DOI: 10.11895/j.issn.0253-3820.140707

苹果产地差异对可溶性固形物近红外光谱检测模型影响的研究

樊书祥[1,2] 黄文倩[2] 郭志明[2] 张保华[2] 赵春江*[1,2] 钱 曼[1,2]
[1](西北农林科技大学机械与电子工程学院, 杨凌 712100)
[2](北京市农林科学院, 北京农业智能装备技术研究中心, 北京 100097)

摘 要 为更好地利用近红外光谱预测苹果可溶性固形物含量,减少产地差异对近红外光谱检测模型的影响,以 4 种不同产地的富士苹果为研究对象,采用基于 *x*-*y* 共生距离的样本划分方法分别对不同产地的苹果选取代表性样本作为校正集,利用偏最小二乘算法,建立和比较单一产地和混合产地下的苹果可溶性固形物近红外光谱检测模型,并结合竞争性自适应重加权算法(CARS) 和连续投影算法(SPA) 对苹果可溶性固形物的建模变量进行筛选。相比单一产地和其它混合产地模型,混合所有 4 种苹果产地的校正集样本建立的模型取得了最好的预测结果,另外,结合 CARS-SPA 筛选的 16 个特征波长,模型得到了进一步简化,其预测相关系数和预测均方根误差分别为 0.978 和 0.441°Brix。结果表明,利用多个产地的苹果样本建立的混合模型,结合有效特征波长,可提高对苹果可溶性固形物含量的预测精度,减小产地差异对可溶性固形物近红外光谱检测的影响。

关键词 苹果; 产地; 近红外光谱; 可溶性固形物

1 引 言

可溶性固形物(Soluble solids content, SSC) 是包括可溶性糖、酸、纤维素等成分的综合型指标,是评价苹果内部品质的重要参数[1]。苹果可溶性固形物含量的快速有效检测对于苹果的生产流通,保证采后的果品品质至关重要。与传统破坏性检测方法相比,近红外光谱技术以其无损、快速、低成本的优点,在水果品质与安全检测方面得到了越来越广泛的应用[2]。

国内外学者对苹果可溶性固形物近红外光谱检测,进行了大量研究。Liu 等[3] 利用富士苹果的傅里叶近红外光谱实现了苹果可溶性固形物的有效检测。Peris 等[4,5] 先后分析了温度变化、季节和品种差异对苹果可溶性固形物近红外光谱检测模型的影响; Bobelyn 等[6] 以 Golden Delicious 和 Pink Lady 两种苹果为例简要分析了产地差异对可溶性固形物检测模型的影响。与此同时,文献[7,8] 结合特征波长优选算法,在简化苹果可溶性固形物近红外光谱检测模型的同时,提高了模型预测精度。

富士苹果在我国苹果生产中占有重要地位,其产地分布范围广,不同产区因土壤、光照、气候不同,苹果外观特征和内部品质也存在一定差异[9]。赵杰文等[10] 利用富士苹果傅里叶光谱信息结合支持向量机实现了产地分类。但先前的研究在富士苹果产地差异对于可溶性固形物近红外光谱检测模型的影响以及如何减小这种影响方面鲜有报道。因此,本研究以产自新疆阿克苏、山东肥城、山东栖霞和陕西宜川的富士苹果为实验对象,应用基于 *x*-*y* 共生距离(SPXY) 的样本划分方法,选取更具有代表性的样本作为校正集,建立和比较单一产地和混合产地的苹果可溶性固形物近红外光谱检测模型,并结合光谱特征波长变量优选,尝试在提高模型的稳定性和预测精度的基础上简化模型,为在实际生产中准确预测苹果可溶性固形物提供参考。

2 实验部分

2.1 实验材料

实验用样本选自我国富士苹果主产区,挑选无缺陷和损伤的共 368 个样品,其中产自新疆阿克苏

2014-08-19 收稿; 2014-09-28 接受
本文系北京市自然科学基金(No. 6144024) 项目,国家农业智能装备工程技术研究中心开放课题(No. KFZN2012N01-014) , 国家科技支撑计划(No. 2014BAD21B00) 项目资助
* E-mail: zhaocj@nercita.org.cn

76 个，山东栖霞 72 个，山东肥城 160 个以及陕西宜川 60 个。将苹果表面清洗干净，依次编号，标记采集区域。实验前，将苹果样品从冰箱取出，放置 12 h，使样本温度与室温达到一致，以避免温度对光谱测量结果产生影响[11]。

2.2 光谱采集与可溶性固形物实际值测量

实验采用 Antaris Ⅱ 傅立叶变换近红外光谱仪(Thermo Science Co.，USA)采集苹果样品在标记点区域的光谱信息，波段范围设置为 3800 ~ 14000 cm^{-1}，扫描次数 32，分辨率 4.0 cm^{-1}。光谱测量完成后使用数字阿贝折光仪(ARIAS 500，Reichert Technologies，New York，USA)进行 SSC 含量测定。每个样品从对应标记部位切取一定厚度果肉，经纱布过滤挤汁滴于折光仪镜面，读取并记录读数。

2.3 样本划分方法

基于 x-y 共生距离的样本划分方法(Sample set partitioning based on joint x-y distances，SPXY)以 Kennard-Stone 算法为基础，同时考虑样本的 x 变量(光谱数据)和 y 变量(SSC 值)的欧氏距离[12]。为了确保样本在 x 和 y 空间的具有相同的权重，标准化的 xy 的距离公式为：

$$d_{xy}(p,q) = \frac{d_x(p,q)}{\max_{p,q\in[1,N]} d_x(p,q)} + \frac{d_y(p,q)}{\max_{p,q\in[1,N]} d_y(p,q)}; \quad p,q \in [1,N] \tag{1}$$

其中，N 为样本总数，$d_x(p,q)$ 和 $d_y(p,q)$ 表示任意两个样本 p，q 之间在 x 变量和 y 变量的欧式距离。其最大优势在于能够有效覆盖多维空间，获得的校正集样本具有较强代表性[13]。因此应用 SPXY 算法分别对 4 个产地的苹果样本进行校正集和预测集的划分。

2.4 可溶性固形物检测模型的建立

偏最小二乘(PLS)算法稳定性好且抗干扰能力强[14]，分别利用 PLS 建立单一产地的苹果可溶性固形物近红外光谱检测模型，以及混合 2 种产地、3 种产地和所有 4 种产地苹果样本的混合产地可溶性固形物近红外光谱检测模型。为衡量所建模型的预测精度，利用建立好的模型分别预测 4 个单一产地下的预测集样本及所有产地的预测集样本。

2.5 特征波长筛选

因光谱变量之间存在大量的冗余和共线性信息[15]，采用竞争性自适应重加权算法(Competitive adaptive reweighted sampling，CARS)和连续投影算法(Successive projections algorithm，SPA)筛选苹果可溶性固形物近红外光谱特征波长。CARS 采用自适应加权采样技术保留 PLS 模型中回归系数绝对值大的波长变量，并根据交互验证均方根误差最小值获取与所测组分性质相关的波长变量[16]。SPA 利用向量的投影分析，在大量波长变量之间筛选含有最少冗余信息的变量组，使变量之间共线性达到最小[17]。

3 结果与分析

3.1 苹果样本的光谱分析及样本划分

为减小光谱仪首尾噪声影响，选择 4000 ~ 10000 cm^{-1} 共 3112 个波长点进行分析。图 1 为 368 个苹果样品的原始光谱信息，仅从图 1 难以看出不同产地苹果样本的光谱信息差异。对所有苹果样本光谱数据进行主成分分析可得，前 6 个主成分(Principal component，PC)可代表原始光谱 99.99% 的信息。应用 Kruskal-Wallis 检验对前 6 个主成分进行差异性检验，当 $p < 0.05$ 时说明差异性显著。该检验方法为非参数检验，对总体分布的正态性和方差齐性不作要求[18]。检验结果如表 1 所示，所有主成分对应的 $p < 0.05$，说明不同产地下的苹果光谱信息存在明显差别。

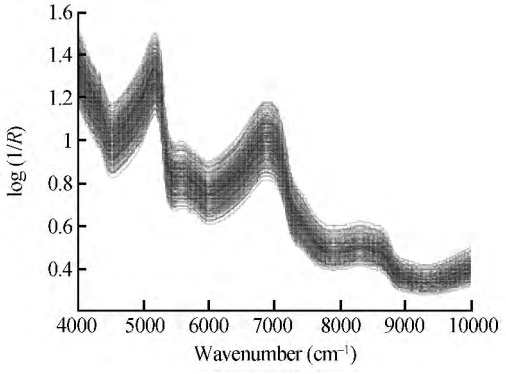

图 1 苹果样本原始光谱图

Fig. 1 Near infrared specra of apple sampls

样本划分前经异常值检验未发现异常样本。根据 SPXY 方法，分别从新疆阿克苏(AKS)、山东肥城(FC)、山东栖霞(QX)和陕西宜川(YC)4个产地的苹果样本中依次挑选出用于建模的校正集样本。

表 1　前 6 个主成分的 Kruskal-Willis 检验结果
Table 1　Kruskal-Willis test results of first six principal component (PC) scores

	PC1	PC2	PC3	PC4	PC5	PC6
卡方 Chi-sauare	46.765	224.709	14.010	106.840	68.104	78.819
自由度 Degree of freedom	3	3	3	3	3	3
渐进显著性 Asymptotic significance (p)	0.000	0.000	0.003	0.000	0.000	0.000

以 76 个阿克苏苹果为例,首先确定校正集的样本数为 60,根据公式(1) 计算所有 76 个样本中任意两样本之间的距离,获取距离最大的两样本点,然后从剩余样本集中选取到已获得的样本点距离最远的样本,重复此过程直到获取 60 个样本为止,剩余的 16 个样本作为预测集。不同产地苹果样本的划分及对应的 SSC 测量值分布如表 2 所示。从表 2 可知,每个产地下校正集样本的可溶性固形物含量分布范围均大于预测集样本,有利于构建更加稳定可靠的检测模型。

表 2　苹果可溶性固形物含量实测值的统计结果
Table 2　Statistic values of soluble solids content (SSC) (°Brix) of apples

产地 Origin	数据集 Data set	样本数 Samples	最小值 Minimum	最大值 Maximum	平均值 Mean	标准差 Standard deviation (SD)
Akesu (AKS)	校正集 Calibration set	60	12.76	21.89	18.16	2.34
	预测集 Prediction set	16	14.38	19.86	17.91	1.70
Feicheng (FC)	校正集 Calibration set	120	10.81	17.13	13.47	1.37
	预测集 Prediction set	40	11.34	14.69	13.29	0.94
Qixia (QX)	校正集 Calibration set	60	10.89	17.21	13.92	1.28
	预测集 Prediction set	12	12.74	15.68	14.17	0.88
Yichuan (YC)	校正集 Calibration set	50	10.79	18.33	14.73	1.54
	预测集 Prediction set	10	13.18	16.49	14.87	1.10

3.2　单一产地的可溶性固形物检测模型预测结果

分别利用产自新疆阿克苏、山东肥城、山东栖霞、陕西宜川的校正集样本建立其对应的单一产地可溶性固形物近红外光谱检测模型,其预测相关系数(R_p) 和预测均方根误差(Root mean square error of prediction, RMSEP) 如表 3 所示。对于同一产地的苹果样本,其校正集建立的单一产地模型对其预测集样本均实现了较好预测。但在实际生产中,对未知样本进行预测时,若先根据其光谱信息判定其产地信息,再根据对应产地下的模型预测可溶性固形物,一定程度上可提高预测精度,但工作量大,不利于实际生产。比较可知,利用某单一产地下的检测模型预测其它不同产地苹果的可溶性固形物时会产生较大误差,R_p 也有不同程度下降。因此,建立混合产地模型更具现实意义。

表 3　单一产地的可溶性固形物检测模型预测结果
Table 3　Prediction results of local origin models for prediction of SSC in apples

校正集 Calibration set	预测集 Prediction set									
	AKS		FC		QX		YC		AKS-FC-QX-YC	
	R_p	RMSEP	R_p	RMSEP	R_p	RMSEP	R_p	RMSEP	R_p	RMSEP
AKS	0.903	0.742	0.827	1.382	0.703	0.683	0.685	1.333	0.909	1.180
FC	0.887	1.023	0.918	0.373	0.845	0.583	0.840	2.287	0.928	1.005
QX	0.856	1.864	0.662	1.120	0.852	0.464	0.657	2.181	0.913	1.414
YC	0.875	0.943	0.742	0.998	0.806	0.585	0.876	0.517	0.933	0.885

R_p: 预测相关系数(Correlation coefficient of predication); RMSEP: 预测均方根误差(Root mean square error of prediction)。

3.3　混合产地的可溶性固形物检测模型预测结果

将不同产地苹果的校正集样本混合,建立混合产地的苹果可溶性固形物近红外光谱检测模型,对 4 个单一产地的预测集样本,以及所有产地的预测集样本的预测结果见表 4。随着校正集苹果产地混合数量的增加,模型的预测精度不断提升。混合所有 4 种苹果产地的校正集建立的模型对各预测集样本均取得了较好结果。通过以上对比可知,当校正集包含更多产地的苹果样本光谱信息时,建立的模型对

未知产地苹果的可溶性固形物预测会取得更好的结果，减小苹果的产地差异对于可溶性固形物近红外光谱检测的影响。

表 4 混合产地可溶性固形物检测模型预测结果

Table 4 Prediction results of hybrid origin models for prediction of SSC in apples

校正集 Calibration set	预测集 Prediction set									
	AKS		FC		QX		YC		AKS-FC-QX-YC	
	R_p	RMSEP	R_p	RMSEP	R_p	RMSEP	R_p	RMSEP	R_p	RMSEP
AKS-FC	0.917	0.730	0.900	0.429	0.867	0.556	0.854	1.157	0.955	0.650
AKS-QX	0.900	0.739	0.890	1.562	0.847	0.483	0.697	1.329	0.920	1.275
AKS-YC	0.916	0.694	0.806	0.815	0.894	0.417	0.869	0.529	0.960	0.709
FC-QX	0.889	0.884	0.929	0.346	0.894	0.385	0.815	1.777	0.943	0.806
FC-YC	0.901	1.289	0.926	0.353	0.861	1.033	0.961	0.369	0.959	0.762
QX-YC	0.899	0.813	0.848	0.933	0.860	0.459	0.859	0.539	0.951	0.807
AKS-FC-QX	0.916	0.758	0.938	0.325	0.916	0.337	0.857	1.477	0.953	0.685
AKS-FC-YC	0.921	0.697	0.925	0.373	0.875	0.802	0.921	0.440	0.970	0.543
AKS-QX-YC	0.918	0.672	0.815	0.566	0.798	0.536	0.901	0.460	0.963	0.573
FC-QX-YC	0.912	0.891	0.923	0.363	0.914	0.349	0.953	0.349	0.975	0.514
AKS-FC-QX-YC	0.928	0.689	0.922	0.368	0.890	0.390	0.948	0.422	0.976	0.461

3.4 基于特征波长的混合产地模型

为简化模型，提高模型预测精度，在 4000 ~ 10000 cm^{-1} 全波段范围内，采用 CARS 算法对混合有 4 种苹果产地的校正集样本的光谱进行可溶性固形物特征波长筛选。因每运行一次 CARS 算法其最优采样次数略有不同，尝试运行 CARS 算法 50 次，选取交互验证均方根误差(Root mean square error of cross validation，RMSECV)最小的一次，图 2 为此次 CARS 算法对可溶性固形物特征波长的筛选过程。由图 2 可见，当采样次数为 47 次时，其 RMSECV 达到最小值，此时对应建模变量数为 102 个。将挑选的特征波长作为输入变量，建立苹果可溶性固形物预测模型，结果如表 5 所示。与全波段所建模型相比，其预测结果略有提升，建模变量数由全波段建模的 3112 减少到 102，模型得到了大大简化。

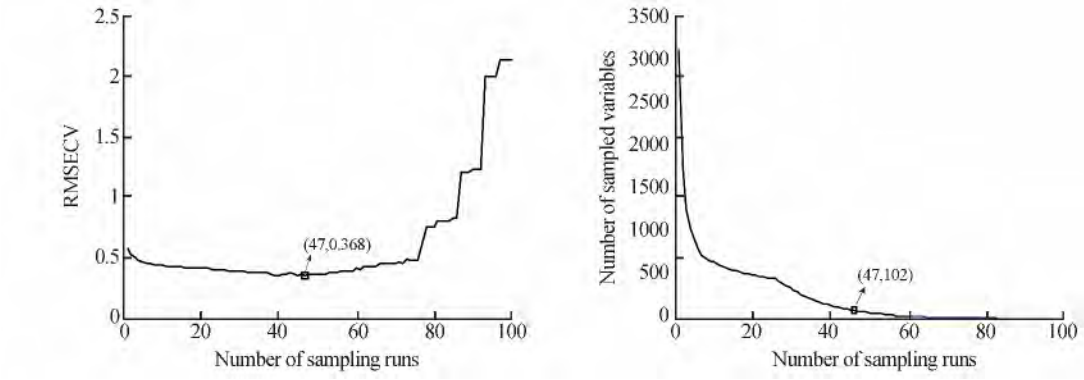

图 2 基于 CARS 算法的特征变量筛选

Fig. 2 Plot of variable selection by competitive adaptive reweighted sampling algorithm (CARS)

表 5 苹果可溶性固形物不同偏最小二乘模型预测结果

Table 5 Prediction results of different PLS models for prediction of SSC in apples

模型 Model	变量数 Variables	预测集 Prediction set									
		AKS		FC		QX		YC		AKS-FC-QX-YC	
		R_p	RMSEP	R_p	RMSEP	R_p	RMSEP	R_p	RMSEP	R_p	RMSEP
PLS	3112	0.928	0.689	0.922	0.368	0.890	0.390	0.948	0.422	0.976	0.461
CARS-PLS	102	0.928	0.678	0.934	0.341	0.898	0.372	0.938	0.428	0.978	0.446
CARS-SPA-PLS	16	0.932	0.640	0.922	0.358	0.928	0.332	0.924	0.463	0.978	0.441

SPA: Successive projections algorithm.

CARS 算法剔除了大量无关信息,但其挑选的波长仍存在一定共线性,且对于实际生产建模变量依然众多。因此采用 SPA 算法对经 CARS 选择后的 102 个变量进一步优选,得到 4013, 4302, 4458, 4539, 4898, 5029, 5264, 5299, 6007, 6282, 6620, 7312, 8641, 8745, 9295 和 9497 cm^{-1}共 16 个特征波长变量,并建立对应的苹果可溶性固形物近红外光谱检测模型,结果如表 5 所示。与 CARS-PLS 模型相比,基于 CARS-SPA 筛选的 16 个特征波长建立的模型更为简单,其对所有产地的预测集样本检测时 R_p = 0.978,RMSEP 为 0.441° Brix。通过比较还发现,CARS-SPA-PLS 模型对于每种产地的苹果单独预测时其预测相关系数均大于 0.92。该模型对不同产地预测集样本的预测值和其实际测量值之间的散点图如图 3 所示。通过特征波长筛选,在保证模型精度的前提下,模型得到了进一步简化,为实现今后苹果可溶性固形物在线检测提供参考。

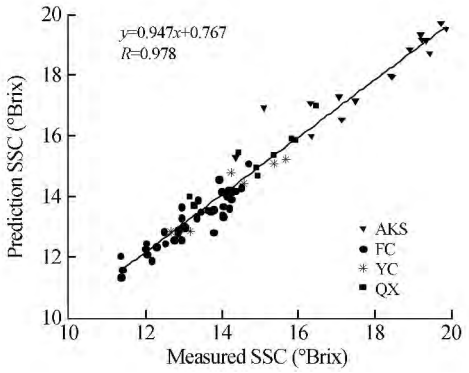

图 3　CARS-SPA-PLS 模型的预测集样本的实际值和预测值散点图

Fig. 3　Measured versus predicted values for SSC by the CARS-SPA-PLS model

4　结　论

本研究尝试以 4 种不同产地的富士苹果为研究对象,探讨了苹果的产地差异对近红外光谱检测模型的影响。相比单一产地和其它混合产地模型,混合 4 种产地苹果的校正集建立的模型取得了理想的预测结果,结合 CARS-SPA 筛选的 16 个特征波长变量,模型得到了进一步的简化,其预测相关系数和预测均方根误差分别为 0.978 和 0.441° Brix。结果表明,含有更多产地的苹果样本建立的校正模型,结合有效筛选的特征波长,可以实现苹果可溶性固形物含量的准确预测,减小苹果产地差异对可溶性固形物近红外光谱检测的影响,为实际生产中利用近红外光谱技术实现苹果可溶性固形物含量的准确、在线检测提供理论基础。

References

1　Mendoza F, Lu R, Ariana D Cen H, Bailey B. *Postharvest Biol. Tec.*, **2011**, 62(2): 149 – 160

2　Nicolaï B M, Beullens K, Bobelyn E, Peirs A, Saeysa W, Theron K I, Lammertyn J. *Postharvest Biol. Tec.*, **2007**, 46(2): 99 – 118

3　Liu Y, Ying Y. *Postharvest Biol. Tec.*, **2005**, 37(1): 65 – 71

4　Peirs A, Scheerlinck N, Nicolaï B M. *Postharvest Biol. Tec.*, **2003**, 30(3): 233 – 248

5　Peirs A, Tirry J, Verlinden B, Verlinden B, Darius P, Nicolaï B M. *Postharvest Biol. Tec.*, **2003**, 28(2): 269 – 280

6　Bobelyn E, Serban A S, Nicu M, Lammertyn J, Nicolaï B M, Saeys W. *Postharvest Biol. Tec.*, **2010**, 55(3): 133 – 143

7　Zou X, Zhao J, Huang X, Li Y. *Chemometr. Intell. Lab. Syst.*, **2007**, 87(1): 43 – 51

8　OUYANG Ai-Guo, XIE Xiao-Qiang, ZHOU Yan-Rui, LIU Yan-De. *Spectroscopy and Spectral Analysis*, **2012**, 32(10): 2680 – 2684
欧阳爱国, 谢小强, 周延睿, 刘燕德. 光谱学与光谱分析, **2012**, 32(10): 2680 – 2684

9　GUO Zhi-Ming, HUANG Wen-Qian, PENG Yan-Kun, WANG Xiu, TANG Xiu-Ying. *Chinese J. Anal. Chem.*, **2014**, 42(4): 513 – 518
郭志明, 黄文倩, 彭彦昆, 王 秀, 汤修映. 分析化学, **2014**, 42(4): 513 – 518

10　ZHAO Jie-Wen, HU Huai-Ping, ZOU Xiao-Bo. *Transactions of the Chinese Society of Agricultural Engineering*, **2007**, 23(4): 149 – 152
赵杰文, 呼怀平, 邹小波. 农业工程学报, **2007**, 23(4): 149 – 152

11　FAN Shu-Xiang, HUANG Wen-Qian, LI Jiang-Bo, ZHAO Chun-Jiang, ZHANG Bao-Hua. *Spectroscopy and Spectral Analysis*, **2014**, 34(8): 2089 – 2093
樊书祥, 黄文倩, 李江波, 赵春江, 张保华. 光谱学与光谱分析, **2014**, 34(8): 2089 – 2093

12 SHANG Liang, GU Jing-Si, GUO Wen-Chuan. *Transactions of the Chinese Society of Agricultural Engineering*, **2013**, 29(17): 257 - 264
商 亮, 谷静思, 郭文川. 农业工程学报, 2013, 29(17): 257 - 264
13 Galvão R K H, Araujo M C U, José G E, Pontes M J C, Silva E C, Saldanha T C B. *Talanta*, **2005**, 67(4): 736 - 740
14 Li J, Huang W, Zhao C, Zhang B. *J. Food Eng.*, **2013**, 116(2): 324 - 332
15 ZHANG Chu, LIU Fei, KONG Wen-Wen, ZHANG Hai-Liang, HE Yong. *Transactions of the Chinese Society of Agricultural Engineering*, **2013**, 29(20): 270 - 277
张 初, 刘 飞, 孔汶汶, 章海亮, 何 勇. 农业工程学报, 2013, 29(20): 270 - 277
16 Li H, Liang Y, Xu Q, Cao D. *Anal. Chim. Acta*, **2009**, 648(1): 77 - 84
17 Araújo M C U, Saldanha T C B, Galvão R K H, Yoneyama T, Chame H C, Visani V. *Chemometr. Intell. Lab.*, **2001**, 57(2): 65 - 73
18 Yao Y, Chen H, Xie L, Rao X. J. *Food Eng.*, **2013**, 119(1): 22 - 27

Assessment of Influence of Origin Variability on Robustness of Near Infrared Models for Soluble Solid Content of Apples

FAN Shu-Xiang[1,2], HUANG Wen-Qian[1], GUO Zhi-Ming[2], ZHANG Bao-Hua[2], ZHAO Chun-Jiang*[1,2], QIAN Man[1,2]
[1](*College of Mechanical and Electronic Engineering, Northwest Agricultural and Forestry University, Yangling*, 712100, *China*)
[2](*Beijing Research Center of Intelligent Equipment for Agriculture, Beijing Academy of Agriculture and Forestry Sciences, Beijing* 100097, *China*)

Abstract In order to improve the precision and robustness in determination of soluble solids content (SSC) of 'Fuji' apple by NIR spectroscopy and eliminate the effect of origin variability on the accuracy of NIR calibration models for the SSC, sample set partitioning based on joint *x*-*y* distances (SPXY) was used to select representative subset from the apple samples of 4 different origins. As a comparison, partial least square (PLS) was used to establish local origin and hybrid origin models for the prediction of SSC in apple. Competitive adaptive reweighted sampling (CARS) and successive projections algorithm (SPA) were implemented to select effective variables of the NIR spectroscopy of SSC of apple. The results indicated that the PLS model established based on the 4 origin apple samples performed better than local origin and other hybrid origin models. The model could be effectively simplified using 16 characteristic variables selected by CARS-SPA method from full-spectrum which had 3112 wavelengths. The correlation coefficient (R_p) and root mean square error of prediction (RMSEP) were 0.978 and 0.441 °Brix, respectively for SSC. It was found that the model developed by more samples of different origins combined with effective wavelengths showed good prediction ability for apple sample of unknown origin, which indicated that it could significantly reduce the origin effect on the robustness of NIR models for SSC of apple.

Keywords Apple; Origin; Near infrared spectrum; Soluble solid content

(Received 19 August 2014; accepted 28 September 2014)

This work was supported by the Natural Science Foundation of Beijin, China (No. 6144024), Open Project of National Reserch Center of Inlelligent Equipment fro Agriculture (No. KFZN2012N01-014) and the Key Projects in the National Science & Technology Pillar Program (No. 2014BAD21B00)