

中图分类号: S532 文献标识码: A 文章编号: 1672-3635(2021)06-0507-07

DOI: 10.19918/j.cnki.1672-3635.2021.06.004

基于近红外光谱技术马铃薯蛋白质含量定标模型的构建

李 赞^{1,2}, 高红秀², 金 萍², 石 瑛^{2*}

(1. 江苏农林职业技术学院, 江苏 镇江 212400; 2. 东北农业大学农学院, 黑龙江 哈尔滨 150030)

摘 要: 马铃薯是世界重要的粮食作物之一。为了提高马铃薯品质育种效率, 研究选用蛋白质含量差异较大的多年份、多品种、多地点种植的411份马铃薯块茎为材料, 使用InfraXact Lab型近红外光谱分析仪分别对其进行光谱扫描, 并用国标化学法对其蛋白质含量进行测定。通过近红外定标软件(WinISI III), 采用一阶导数的数学处理(1, 4, 4, 1)、标准正态变换和去趋势(SNV+ Detrend)最佳组合的预处理方法, 结合改进偏最小二乘法(Modified partial least squares, MPLS)回归技术构建了马铃薯块茎蛋白质含量的近红外定标模型, 并对预测模型的准确性进行了验证。其定标标准偏差(Standard error of calibration, SEC)、交叉检验标准误差(Standard error of cross-validation, SECV)和交叉验证相关系数(1 minus the variance ratio, 1-VR)分别为0.566、0.632和0.912。97个验证集样品用于外部检验马铃薯块茎中的蛋白质含量的模型预测值与化学值之间的相关系数为0.931, 表明利用近红外光谱分析技术可以测定马铃薯块茎中蛋白质含量。

关键词: 马铃薯; 蛋白质; 近红外光谱法; 定标模型

Establishment of Potato Protein Content Calibration Model Based on Near Infrared Spectroscopy

LI Zan^{1,2}, GAO Hongxiu², JIN Ping², SHI Ying^{2*}

(1. Jiangsu Vocational College of Agriculture and Forestry, Zhenjiang, Jiangsu 212400, China;

2. College of Agronomy, Northeast Agricultural University, Harbin, Heilongjiang 150030, China)

Abstract: Potato is one of the most important food crops in the world. In order to improve the efficiency of potato quality breeding, 411 potato tuber samples of various varieties planted in some years and places with large differences in protein content were selected as materials, and were scanned by InfraXact near infrared spectroscopy (NIRS) analyzer and their protein contents were also determined by national standard chemical method. The near infrared calibration model of potato tuber protein content was established by using the near-infrared calibration software (WinISI III), the preprocessing method of the best combination of mathematical treatment of the first derivative (1,4,4,1), standard normal variate and detrend (SNV + Detrend) transformation, combined with the modified partial least squares (MPLS) regression technology, and the accuracy of the prediction model was verified. The standard error of calibration (SEC), standard error of cross-validation (SECV) and 1 minus the variance ratio (1-VR) of the calibration was 0.566, 0.632 and 0.912, respectively. A ninety-seven set of samples were selected for external validation, and the correlation coefficient between model prediction and the chemistry value for protein content of potato tubers was 0.931, suggesting that the

收稿日期: 2021-12-23

基金项目: 国家重点研发计划项目(2017YFD0101906)。

作者简介: 李赞(1985-), 女, 硕士, 高级农艺师, 主要从事作物遗传育种研究。

*通信作者(Corresponding author): 石瑛, 副研究员, 主要从事马铃薯遗传育种及栽培研究, E-mail: yshi@neau.edu.cn。

near infrared spectral analysis technology could be used to determine the protein contents in potato tubers.

Key Words: potato; protein; near infrared reflectance spectroscopy; calibration model

马铃薯在全世界范围内一直是重要的粮食作物之一。随着人们生活水平的不断提升,对马铃薯的品质要求也逐渐提高,蛋白质含量是研究马铃薯品质中重要的指标之一。传统测试马铃薯块茎蛋白质含量的方法一般采用凯氏定氮法,其作为一种国际通用的测试方法,技术比较成熟且标准化,但应用过程中存在一些明显不足。该方法需要对检测样品进行前处理,操作过程消耗时间长(至少需要2 h才能完成),工作量较大,费工费时,且所使用试剂具有强烈的腐蚀性,对环境污染大。因此,需要找到一种快速和准确检测马铃薯蛋白质含量的有效方法,为马铃薯蛋白质含量的评价提供依据。

采用近红外方法测定蛋白质含量可以大幅度减少操作过程中消耗的各类材料,去掉了很多繁琐且有危险的工作程序,降低了有害气体的污染和对操作人员的伤害,大幅提升测试分析的工作效率,并显著降低测试成本。近红外光谱分析技术实际上是一个二级分析方法,在对未知样品进行分析之前,必须选择一组具有代表性的样品作为一个定标集,对应其中的每个样品测量光谱及对应的组分或性质,并采用多元校正的方法将已测量的光谱与对应的性质或组成数据关联,建立该组分的定标模型,然后进行未知样品光谱数据的采集,并将其与校正模型相互对应,计算出此未知样品的组分。因此,近红外光谱分析技术的广泛应用,最重要的是建立定标模型,定标模型的合理性直接影响检测结果的准确度和稳定性。

近红外光谱分析技术在测定农副产品(如饲料、谷物、肉、蛋、奶、水果和蔬菜)的品质(包括蛋白质、脂肪、纤维、灰分、氨基酸等)方面得到广泛使用^[1-7],已成为粮食品质分析的重要手段。近年来,已有不少作物利用近红外技术测定蛋白质含量。研究者利用近红外光谱技术分别对小麦^[8]、水稻^[9]、大豆^[10]、玉米^[11]和食用向日葵子仁^[12]中的蛋白质含量进行了定标模型的建立,并取得了良好的效果。由此可见,利用近红外技术对谷物品质建

立定标模型的方法已经相当成熟,并已有相关报道近红外技术应用于马铃薯研究中^[13,14],但利用近红外技术在马铃薯蛋白质含量分析测试方法或建立相关模型方面的报道还很少。本研究以不同年份、不同区域马铃薯块茎为试验材料,进行马铃薯块茎蛋白质含量近红外光谱定标模型的建立,并对其准确性进行验证,以期提供一种准确、快速、无需预处理测定马铃薯块茎中蛋白质含量的分析方法,为以后的相关研究与应用提供理论指导。

1 材料与方法

1.1 试验材料

供试材料为东北农业大学马铃薯育种基地的无性系种质材料,马铃薯品种审定区域试验、生产试验的品种(系),来自于克山、讷河等地的马铃薯品种(系)。样品数目为986份,按不同年份、地域、生长季节等条件收集有代表性的马铃薯样品,将新鲜样品切碎,放在105℃的恒温箱内杀青30 min,然后将温度调到70.5℃,继续烘干14~16 h,使样本材料至恒重,将样品粉碎至30~40目,装于小塑料袋中密封保存。

1.2 试验方法

1.2.1 样品扫描

采用福斯公司生产的近红外分析仪Infraxact对样品进行光谱扫描,扫描前,先将光谱仪开机预热1 h。波长范围570~1 850 nm,每份样品重复3次。然后使用WinISI III软件对光谱进行平均,生成平均光谱文件。

1.2.2 蛋白质含量化学值测定

马铃薯蛋白质含量化学测定方法为凯氏定氮法(GB 5009.5—2016)^[15],使用瑞典福斯公司的2300型全自动凯氏定氮仪进行测定。每个样品采用双平行分析,测定结果取平均值。

1.2.3 剔除超常和过剩样品,确定定标样品集

采用主成分分析技术(聚类分析技术)将光谱数据进行压缩,并分解为主成分和得分矩阵数

据。然后利用得分矩阵数据, 比较各样品光谱间的差异, 以及某样品与主组群样品组间的差异, 以此确定相似样品及超常样品, 从而可确定参与定标的最好样品。首先, 利用光谱文件创建得分文件, 计算出数据的平均值和每一个样品到平均值的距离。边界是数据集的3倍标准偏差。然后从剔除超常样品后的光谱文件中选择代表性样品, 即剔除过剩样品。过剩样品剔除限是0.6, 0.6的定义为以某一个样品为中心, 在半径为0.6以内的样品将被认为是与此样品相似, 其光谱的性质则不能增加定标集样品的变异范围, 即作为过剩样品, 不可参加定标样品集。

1.2.4 定标模型的建立

首先将测定的样品化学值输入到定标集的光谱文件中, 使每个样品的近红外光谱与化学值一一对应, 然后用软件中的改进最小二乘法(Modified partial least squares, MPLS)回归技术法建立马铃薯蛋白质含量的近红外分析模型, 预处理方法None(无散射处理)和SNV + Detrend(标准正常化 + 散射处理), 导数处理参数选择分别为0.0.1.1、1.4.4.1。观察统计数据列交叉验证相关系数(1

minus the variance ratio, 1-VR)和交叉验证误差(Standard error of cross-validation, SECV), 找出1-VR值最大, 而SECV值最小的即为最佳定标模型, 这两组数据基本能反应定标模型对其他未知样品的预测性能。

1.2.5 定标模型的验证

在定标方程建立后, 以一组没有参与定标的样品集作为验证集, 对该方程的预测性能进行验证。验证样品集样品应具有代表性, 其成分应覆盖在一定范围, 并且其传统实验室参考数据须准确可靠, 才能保证验证结果的合理性^[16]。定标验证工作是通过WinISI软件的Monitor Program程序进行的, 其验证结果表征为实验室数据和近红外预测数据相互比较所计算出的一系列统计结果。

2 结果与分析

2.1 马铃薯测试样品的近红外光谱图

在收集样品光谱后, 首先观察每一样品的吸收图谱, 对于异常图谱要重新进行扫描或作剔除处理。样品的近红外光谱图如图1所示, 马铃薯近红外光谱图有明显的吸收峰。

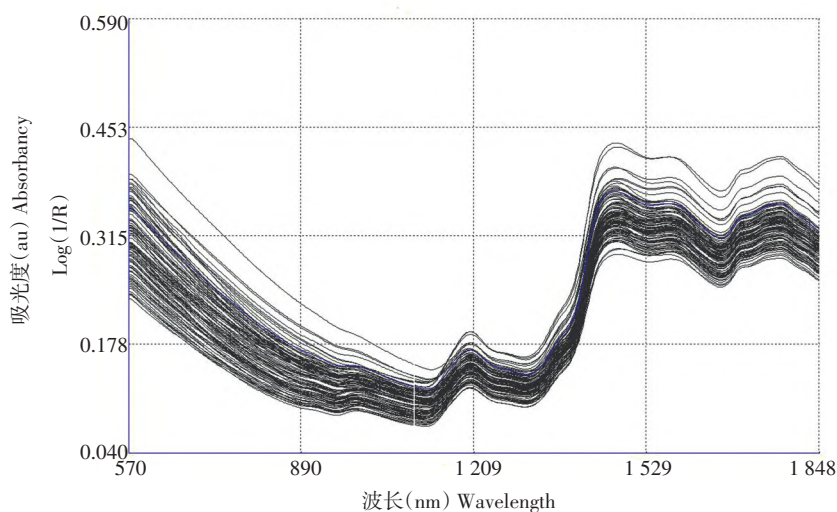


图1 马铃薯测试样品近红外光谱图

Figure 1 Near infrared spectrogram of potato samples

2.2 确定定标样品集

采用主成分分析PCA法, 根据马氏距离或相

关性去除超常样品和过剩样品, 超常样品剔除限是3.0, 过剩样品剔除限是0.6, 最终确定定标样

品集为411份。样品蛋白质含量分布见图2, 样品化学含量的梯度分布较均匀, 基本覆盖了马铃薯的化学指标含量范围, 有较好的代表性, 满足建标的需要。本试验随机选取100份样品组成验证集, 其余311份样品自动生成定标集, 其中最小值为6.77, 最大值为23.21。定标样品集及验证集蛋白质含量的分布见表1。

2.3 定标模型的建立

将测定的样品化学值输入到定标集的光谱文件中, 用MPLS法建立蛋白质含量的近红外分析模型, 预处理方法分别为无散射处理(None)和去

散射处理(SNV + Detrend), 导数处理参数选择分别为0.0.1.1、1.4.4.1。观察统计数据列1-VR和SECV, 找出1-VR值最大的, 而SECV值最小的即为最佳定标模型。最终采用一阶导数的数学处理(1, 4, 4, 1)、去散射处理(SNV + Detrend)组合的预处理方法为最优定标模型。马铃薯蛋白质定标方程参数如表2所示, 其定标标准偏差(SEC)、交叉检验标准误差(SECV)和交叉验证相关系数(1-VR)分别为0.566、0.632和0.912, 说明所建的定标模型可用于马铃薯块茎蛋白质含量的快速检测, 该模型可代替常规测试方法使用。

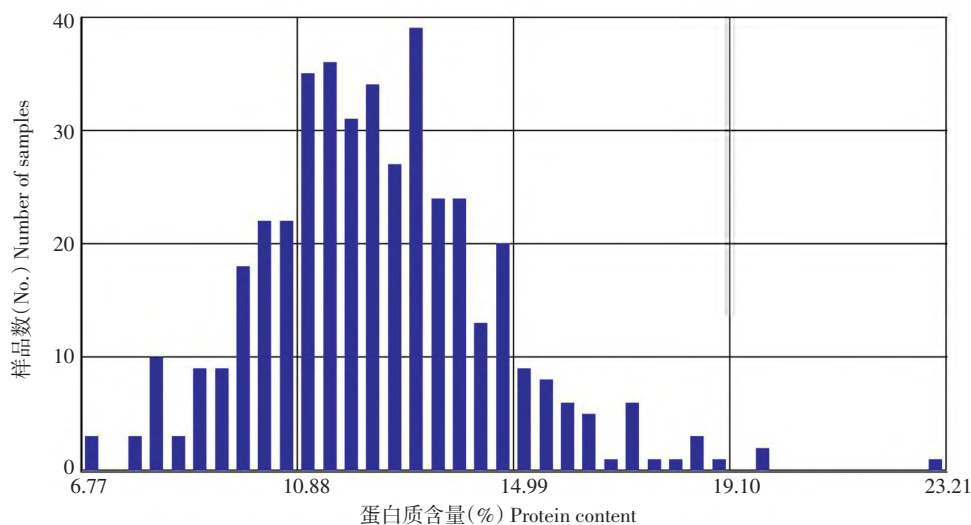


图2 马铃薯测试样品蛋白质含量柱形图

Figure 2 Histogram of protein contents of potato test samples

表1 蛋白定标集及验证集化学分析数据

Table 1 Chemical analysis of protein content calibration and validation sets

组分 Constituent	最小值 Minimum	最大值 Maximum	平均值 Mean	标准差 Standard deviation	样本数 Number of sample
定标集 Calibration	6.77	23.21	12.29	2.20	311
验证集 Validation	7.16	19.54	12.52	2.19	100

表2 马铃薯蛋白质组分定标模型参数

Table 2 Parameters of calibration model for potato protein contents

组分 Constituent	定标样本数 N	定标样品集平均浓度 Mean	定标样品集浓度最大值 Est.Max	定标标准偏差 SEC	交叉验证误差 SECV	交叉验证相关系数 1-VR
Protein	295	12.315	18.692	0.566	0.632	0.912

2.4 定标模型的验证

定标模型建立后用100份没有参与定标的样品来评估定标方程的预测性能。得到预测结果与常规方法测定结果及其偏差见表3以及马铃薯蛋白质预测值与化学值相关图(图3)。

在表3中,85号样品、89号样品和98号样品的化学值和预测值之间的差值偏大,视为异常样品,

作剔除处理,剔除异常样品后对剩下的97个样品的化学值和预测值进行相关性分析,结果如图3所示。

与常规化学分析测量结果之间的相关系数(R)为0.931,斜率为0.986,其斜率和相关系数均接近于1,结果表明近红外预测马铃薯蛋白质含量与传统方法结果无显著差异,所建的模型用于马铃薯蛋白质含量检测是准确可靠的。

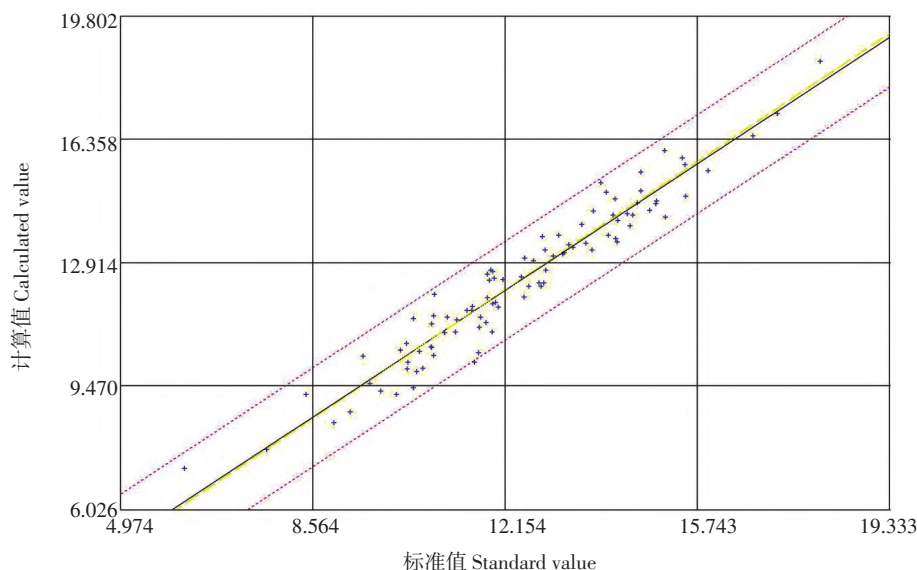


图3 近红外预测值与实验室分析值相关分析

Figure 3 Correlation analysis between near infrared predicted value and chemical analysis value

3 讨 论

试验建立了一个马铃薯块茎蛋白质含量的近红外定标模型,并对构建的定标方程的预测性能进行了评估。试验结果表明,马铃薯块茎蛋白质含量定标模型的SECV值为0.632,而1-VR值为0.912,蛋白质含量的验证参数SEP值为0.558, R 值为0.931(图3),说明所建模型与凯氏定氮法测定的蛋白质含量无显著差异,结果可靠、理想,检测精度高、重复性好,可以用于今后马铃薯蛋白质含量的快速测定。

应用近红外技术不仅可以大大缩短品质育种工作中的材料筛选时间,而且可以节省大量的人力、物力和财力,并且减少了很多工序,提高了工作效率,降低了有害气体的污染和对实验操作

人员的伤害。但是,近红外光谱技术是通过样品近红外光谱与化学值之间的定标模型来预测未知样品的组分含量,实际上是一个二级分析方法^[16]。影响定标准确度的因素很多,如参与定标的样品数量不足,不具代表性;定标样品差异性不显著造成定标不具代表性;样品近红外扫描数据差;定标所使用的实验室数据分析不精确;非线性因素对定标的影响等^[17]。特别是在定标集的样品选择上并不是样品数量越多越好,应选择具有代表性的,芦永军等^[18]研究表明,采用相似样品剔除算法从178个玉米粉样品中成功提取了94个优选样品,通过对178个样品和94个优选样品分别进行定标试验发现优选样品保持了由原始样品集参与定标所达到的定标精度,给出了满意的定标结果。样本比例分配也应适当,韩春亮等^[19]研究表

表3 化学法测定值和近红外预测值的比较

Table 3 Comparisons of values measured by chemical analysis and near infrared predicted

样品序号 Sample No.	化学值 Chemical analysis value	预测值 Predicted value	绝对误差 Absolute error	样品序号 Sample No.	化学值 Chemical analysis value	预测值 Predicted value	绝对误差 Absolute error
1	10.31	10.81	-0.50	51	13.65	12.85	0.80
2	11.40	11.69	-0.29	52	7.68	7.69	-0.01
3	12.26	12.58	-0.32	53	8.44	8.95	-0.51
4	14.57	14.96	-0.39	54	10.46	10.19	0.27
5	14.24	14.52	-0.28	55	10.39	11.65	-1.26
6	9.32	9.82	-0.50	56	10.12	10.33	-0.21
7	13.69	14.08	-0.39	57	12.35	12.78	-0.43
8	12.45	12.10	0.35	58	18.55	18.02	0.52
9	10.13	11.57	-1.44	59	13.26	13.77	-0.51
10	14.63	14.97	-0.34	60	9.22	8.42	0.80
11	14.37	14.84	-0.47	61	9.52	9.62	-0.10
12	10.98	11.90	-0.92	62	14.24	14.16	0.08
13	11.93	11.82	0.11	63	15.48	15.93	-0.45
14	12.71	12.90	-0.19	64	13.27	12.89	0.38
15	13.49	14.23	-0.74	65	11.59	11.44	0.15
16	13.69	13.15	0.54	66	14.76	15.52	-0.76
17	14.29	14.43	-0.14	67	11.57	11.53	0.04
18	13.46	13.65	-0.19	68	14.70	14.20	0.50
19	13.10	13.05	0.05	69	12.60	11.81	0.79
20	11.96	12.50	-0.54	70	9.41	10.42	-1.01
21	11.21	10.77	0.44	71	13.98	13.58	0.40
22	11.80	11.97	-0.17	72	12.71	11.87	0.84
23	10.43	10.54	-0.11	73	11.36	10.43	0.93
24	10.31	9.50	0.81	74	11.24	11.79	-0.55
25	14.93	14.68	0.25	75	7.16	6.16	1.00
26	14.18	15.13	-0.95	76	10.54	10.77	-0.23
27	11.67	12.02	-0.35	77	8.74	9.25	-0.51
28	14.36	13.79	0.57	78	12.42	11.86	0.56
29	11.44	10.82	0.62	79	13.24	11.56	1.68
30	15.65	15.51	0.14	80	11.11	11.65	-0.54
31	12.55	14.25	-1.70	81	12.25	12.81	-0.56
32	16.05	15.12	0.93	82	13.59	14.21	-0.62
33	10.98	11.21	-0.23	83	13.16	13.22	-0.06
34	12.35	12.87	-0.52	84	9.86	10.49	-0.63
35	12.04	10.82	1.22	85	19.54	15.89	3.65*
36	15.15	13.93	1.22	86	16.46	16.77	-0.31
37	15.45	14.68	0.77	87	11.40	11.07	0.33
38	17.08	17.23	-0.15	88	10.65	10.30	0.35
39	9.97	10.61	-0.64	89	12.12	14.91	-2.79*
40	15.84	15.46	0.38	90	9.22	10.12	-0.90
41	9.94	10.31	-0.37	91	14.59	14.62	-0.03
42	14.10	14.25	-0.15	92	13.95	14.48	-0.53
43	13.20	13.25	-0.05	93	13.42	13.34	0.08
44	10.95	11.02	-0.07	94	12.52	12.45	0.07
45	13.04	11.28	1.76	95	12.48	11.94	0.54
46	12.67	11.92	0.75	96	10.56	10.76	-0.20
47	12.97	12.68	0.29	97	13.35	13.41	-0.06
48	13.05	12.51	0.54	98	12.97	14.97	-2.00*
49	14.89	14.03	0.86	99	11.70	11.54	0.16
50	11.76	11.91	-0.15	100	11.32	11.25	0.07

注: *表示异常样品。

Note: * represents abnormal sample.

明以70%的比例样本作为定标模型的建立, 其余30%比例样本作为该模型的验证样本, 可以获得更好的预测效果。

定标模型并不是一劳永逸的, 由于自然样品其成分随着种植季节、施肥量、降雨量和种植条件的变化而发生相应变化, 因此, 定标方程应定期补充新样品的扫描光谱和化学分析数据进行逐步调整或升级, 目的是使定标方程不断适用待测样品的变化。如果样品的验证效果符合要求, 则不需要进行定标调整, 如果验证效果不符合要求, 则从实验室成分分析的准确性以及定标模型的适用性等方面寻找问题根源, 并进行相应的再次验证, 直到符合定标要求, 该模型才可以使用^[16]。在定标的过程中如遇到超常样品, 应对超常样品进行化学分析, 然后将样品添加到原定标样品系对模型进行升级, 对模型进行升级将使模型的预测性能更稳定。

目前, 天然样品近红外定标最常使用的定标技术为改进最小二乘法回归(MPLS), 很多广泛应用的商品化软件中都采用此种建模方式。但当选择的校正集样本中出现奇异点(即超常样品), 或个别样品的性质范围已经超出校正集样本的范围时, 则会出现较大偏差的可能。随着技术的不断革新, 人工神经网络(ANN)技术解决了定标面临的吸收非线性问题, 适用于处理大样品数据库, 至少需要1 000份样品, 因此其模型适用范围广, 可以减少或降低定标模型的调整工作, 在很多领域的应用中已取得了良好效果。如全球通用谷物定标开发, 即使样品地域或收购季节和品种变化时仍然获得较好结果。但是在光谱模型的构建过程中, 必须投入相对较多的材料和时间成本, 才能得到更加准确的校正模型。所以, 如何实现技术优化和更有效的模型共享, 仍是将近红外光谱技术研究及广泛应用的重要课题。

【参 考 文 献】

- [1] 林家永. 红外光谱分析技术在玉米品质分析中的研究进展[J]. 中国粮油学报, 2010, 25(4): 108-115.
- [2] 林家永, 范维燕, 邢郇. 近红外光谱技术在稻米品质分析中的应用[J]. 粮食与食品工业, 2008, 15(6): 48-53.
- [3] 苗雪雪, 苗莹, 龚浩如, 等. 不同偏最小二乘法的近红外光谱技术测定大米中水分的研究分析[J]. 科学学报, 2019, 35(5): 643-649.
- [4] 刘红梅, 肖正午, 申涛, 等. 稻米直链淀粉含量近红外检测模型的建立[J]. 湖南农业大学学报: 自然科学版, 2019, 45(2): 189-193.
- [5] 曲艺伟, 张鹤, 韩笑, 等. 花生脂肪酸近红外模型的建立[J]. 分子植物育种, 2019, 17(2): 568-578.
- [6] 李路, 黄汉英, 赵思明, 等. 大米蛋白质、脂肪、总糖、水分近红外检测模型研究[J]. 中国粮油学报, 2017, 32(7): 121-126.
- [7] 田翔, 刘思辰, 王海岗, 等. 近红外漫反射光谱法快速检测谷子蛋白质和淀粉含量[J]. 食品科学, 2017, 38(16): 140-144.
- [8] 张玉荣, 付玲, 周显青, 等. 基于近红外光谱技术测定小麦蛋白质模型的建立[J]. 粮食与饲料工业, 2013(4): 15-18.
- [9] 李君霞, 张洪亮, 严衍禄, 等. 水稻蛋白质近红外定量模型的创建及在育种中的应用[J]. 中国农业科学, 2006, 39(4): 836-841.
- [10] 王翠秀, 曹见飞, 顾振飞, 等. 基于近红外光谱大豆蛋白质、脂肪快速无损检测模型的优化构建[J]. 大豆科学, 2019, 38(6): 968-976.
- [11] 王铁固, 刘新香, 库丽霞, 等. 近红外反射光谱测定玉米完整子粒蛋白质和淀粉含量的校正模型[J]. 玉米科学, 2008, 16(3): 57-59.
- [12] 周菲, 王文军, 黄绪堂, 等. 食用向日葵子仁蛋白质含量近红外光谱模型的建立[J]. 作物杂志, 2013(6): 73-75.
- [13] 刘翠翠, 高红秀, 李赞, 等. 马铃薯块茎钾含量近红外模型的建立[J]. 中国马铃薯, 2011, 25(1): 65-68.
- [14] 张良, 赵雪君, 石瑛. 马铃薯直链淀粉含量近红外模型的建立[J]. 作物杂志, 2014(2): 73-76.
- [15] 国家食品药品监督管理总局, 国家卫生和计划生育委员会. GB 5009.5—2016 食品中蛋白质的测定[S]. 北京: 中国标准出版社, 2016.
- [16] 高红秀, 金萍, 杨亮, 等. 近红外光谱分析技术定标模型建立及优化[J]. 实验技术与管理, 2014, 31(2): 57-59.
- [17] 李勇, 魏益民, 王锋. 影响近红外光谱分析结果准确性的因素[J]. 核农学报, 2005, 19(3): 236-240.
- [18] 芦永军, 曲艳玲, 朴仁官, 等. 近红外光谱分析技术定标和预测中的相似样品剔除算法[J]. 光谱学与光谱分析, 2004(2): 158-161.
- [19] 韩春亮, 郑利宇, 崔凤霞. 近红外光谱的原理及应用[J]. 河南教育学院学报: 自然科学版, 2009, 18(4): 19-21.