultivar discrimination method for maize seeds based on near infrared spectroscopy and biomimetic pattern recognition. Beijing: Institute of Semiconductors, Chinese Academy of Sciences, 2010. (in Chinese)

[16] Agelet L E, Hurburgh C R. Limitations and current applications of near infrared spectroscopy for single seed analysis. Talanta, 2014; 121: 288.

[17] Jia S Q, Guo T T, Tang X T, Si G, Yan Y L, An D. Study on spectral measurement methods in identification of maize variety authenticity based on near-infrared spectra of single kernels. Spectroscopy and Spectra Analysis, 2012; 32: 103. (in Chinese)

[18] Guo T T, Wu W J, Su Q, Wang S J, An D. Effects of spectral pretreatment and wavelength selection on discrimination of maize seed varieties by NIR spectroscopy. Transactions of the CSAM. 2009; 40(Supp l): 87. (in Chinese)

[19] Cocchi M, Corbellini M, Foca G, Lucisano M, Pagani MA, Tassi L, et al. Classification of bread wheat flours in different quality categories by a wavelet-based feature selection/classification algorithm on NIR spectra. Anal. Chim. Acta, 2005; 544: 100.

[20] Janni J. Pioneer Hi-Bred International Inc. Patent 7274457 B2, 2007.

[21] Janni J, Weinstock B A, Hagen L, Wright S. Novel near-infrared sampling apparatus for single kernel analysis of oil content in maize. Appl. Spectrosc, 2008; 62(4): 423.

[22] Armstrong P R. Rapid single-kernel NIR measurement of grain and oil-seed attributes. Appl. Eng. Agric, 2006; 22(5): 767.