SPXY算法的西瓜可溶性固形物近红外光谱检测

王世芳[1]，韩　平[1*]，崔广禄[2]，王　冬[1]，刘珊珊[1]，赵　跃[2]

1. 北京农业质量标准与检测技术研究中心，北京　100097
2. 北京市大兴区农业技术推广站，北京　102600

摘　要　可溶性固形物(SSC)是一种综合参数，主要包括糖、酸、纤维素、矿物质等成分，对评价果实成熟度和品质具有重要意义，影响果实口感、风味及货架期。西瓜可溶性固形物含量的无损快速检测对西瓜成熟度的确定、贮藏及运输过程中西瓜内部品质监控具有十分重要的意义，有助于提高西瓜生产效益和市场竞争力。在西瓜可溶性固形物含量的快速无损近红外光谱检测中，近红外漫透射的方式所需光源的能量大，同时大功率透射会对水果的内部品质产生影响；采用近红外漫反射方式的研究较少，但漫反射采集所需的能量小，有助于实现仪器小型便携化，成本低，同时避免透射引起的水果品质变化。以小型西瓜为研究对象，利用JDSU便携式近红外光谱仪采集西瓜样品瓜梗、瓜脐、赤道部位的近红外反射光谱，在976，1 186和1 453 nm附近有明显的吸收，利用偏最小二乘回归定量分析方法建立西瓜可溶性固形物的近红外光谱无损预测模型。首先，采用光谱-理化值共生距离(SPXY)算法对西瓜不同检测部位的样品集进行划分，以可溶性固形物含量为y变量，光谱为x变量，利用两种变量同时计算样品间距离，以保证最大程度表征样本分布，有效地覆盖多维向量空间，增加样本间的差异性和代表性，提高模型稳定性。将西瓜样品划分为51个校正集和15个预测集，校正集样本的SSC含量涵盖了预测集样本的SSC含量范围，且变异系数均小于9%，样品集划分合理，有助于建立稳健可靠的预测模型。其次，对比分析西瓜瓜梗、瓜脐、赤道检测部位的近红外反射光谱与可溶性固形物含量之间的定量模型的预测精度，结果得出西瓜赤道部位的反射光谱与可溶性固形物含量相关性较高，预测效果较好，预测集相关系数为0.629，预测集均方根误差为0.49%。对于不同检测部位获取的光谱信息所建立的近红外光谱SSC预测模型的精度问题，一方面与光谱的采集方式有关，另一方面与西瓜的产地、品种、成熟期等因素引起的其性状上的差异有关。在模型建立过程中根据实际情况确定西瓜的检测部位。最后，为提高西瓜赤道部位近红外反射光谱与可溶性固形物含量之间的预测模型精度，采用光谱预处理方法进行优化，结果得出经标准归一化预处理后，建立的偏最小二乘回归预测模型效果最佳，预测集相关系数为0.864，预测集均方根误差为0.33%，模型相关性较好，预测精度得到了很大提升。研究结果表明，近红外反射光谱检测小型西瓜赤道部位能很好预测其可溶性固形物含量，为实际生产中近红外光谱无损快速检测西瓜可溶性固形物含量及小型便携式仪器研发提供了技术储备。

关键词　小型西瓜；近红外反射光谱；SPXY算法；检测部位；可溶性固形物

中图分类号：O657.3　　**文献标识码**：A　　**DOI**：10.3964/j.issn.1000-0593(2019)03-0738-05

引　言

西瓜具有较高的食用价值和药用价值，水分和糖分含量较高，是人们夏天解暑的水果之一。可溶性固形物(soluble solid content，SSC)是一种综合参数，主要包括糖、酸、纤维素、矿物质等成分，对评价果实成熟度和品质具有重要意义，影响果实口感、风味及货架期。西瓜SSC含量的无损快速检测对西瓜成熟度的确定、贮藏及运输过程中西瓜内部品质监控具有十分重要的意义，有助于提高西瓜生产效益和市场竞争力。

近红外光谱检测具有快速、无损且多组分同时检测的优点，在农产品品质检测与评价方面得到广泛的应用。在西瓜SSC近红外漫反射和漫透射光谱检测方面，国内外已开展了

收稿日期：2017-12-25，**修订日期**：2018-04-18

基金项目：国家重点研发计划(2017YFD0801201)和北京市农林科学院所级科技创新团队建设项目(JNKST201620)资助

作者简介：王世芳，女，1989年生，北京农业质量标准与检测技术研究中心实习研究员　　e-mail：wangshifang1302@126.com

*通讯联系人　　e-mail：hanping1016@163.com

一系列研究。介邓飞等[1-2]选取 680～920 nm 波段，对西瓜 SSC 进行漫透射近红外光谱检测，瓜脐部位 SSC 的偏最小二乘回归模型精度优于最小二乘支持向量机模型，R_p 和 RMSEP 分别为 0.823 和 0.652%。韩东海等[3]采用漫透射可见近红外光谱采集西瓜顶部和赤道部位深层、中层和浅层的光谱信息，发现中层和浅层信息为光谱预测的主要信息，中层区域为最佳基础信息采集区域。钱曼等[4]选取 489～1 156 nm 波段，对西瓜瓜脐、瓜梗、赤道部位 SSC 进行漫透射近红外光谱检测，结果表明所有检测部位的校正集样本参与模型的建立能够得到较优的预测效果，同时利用自适应重加权算法对西瓜 SSC 近红外光谱变量进行特征波长筛选，采用筛选出的 42 个特征波长所建立的模型预测精度提高，模型得到简化，R_p 达到了 0.890 以上，RMSEP 小于 0.721°Brix。Qi 等[5]利用可见-近红外漫透射光谱采集 500～1 010 nm 波段的光谱，建立西瓜 SSC、番茄红素和水分的偏最小二乘模型，采用 X 射线定标线性方程建立体积和质量的模型，综合分析，评价西瓜成熟度。Elena 等[6]利用近红外漫反射光谱在线检测西瓜 SSC，番茄红素和 β-胡萝卜素含量，预测效果较好，适用于水果市场。对于西瓜 SSC 含量的快速无损近红外光谱检测，许多研究学者采用漫透射的方式，主要原因是西瓜瓜皮质地坚硬，对反射光的采集比较困难，但透射所需光源的能量大，同时大功率透射会对水果的内部品质产生影响；采用漫反射方式采集的研究较少，漫反射采集所需的能量小，有助于实现仪器小型便携化，成本低，同时避免透射引起的水果品质变化。

建立西瓜 SSC 的近红外光谱预测模型的实际应用中，为了提高所建模型稳定性和准确性，避免样本间差异过小或相同的情况引起预测模型过拟合或预测效果差，采用光谱-理化值共生距离(sample set partitioning based on joint x-y distances，SPXY)算法对样品集进行划分，使校正集中的样本 SSC 分布较大，增加样本间的差异性和代表性，从而减少校正集的样本数量和建模过程中的运算量。本研究采用近红外反射方式采集西瓜不同部位的光谱，采用 SPXY 算法对样品集进行划分，并采用光谱预处理对较优检测部位的 SSC 模型进行优化，得到较优的预测模型，为实际应用中近红外光谱测定西瓜 SSC 含量无损快速检测及装备的研发提供技术储备和参考依据。

1　实验部分

1.1　材料

供试西瓜于 2017 年 6 月在北京市大兴区农业科技成果展示基地采集。挑选形状规则、大小均匀且表皮无机械损伤的小型西瓜(品种为“L600”)66 个。当天运送至实验室，室温放置 24 h，确保样本温度与室温一致，避免温度对光谱采集及 SSC 含量测定产生影响。

1.2　光谱采集、SSC 和皮厚测定

清除西瓜表面的杂质和灰尘，并对西瓜进行编号，针对每个西瓜标记 5 个光谱采集位置，瓜梗部位和瓜脐部位各标记 1 个点，赤道部位均匀标记 3 个点(间隔约 120°)，如图 1 所示。采用 JDSU 便携式近红外光谱仪(美国 ASD 公司，型号 MicroNIR1700)进行反射光谱采集，测定范围为 908.1～1 676.2 nm，光谱间隔为 6.2 nm，光谱分辨率为 10 nm。采用日本 ATAGO 公司的 PR-101 折射式数字糖度计测定 SSC 含量。从每个西瓜标记部位取出一定厚度果肉，混合榨汁，将果汁滴于折光仪镜面窗口，读取 SSC 值并记录，最大值为 11.60%，最小值为 7.50%，平均值为 9.89%，标准差为 0.84%。对小型西瓜的皮厚进行检测，最大值为 0.71 cm，最小值为 0.45 cm，平均值为 0.58 cm，标准差为 0.05%。且西瓜中 SSC 含量与西瓜皮厚几乎没有相关性。

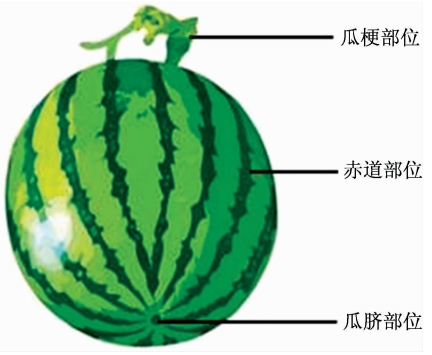

图 1　西瓜部位图

Fig. 1　The figure of watermelon tissues

1.3　光谱处理与数据分析

采用 SPXY 算法对样品集进行划分。SPXY 算法是由 Galvao 等[7]首先提出的，原理是采用理化值和光谱两种变量同时计算样品间距离，以保证最大程度表征样本分布，有效地覆盖多维向量空间，增加样本间的差异性和代表性，提高模型稳定性。

近红外漫反射光谱与西瓜 SSC 之间的定量分析模型采用偏最小二乘(partial least squares regression，PLSR)算法，在 Unscrambler 9.7 软件(挪威 CAMO 公司)中实现。为了减少背景噪声、基线漂移、杂散光等无用信息对原始光谱数据的干扰，进行平滑、导数、标准归一化、多元散射校正、基线校正等光谱预处理，对模型进行优化。模型评价指标有校正集相关系数(R_c)、校正样本均方根误差(RMSEC)、预测集相关系数(R_p)和预测样本均方根误差 (RMSEP)，R_c 值越大，RMSEC 越小，则模型的校正效果越好；R_p 值越大，RMSEP 越小，则模型的预测效果越好。

2　结果与讨论

2.1　西瓜不同部位原始近红外光谱图

对西瓜赤道部位采集的三条光谱取平均值，并将平均光谱作为赤道部位原始光谱进行分析。对不同检测部位采集的西瓜近红外光谱进行平均，见图 2。三个检测部位采集的西瓜样本的谱图变化趋势类似，且在 976，1 186 和 1 453 nm 附近有明显的吸收，1 453 nm 附近的吸收峰值的吸收强度最大，976 和 1 453 nm 附近吸收峰值主要是由水分含量引起的，1 186 nm 附近峰值主要是糖分吸收引起的。从一维谱图

中很难得出与糖分相关性最好的检测部位，需要引入定量分析方法进行分析。

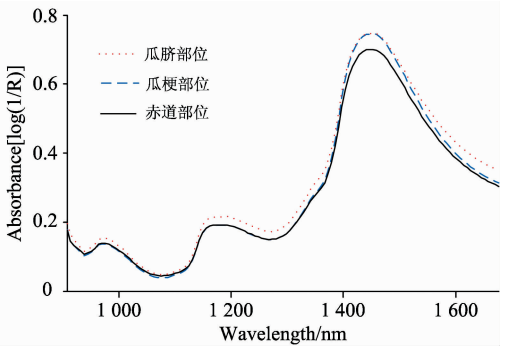

图 2 西瓜不同部位的原始近红外光谱图

Fig. 2 The near infrared spectra of the watermelon in different detection positions

2.2 样品集划分

SPXY 算法在样品集的划分中应用广泛，优于随机算法(random sampling，RS)，Kennard-Stone(KS)算法，双向算法(duplex)等，Guo 等[8] 采用 SPXY 方法将样本集划分为 106 个校正集和 54 个预测集样本，Zhu 等[9] 采用 SPXY 方法将样本集划分为 160 个校正集和 40 个预测集样本，都得到很好的预测效果，说明采用 SPXY 算法对样本集进行划分能够有效地提高模型预测精度。本研究利用 SPXY 算法对样品集进行划分，以 SSC 为 y 变量，近红外光谱值为 x 变量，将西瓜样本划分为 51 个校正集和 15 个预测集样本。SPXY 算法采用 x 变量和 y 变量同时计算样品间距离，样品集的划分结果见表 1。从表 1 中可以看出，校正集样本的 SSC 含量涵盖了预测集样本的 SSC 含量范围，且变异系数均小于 9%，样品集划分合理，有助于建立稳健可靠的预测模型。

表 1 样品集划分

Table 1 The partition of sample set

统计参数	赤道部位		瓜脐部位		瓜梗部位		平均光谱	
	校正集	预测集	校正集	预测集	校正集	预测集	校正集	预测集
最小值/%	8.10	9.10	8.10	8.10	8.10	8.10	8.10	8.10
最大值/%	11.60	11.10	11.60	11.20	11.60	11.20	11.60	11.20
平均值/%	9.93	10.05	9.98	9.87	9.98	9.87	9.98	9.87
标准偏差/%	0.83	0.63	0.77	0.86	0.77	0.86	0.77	0.86
变异系数/%	8.32	6.30	7.69	8.67	7.69	8.67	7.69	8.67

2.3 西瓜不同部位 SSC 定量分析模型对比

采用 PLSR 算法分别建立西瓜不同部位光谱、整果平均光谱与西瓜 SSC 含量的定量分析模型，并对模型预测效果进行比对分析，结果见表 2。从模型校正集和预测集的相关系数来分析，赤道部位和整果平均光谱建立的定量分析模型效果较好，其次是瓜脐部位，瓜梗部位建立的模型效果最差；校正样本和预测样本均方根误差越小，模型效果越好，综合考虑，赤道部位的模型效果预测较好，R_c 和 RMSEC 分别为 0.717 和 0.57%，R_p 和 RMSEP 分别为为 0.629 和 0.49%。相关研究结果表明，西瓜不同采集部位光谱信息的差异会对近红外光谱 SSC 预测模型的精度产生影响。介邓飞等[1] 采用漫透射方式采集了西瓜赤道部位、瓜梗和瓜脐的近红外光谱并建立了 SSC 的定量预测模型，研究结果表明赤道部位获取的光谱信息所建立的预测模型检测精度较差，而采用瓜脐部位获取的光谱信息建立的预测模型略好于瓜梗部位；韩东海等[3] 采用近红外光谱分别建立的瓜顶部位和赤道部位 SSC 的 PLS 预测模型，结果表明瓜顶部位的预测模型精度较好。基于以往的研究结果和本研究结果，作者认为对于不同检测部位获取的光谱信息所建立的近红外光谱 SSC 预测模型的精度问题，一方面与光谱采集的方式(反射和透射)有关，另一方面与西瓜的产地、品种、成熟期等因素引起的其性状上的差异有关，在模型建立过程中根据实际情况确定西瓜的检测部位。

2.4 西瓜赤道部位 SSC 定量分析模型的优化

为了进一步提高西瓜赤道部位 SSC 近红外光谱预测模型精度，采用光谱预处理方法提高光谱的信噪比，对模型进行优化。标准归一化光谱预处理后的西瓜赤道部位的近红外光谱图，见图 3。从图 3 可以看出，经标准归一化预处理的光

表 2 西瓜不同部位可溶性固形物含量近红外光谱定量分析模型结果

Table 2 Results of near infrared spectral quantitative analysis models for soluble solid content of watermelon with different detection positions

部位	PC	R_c	RMSEC/%	R_P	RMSEP/%
赤道	5	0.717	0.57	0.629	0.49
瓜脐	2	0.562	0.65	0.606	0.71
瓜梗	3	0.501	0.77	0.269	0.40
整果	4	0.630	0.66	0.602	0.44

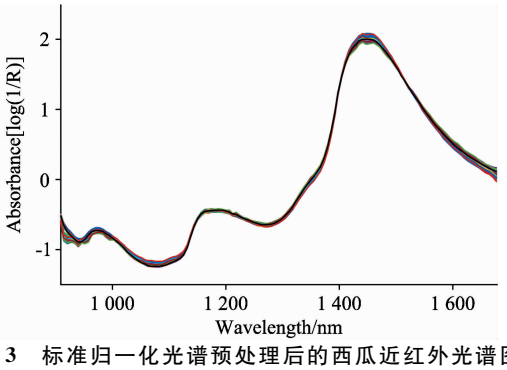

图 3 标准归一化光谱预处理后的西瓜近红外光谱图

Fig. 3 The near infrared spectra of the watermelon by standard normal variable

谱信噪比提高，基线平稳。不同光谱预处理后的模型分析结果，见表 3。

表 3　光谱预处理的西瓜可溶性固形物含量近红外光谱定量分析模型优化结果

Table 3　The optimization results of near infrared spectral quantitative analysis models for soluble solid content of watermelon based on spectral pretreatment

预处理方法	PC	校正集		预测集	
		R_c	RMSEC/%	R_P	RMSEP/%
原始光谱	5	0.717	0.57	0.629	0.49
平滑(3 点)	3	0.648	0.62	0.574	0.51
多元散射校正	6	0.796	0.50	0.621	0.50
基线校正	6	0.757	0.53	0.547	0.56
标准归一化	10	0.945	0.27	0.864	0.33
一阶导数(3 点)	5	0.790	0.50	0.596	0.53
一阶导数(9 点)	6	0.785	0.51	0.675	0.47
一阶导数(15 点)	11	0.904	0.35	0.685	0.47
二阶导数(9 点)	8	0.930	0.30	0.707	0.45
二阶导数(13 点)	5	0.770	0.52	0.738	0.42
二阶导数(17 点)	7	0.850	0.43	0.771	0.39
二阶导数(19 点)	7	0.836	0.45	0.783	0.38
二阶导数(21 点)	7	0.829	0.46	0.770	0.39
标准归一化＋二阶导数(19 点)	7	0.848	0.43	0.786	0.38

从表 3 中可以得出，平滑预处理后，模型效果没有得到改善；多元散射校正、基线校正预处理后，校正集模型效果得到优化，但预测精度并未得到改善；标准归一化预处理后，模型效果得到很大的改善；一阶导数和二阶导数预处理后，模型效果得到改善，但一阶导数(15 点)模型有过拟合现象，二阶导数(19 点)模型效果较好；将模型效果较好的标准归一化和二阶导数(19 点)预处理方法相结合，模型效果比二阶导数(19 点)预处理后的效果有很小的提高，但和标准归一化预处理后得到的模型精度相比，略差。因此，标准归一化预处理后的 PLSR 模型预测效果优于其他模型，R_c 和 RMSEC 分别为 0.945 和 0.27%，R_p 和 RMSEP 分别为 0.864 和 0.33%。经标准归一化预处理后建立的 PLSR 校正模型和预测结果，见图 4。从表 3 的预测结果和图 4 的离散性得出，模型相关性较好，预测精度得到了很大提升。

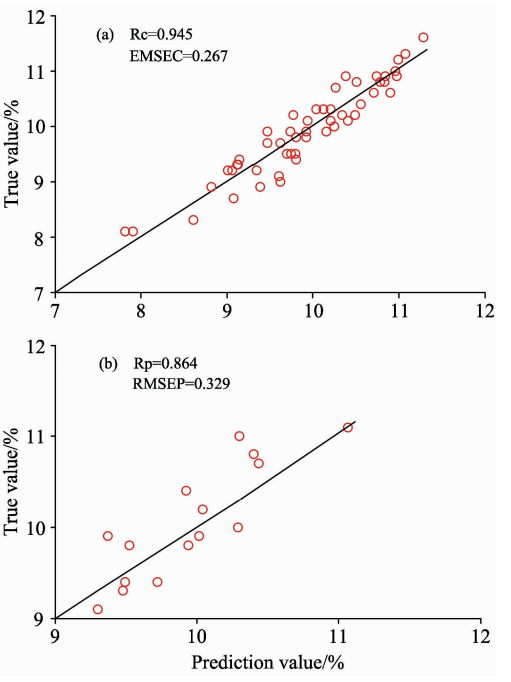

图 4　经标准归一化处理后建立的 PLSR 校正模型(a)和预测结果(b)

Fig. 4　The calibration model (a) and the prediction result (b) of PLSR model by standard normal variable

3　结　论

以小型西瓜为研究对象，采用 SPXY 算法对样品集进行划分，分别建立了西瓜赤道部位、瓜脐部位、瓜梗部位和整果的西瓜可溶性固形物近红外光谱定量分析模型，对比分析得出西瓜赤道部位建立的预测模型精度较好。为提高西瓜赤道部位近红外反射光谱与可溶性固形物含量之间的预测模型精度，采用光谱预处理方法进行优化，结果表明，标准归一化光谱预处理后的 PLSR 模型预测效果优于其他模型，R_c 和 RMSEC 分别为 0.945 和 0.27%，R_p 和 RMSEP 分别为 0.864 和 0.33%。研究结果表明，近红外漫反射光谱可以对西瓜 SSC 含量进行无损快速检测分析，为实际生产中采用近红外漫反射光谱测定西瓜 SSC 含量和低成本小型便携式仪器的研发提供技术储备。

References

[1] JIE Deng-fei, XIE Li-juan, RAO Xiu-qin, et al(介邓飞，谢丽娟，饶秀勤，等). Transactions of the Chinese Society of Agriculture Engineering(农业工程学报), 2013, 29(12): 264.

[2] Jie D F, Xie L J, Fu X P, et al. Journal of Food Engineering, 2013, 118: 387.

[3] HAN Dong-hai, CHANG Dong, SONG Shu-hui, et al(韩东海，常　冬，宋曙辉，等). Transactions of the Chinese Society of Agricultural Machinery(农业机械学报), 2013, 44(7): 174.

[4] QIAN Man, HUAG Wen-qian, WANG Qing-yan, et al(钱　曼，黄文倩，王庆艳，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2016, 36(6): 1700.

[5] Qi S Y, Song S H, Jiang S N, et al. Journal of Innovative Optical Health Sciences, 2014, 7(4): 1350034-1.

[6] Elena T, Stefania C, Irene R, et al. Sensor, 2017, 17(4): 746.

[7] Galvão R K, Araujo M C, José G E, et al. Talanta, 2005, 67(4): 736.

[8] Guo W C, Shang L, Zhu X H, et al. Food & Bioprocess Technology, 2015, 8(5): 1126.
[9] Zhu X H, Fang L J, Gu J S, et al. Food Analytical Methods, 2015, 9(6): 1.

The NIR Detection Research of Soluble Solid Content in Watermelon Based on SPXY Algorithm

WANG Shi-fang[1], HAN Ping[1*], CUI Guang-lu[2], WANG Dong[1], LIU Shan-shan[1], ZHAO Yue[2]
1. Beijing Research Center for Agriculture Standards and Testing, Beijing 100097, China
2. Agricultural Technology Extension Station of Daxing District in Beijing, Beijing 102600, China

Abstract Soluble solid content (SSC), including sugar, acid, fibrin and mineral components, is a comprehensive index for evaluating the fruit maturity and quality, which can affect the taste, flavor and shelf life. Non-destructive and rapid detection of SSC in watermelon is very important for determining the maturity and monitoring the internal quality during storage and transportation, and is helpful to improve production efficiency and market competitiveness of watermelon. For the rapid and non-destructive near infrared (NIR)-based detection of the watermelon SSC, many researchers have used near infrared diffuse transmission method, which requires high light energy and high power transmission, and high power transmission will affect the internal quality. In contrast, the number of researches on near infrared diffuse reflectance method are relatively smaller. It has the advantages of low light energy and low cost, which is in favor of miniaturization and portability of the instruments, and will avoid the fruit quality changes caused by high power transmission. In this study, the greenhouse watermelon was used as the research object, and the near infrared reflectance spectra were collected in the watermelon stem, navel and equator at near 976, 1 186 and 1 453 nm by using JDSU portable near infrared spectrometer. The models between watermelon SSC and near infrared reflectance spectroscopy were established by using partial least square regression (PLSR). Firstly, the sample collection of different parts in the watermelon was divided based on the joint x-y distances (SPXY) method, with SSC as y variables and spectral as x variables. The samples distances were calculated by using x and y variables, and the watermelon samples were divided into 51 calibration sets and 15 prediction sets. The SSC of the calibration sets has a wide distribution range, which covers that of the prediction sets, and can increase the diversity and representativeness of samples and help to build a stable and reliable prediction model. Secondly, the prediction accuracy of quantitative models between the near infrared reflectance spectroscopy and SSC in different detection positions was investigated, and higher correlation and better prediction performance was found in the equator position with prediction correlation coefficient of 0.629 and root mean standard error of prediction of 0.49%. The accuracy of the models between SSC and near infrared spectra information in different watermelon positions was related with the spectrum collection ways and the differences in growing area, variety and maturity. Therefore, the determination of the detection position in the watermelon should be based on the actual situation in the model-building process. Finally, in order to improve the prediction accuracy of the models built for the watermelon equator, the spectra should be pre-processed with the model built for the watermelon equator, and then normalize the results, based on which we can obtain the best prediction model of PLSR. The prediction correlation coefficient was 0.864 and the root mean standard error of prediction was 0.33%, showing higher correlation and improved prediction accuracy. In conclusion, the results indicated that the SSC of the greenhouse watermelon can be accurately predicted based on detecting the equator position by near infrared reflectance spectroscopy. Therefore, it has the potential for improving the rapid and non-destructive testing technology and developing small and portable equipment to detect watermelon SSC by near infrared spectroscopy.

Keywords Watermelon; Near infrared reflectance spectroscopy; SPXY algorithm; Detection position; Soluble solid content

(Received Dec. 25, 2017; accepted Apr. 18, 2018)

* Corresponding author