

不同颜色冬枣可溶性固形物可见-近红外光谱 分析模型构建

郝勇¹, 杜娇君¹, 张书敏², 王起明¹

1. 华东交通大学机电与车辆工程学院, 江西 南昌 330013
2. 南昌海关技术中心, 江西 南昌 330013

摘要 冬枣品质受其品种和生长环境等因素的影响,引起采后化转红指数不同,导致果实的颜色差异较大,从而影响其可溶性固形物(SSC)检测模型的分析精度。采用可见-近红外(Vis-NIR)光谱结合 Norris-Williams 平滑(NWS)、连续小波导数(CWD)、多元散射校正(MSC)、标准正态变量变换(SNV)和 NWS-MSC 五种光谱预处理方法构建不同颜色(红绿相间 MJ, 绿色 GJ 和红色 RJ)冬枣 SSC 的偏最小二乘(PLS)定量分析模型,分别采用 MJ, GJ, RJ, MJ-GJ 和 MJ-GJ-RJ 五个样品集建立冬枣 SSC 的定量分析模型,并采用由 MJ-GJ-RJ 三种颜色冬枣样品组成的测试集进行模型的评价;以不同建模样品集(校正集)的校正相关系数(R_c)和交互验证均方根误差(RMSECV)作为构建最优模型的评价指标;测试集的预测相关系数(R_p)和预测均方根误差(RMSEP)用于模型预测精度的评价。研究结果表明:分别采用 MJ, GJ 和 RJ 的独立样品集进行建模时,模型仅对具有相同颜色的冬枣样品的 SSC 实现了较好的预测;分别在 MJ 样品中加入 GJ 和 GJ-RJ 样品进行 MJ-GJ 和 MJ-GJ-RJ 两个混合样品集的定量模型的构建时, MJ-GJ 模型对 MJ 和 GJ 样品的 SSC 具有较好的预测效果,其 RMSECV, R_c , RMSEP, R_p 分别为 1.108, 0.698, 0.980, 0.724 和 1.108, 0.698, 0.983, 0.822, 而对 RJ 样品的预测误差较大,模型的 RMSECV, R_c , RMSEP, R_p 为 1.108, 0.698, 1.928, 0.597;而 MJ-GJ-RJ 模型对三种颜色的冬枣 SSC 均有较好的预测结果: MJ-GJ-RJ 模型对 MJ 样品的 SSC 模型的 RMSECV, R_c , RMSEP, R_p 为 1.158, 0.796, 1.077, 0.668;对 GJ 样品的 SSC 模型的 RMSECV, R_c , RMSEP, R_p 为 1.158, 0.796, 1.140, 0.841;采用蒙特卡罗无信息变量消除(MCUVE)方法进一步对 MJ-GJ-RJ 样品集光谱的特征变量进行优选后,模型的 R_c 和 R_p 分别由原来的 0.796 和 0.864 提高到 0.884 和 0.922,模型的 RMSECV 和 RMSEP 分别由 1.158 和 0.946 减小到 0.886 和 0.721,模型具有较好的分析精度。采用可见-近红外光谱对不同颜色冬枣的 SSC 进行分析时,当建模集样品与测试集样品颜色属性相似或选择性质相似的建模变量进行模型构建时,模型具有更好的通用性。

关键词 可见-近红外光谱; 偏最小二乘; 冬枣; 可溶性固形物; 蒙特卡罗无信息变量消除

中图分类号: O657 **文献标识码:** A **DOI:** 10.3964/j.issn.1000-0593(2021)11-3385-07

引言

冬枣又名冻枣、雁来红或苹果枣,是目前公认的鲜食优质栽培品种。冬枣营养极其丰富,含有天门冬氨酸、苏氨酸、丝氨酸等 19 种人体必需的氨基酸,且维生素 C 的含量尤其丰富,有“活维生素丸”之美誉,营养价值为百果之冠^[1]。冬枣的可溶性固形物(soluble solids content, SSC)是其重要的

内部品质和成熟度评价指标。冬枣生长环境的温度、水分、光照、风和土壤等因素均会对其采后转红指数产生影响,造成冬枣颜色和 SSC 差异,而消费者常会将颜色与其 SSC 进行关联来判断冬枣的口感。因此,对于不同颜色冬枣 SSC 含量的研究是非常必要的。

SSC 含量的测定方法主要分为折射法和光谱分析法,折射法的主要缺点是需破坏被测样品;而可见-近红外(visible and near-infrared spectroscopy, Vis-NIR)光谱技术是一种快

收稿日期: 2020-10-28, 修订日期: 2021-03-02

基金项目: 国家自然科学基金项目(31960497, 21265006)资助

作者简介: 郝勇, 1978 年生, 华东交通大学机电与车辆工程学院副教授 e-mail: haonm@163.com

速、无损、绿色的现代检测技术,根据全波段或者特征波长给出的信息与样品属性或含量间的相关关系对样品进行定性和定量分析。然而,由于 Vis-NIR 光谱常含有较多冗余信息及干扰,影响了模型的精度和稳定性,因此,需要进行光谱预处理及变量筛选,实现对干扰信息的抑制或滤除,从而提高后续模型的分析精度。李勇等讨论了油菜籽样品颜色对波长为 700~2 500 nm 谱区的近红外光谱分析准确度的影响,认为样品表面颜色越深,吸光度越大,在短波处最为明显^[2];姚鑫森等研究了玉米子粒颜色对 850~1 050 nm 的光谱波段的淀粉含量的影响,结果表明样品的外观颜色是影响其分析精度的因素之一,采用平滑求导等光谱预处理方法可提高模型的预测性能^[3]。郭成等采用无信息变量消除(UVE)方法对无花果 SSC 的 PLS 模型建模变量进行优选,其预测相关系数可达到 0.89^[4]。Tian 等采用光谱预处理和随机森林变量选择方法对苹果 SSC 在线预测模型进行优化,变量减少为原始变量的 86%,而 RMSEP 减小了 17%^[5]。

本工作采用 Vis-NIR 光谱结合 Norris-Williams 平滑(Norris-Williams smoothing, NWS)、连续小波导数(continuous wavelet derivative, CWD)、多元散射校正(multiplicative scattering correction, MSC)、标准正态变量变换(standard normal variate, SNV)和 NWS-MSC 五种预处理方法^[6-8]和蒙特卡罗无信息变量消除(Monte Carlo uninformative variable elimination, MCUVE)方法^[8]对不同颜色(红绿相间 MJ, 绿色 GJ 和红色 RJ)冬枣样品的 SSC 进行偏最小二乘(partial least squares, PLS)定量分析模型的构建和优化,以期实现不同颜色冬枣 SSC 的准确分析^[9]。

1 实验部分

1.1 材料

样品(大荔冬枣)来源于陕西某枣园,冬枣样品如图 1 所示。采集了同一批次、取大小均等且表皮颜色不同的三种冬枣(红绿相间 MJ, 绿色 GJ 和红色 RJ)各 45 个样品,共计 135 个。采用冷库保鲜处理,用蓄冷箱送至实验室。



图 1 不同颜色大荔冬枣样品

Fig. 1 Samples of Dali jujube with different colours

1.2 光谱采集

光谱采集装置如图 2 所示,采用透射方式进行光谱的采集, Ocean Optics USB2000 Vis-NIR 光谱仪(中国海洋光学公司)用于冬枣光谱的检测, 2 盏 100 W 的飞利浦卤素灯作为光源, 光谱范围为 340~1 027 nm, 光谱分辨率约为 0.4 nm。

光谱采集前,将三种颜色的冬枣样品从冷藏室取出并置

于室温下 8 h,相对湿度为 25%~30%;采集距离为 10 cm,随机翻转冬枣位置 5 次进行光谱的采集,5 次测量的平均光谱作为最终的分析光谱。采用 SpectraSuite 软件进行光谱数据的采集和存储。

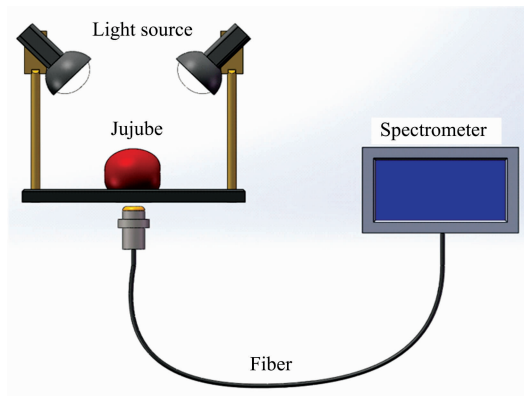


图 2 光谱采集装置示意图

Fig. 2 Schematic diagram of spectrum acquisition device

三种颜色冬枣样品原始光谱的平均光谱如图 3 所示,从图中可知,不同颜色的冬枣样品光谱谱形相似,主要不同之处在于光谱响应强度间的差异。此外,光谱图不仅包含其组分信息,还包括光谱仪相关器件引入的噪声、杂散光和样品背景、颜色等干扰信息引起的光谱反射率的变化,因此,需要对光谱进行预处理,以提高后续模型的分析精度。

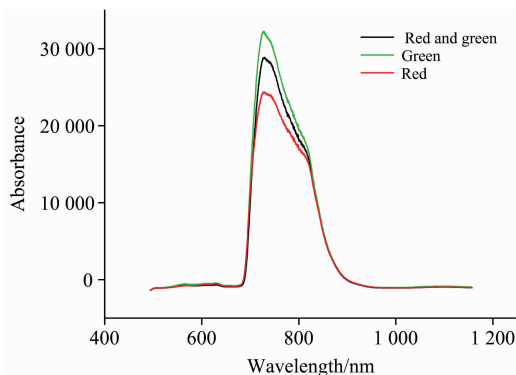


图 3 冬枣样品平均光谱图

Fig. 3 Average spectra of jujube samples

1.3 模型的建立与评价

每个冬枣样品均匀地选择 5 个点进行 SSC 的测试,其平均值作为该样品的最终 SSC 的标准参考值。采用 SSC 均匀分布的方式将样品集划分为校正集和测试集,每种颜色冬枣样品校正集和测试集样品数量比为 2:1,即每种颜色的 45 个冬枣样品中,30 个样品用于建模,15 个样品用于模型的测试。

为研究不同颜色冬枣 SSC 构建的 PLS 模型的分析精度,分别建立不同颜色冬枣 SSC 的单一模型和混合模型。单一模型即单种颜色冬枣 SSC 建模,混合模型即两种或三种颜色冬枣 SSC 混合建模。为了描述方便,单一模型校正集 MJ_c, GJ_c 和 RJ_c 分别表示建模样品集是为红绿相间冬枣、绿色冬枣和

红色冬枣；混合模型校正集 MJ_c-GJ_c 和 MJ_c-GJ_c-RJ_c 分别表示建模样品是由红绿相间冬枣-绿色冬枣和三种颜色的冬枣组成；MJ_v, GJ_v, RJ_v 和 MJ_v-GJ_v-RJ_v 分别表示测试样品集

是由红绿相间冬枣、绿色冬枣、红色冬枣和三种颜色冬枣组成；冬枣样品 SSC 的统计信息如表 1 所示。

表 1 冬枣样品及其 SSC 含量统计信息

Table 1 Statistics of SSC contents in jujube samples

Type of color	Calibration set	¹ n _c	SSC(°Brix)	Prediction set	² n _v	SSC(°Brix)
MJ	MJ _c	30	10.93~15.67	MJ _v	15	10.93~15.67
GJ	GJ _c	30	11.00~17.10	GJ _v	15	11.50~16.90
RJ	RJ _c	30	10.60~19.20	RJ _v	15	11.10~18.30
MJ-GJ	MJ _c -GJ _c	60	10.93~17.10	MJ _v -GJ _v	30	10.93~16.90
MJ-GJ-RJ	MJ _c -GJ _c -RJ _c	90	10.60~19.20	MJ _v -GJ _v -RJ _v	45	10.93~18.30

注: ¹n_c: 建模集样本数; ²n_v: 测试集样本数

Note: ¹n_c is the number of samples for modeling set and ²n_v is the number of samples for test set

采用模型的校正相关系数 (correlation coefficient of calibration set, R_c) 和交叉验证均方根误差 (root mean square error of cross-validation, RMSECV) 作为评价指标对光谱的预处理方法和变量进行优选, 采用测试集的预测相关系数 (correlation coefficient of prediction set, R_p) 和预测均方根误差 (root mean square error of prediction, RMSEP) 作为模型预测精度的评价指标。其中 R_c 越大, RMSECV 越小, 模型的精度越高; R_p 越大, RMSEP 越小, 模型的预测能力越好。Matlab R2018a 用于光谱数据的预处理、变量优选以及 PLS 模型的建立。

2 结果与讨论

2.1 不同预处理方法对冬枣 SSC 模型精度的影响

样品光谱采集时易受光谱仪噪声和杂散光等因素的干扰, 引起光谱反射率的变化。因此, 需要对样品光谱进行预处理, 抑制或滤除干扰信息以提高模型的分析精度。为了研究不同预处理方法对冬枣 SSC 模型精度的影响, 分别采用 NWS, CWD, MSC, SNV 和 NWS-MS 五种预处理方法对光谱进行信息变换和提取, 并计算其 R_c 和 RMSECV 对模型精度进行评价, PLS 模型的校正结果如表 2 所示。由表可知, 校正集为 MJ_c, GJ_c, RJ_c, MJ_c-GJ_c, MJ_c-GJ_c-RJ_c 建立的五种 PLS 模型分别选用 NWS, CWD, NWS, NWS-MS 和 SNV 方法进行预处理时, 模型的精度均有明显提升。

光谱的吸收主要反映了有机物中 C—H, O—H 和 N—H 等含氢基团信息, 而 SSC 主要是指可溶性糖类, 包含重要的一OH 基团信息。在冬枣样品 Vis-NIR 光谱中, 波长 725 nm 处具有明显的由于 O—H 键伸缩振动产生的吸收峰。图 4(a) 为采集的三种颜色共 135 个冬枣样品的原始可见-近红外光谱图, 图 4(b) 为采用优选的 SNV 方法进行预处理后的光谱图。由图可知, 预处理后在一定程度上消除了原始光谱中的背景噪声及其他影响。

2.2 建模样品集组成对冬枣 SSC 的 PLS 模型预测精度的影响分析

2.2.1 单一颜色冬枣 SSC 模型分析

表 2 不同预处理方法的冬枣可溶性固形物 PLS 模型的校正结果

Table 2 Modelling results of PLS models for jujube SSC with different pretreatment methods

Samples	Pretreatment methods	Factor	RMSECV	R _c
MJ _c	RAW	8	0.898	0.700
	NWS	14	0.843	0.756
	CWD	12	0.948	0.672
	MSC	10	0.958	0.652
	SNV	10	0.969	0.643
	NWS-MS	11	0.896	0.722
GJ _c	RAW	19	0.825	0.854
	NWS	10	0.718	0.893
	CWD	9	0.659	0.910
	MSC	9	0.774	0.873
	SNV	20	0.739	0.884
	NWS-MS	10	0.717	0.897
RJ _c	RAW	7	1.268	0.816
	NWS	7	1.211	0.835
	CWD	6	1.337	0.796
	MSC	8	1.376	0.784
	SNV	8	1.325	0.799
	NWS-MS	8	1.340	0.801
MJ _c -GJ _c	RAW	6	1.362	0.496
	NWS	10	1.329	0.578
	CWD	8	1.245	1.606
	MSC	8	1.124	0.676
	SNV	8	1.132	0.671
	NWS-MS	11	1.108	0.698
MJ _c -GJ _c -RJ _c	RAW	10	1.351	0.786
	NWS	11	1.318	0.740
	CWD	8	1.255	0.757
	MSC	8	1.187	0.785
	SNV	8	1.156	0.796
	NWS-MS	10	1.168	0.793

由于三种颜色的冬枣样品光谱谱形相似, 可用不同颜色冬枣样品的预测集对冬枣 SSC 模型进行评价。分别对 MJ_c,

GJ_c 和 RJ_c 样品集进行建模, 采用 MJ_v , GJ_v 和 RJ_v 三个测试集样品对模型的预测能力进行评价。单一颜色冬枣样品 SSC 的 PLS 模型的预测结果如表 3 所示。

由表 3 可知, 对于单一颜色冬枣的 PLS 模型, 当预测集样品与建模集样品的颜色性质相同时, 其模型的预测结果较好, 即校正集 MJ_c 中, MJ_v 为预测集时预测结果最佳; 校正集 GJ_c 中, GJ_v 为预测集时预测结果最佳; 校正集 RJ_c 中,

RJ_v 为预测集时预测结果最佳。

2.2.2 混合颜色冬枣 SSC 模型分析

分别对混合颜色 MJ_c - GJ_c 和 MJ_c - GJ_c - RJ_c 样品集进行建模, 采用 MJ_v , GJ_v , RJ_v 和 MJ_v - GJ_v - RJ_v 四个测试集样品对模型的预测能力进行评价。混合颜色冬枣样品 SSC 的 PLS 模型的预测结果如表 4 所示。

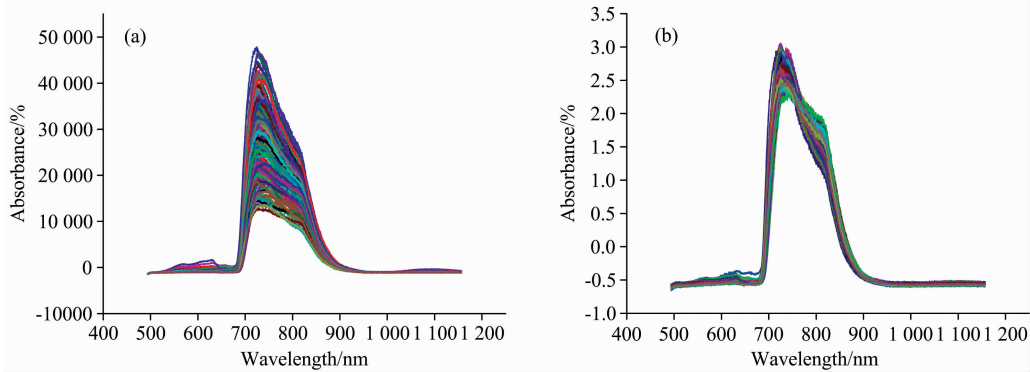


图 4 冬枣样品的可见-近红外光谱图与 SNV 预处理后的光谱图

(a): 原始可见-近红外光谱; (b): SNV 预处理后可见-近红外光谱

Fig. 4 Original Vis-NIR spectra of Jujube samples and spectra after pretreatment with SNV

(a): Original Vis-NIR spectra; (b): Spectra after pretreatment with SNV

表 3 单一颜色冬枣样品 SSC 的 PLS 模型的预测结果

Table 3 Predicted results of PLS models for SSC of single color jujube

Samples for modeling	Pretreatment methods	Factor	RMSECV	R_c	Prediction set	RMSEP	R_p
MJ_c	NWS	14	0.843	0.752	MJ_v	1.314	0.816
					GJ_v	2.112	0.657
					RJ_v	2.792	0.405
GJ_c	CWD	9	0.659	0.910	MJ_v	1.769	0.611
					GJ_v	0.875	0.822
					RJ_v	1.276	0.812
RJ_c	NWS	8	1.211	0.835	MJ_v	1.517	0.736
					GJ_v	1.578	0.831
					RJ_v	0.960	0.923

表 4 混合颜色冬枣样品 SSC 的 PLS 模型的预测结果

Table 4 Predicted results of PLS models for SSC of mixed color jujube samples

Samples for modeling	Pretreatment methods	Factor	RMSECV	R_c	Prediction set composition	RMSEP	R_p
MJ_c - GJ_c	NWS-MSD	11	1.108	0.698	MJ_v	0.980	0.724
					GJ_v	0.983	0.822
					RJ_v	1.928	0.597
					MJ_v - GJ_v - RJ_v	1.378	0.704
MJ_c - GJ_c - RJ_c	SNV	8	1.158	0.796	MJ_v	1.077	0.668
					GJ_v	0.881	0.861
					RJ_v	1.140	0.841
					MJ_v - GJ_v - RJ_v	0.946	0.864

由表 4 可知, 当预测集中只含有与校正集相同颜色的样品时, 其预测结果较好, 即校正集为 MJ_c - GJ_c 时, 模型对 MJ_v 和 GJ_v 的预测精度较 RJ_v 的好; 用预测集为 MJ_v - GJ_v - RJ_v 评价模型时, 虽然校正集 MJ_c - GJ_c 中含有红绿相间和绿

色冬枣样品, 但无红色冬枣样品, 因此, 预测精度较差; 当在 MJ_c - GJ_c 中增加 RJ_c 组成 MJ_c - GJ_c - RJ_c 建模集, 模型对 MJ_v - GJ_v - RJ_v 的预测误差由 1.378 减小到 0.946。

图 5 所示为采用不同颜色的冬枣样品组成的五个校正集

建立冬枣 SSC 的 PLS 模型对由三种颜色冬枣样品组成的测试集进行预测的 RMSEP 变化图, 由图中可知, 当建模集样品与测试集样品性质相同时, 模型均得到较好的预测结果,

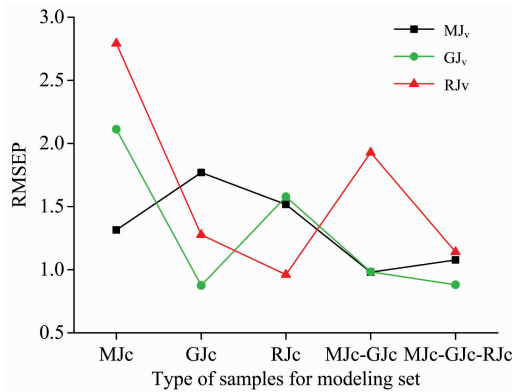


图 5 不同颜色样品建模集建立的冬枣 SSC 的 PLS 模型的 RMSEP 变化图

Fig. 5 RMSEP variation diagram of PLS models for jujube SSC established by samples with different colours for modeling sets

RMSEP 均比较小。

2.3 变量筛选方法对模型预测精度的影响分析

通过对校正集 MJ_c, GJ_c 和 RJ_c 的单一模型, 及 MJ_c-GJ_c, MJ_c-GJ_c-RJ_c 的混合模型经预处理方法优选后采用 PLS 建模分析可知, 当测试集样品与校正集样品性质相似时, 模型具有更好的预测能力。为了进一步对模型进行优化, 采用 MCUVE 方法进行变量选择后, 应用所筛选的特征变量建立 PLS 模型的结果如表 5 所示^[10]。

由表 5 中可知, 冬枣样品的 Vis-NIR 光谱包含有 2 048 个建模变量, 选用 MCUVE 变量筛选后, 模型的建模变量均减少, 且精度得到不同程度的提高。对于冬枣校正集为 MJ_c-GJ_c-RJ_c 模型, 优化后的模型保留了 200 个有效建模变量。图 6 是校正集为 MJ_c-GJ_c-RJ_c 时样品经过 MCUVE 选择的波长信息, 选择的变量主要集中于近红外短波光谱区域(770~1 100 nm), 剔除大部分位于可见光谱区域(390~770 nm)的样本点, 从而减少了表面颜色对分析模型的影响, 大大提高了不同颜色冬枣混合建模时变量的相似性, 从而使模型的精度及预测能力得以提高, 模型的 RMSECV 从 1.158 降低到 0.886, RMSEP 从 0.946 降低到 0.721。

表 5 变量优选后模型的分析结果

Table 5 Analysis results of the models with modelling variables optimization

Calibration set	Selected the number of variables	Factor	RMSECV	R _c	Prediction set	RMSEP	R _p
MJ _c	2 040	13	0.850	0.748	MJ _v	1.070	0.824
GJ _c	340	9	0.515	0.947	GJ _v	0.443	0.958
RJ _c	180	11	0.725	0.943	RJ _v	0.980	0.947
MJ _c -GJ _c	160	14	0.903	0.812	MJ _v	0.797	0.848
	300	14	0.808	0.851	GJ _v	0.605	0.933
MJ _c -GJ _c -RJ _c	200	10	0.886	0.884	MJ _v -GJ _v -RJ _v	0.721	0.922

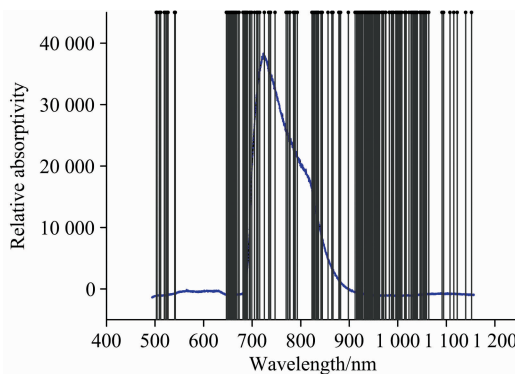


图 6 MCUVE 方法选择变量的分布

Fig. 6 Distribution of selected variables by MCUVE method

2.4 冬枣 SSC 最优光谱分析模型

不同颜色的冬枣样品进行混合建模时, 通过光谱预处理和变量优选方法可以得到预测精度较高的分析模型。图 7 所示为利用优选变量建立的混合冬枣 PLS 模型的 RMSECV 和 RMSEP 随主成分数的变化图, 从图中可以, 两者相差较小且变化规律具有较好的一致性, 表明模型拟合的较合理^[11]。

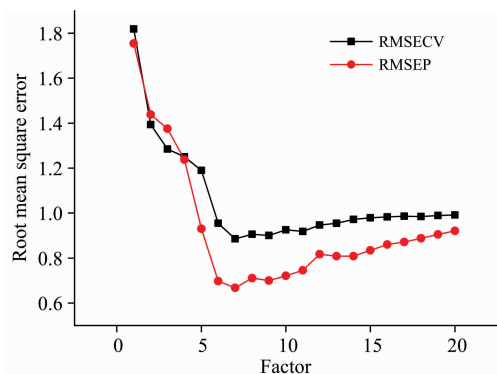


图 7 混合冬枣模型的 RMSECV 和 RMSEP 随 PLS 因子数变化图

Fig. 7 Variations of RMSECV and RMSEP of mixed jujube model with PLS factors

最优混合模型对测试集样品 SSC 的预测值和参考值的相关关系如图 8 所示。由图可知, 冬枣 SSC 的参考值和模型预测值之间存在较好的相关关系, 预测样品集的 R_p 和 RMSEP 分别为 0.922 和 0.721。

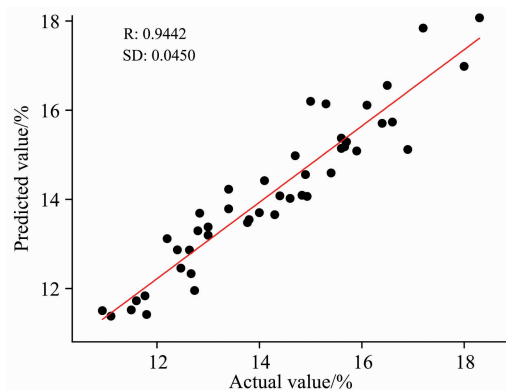


图 8 SSC 实测值与模型预测值的相关关系

Fig. 8 Correlation of SSC measured value and model predicted value

3 结 论

对不同颜色冬枣 SSC 的 Vis-NIR 光谱分析模型构建进行研究, 采用不同的光谱预处理方法和 MCUVE 变量筛选方法对冬枣 SSC 的 PLS 定量分析模型进行优化, 以期实现不同颜色冬枣 SSC 含量准确分析的通用模型的构建。结果表明, 分别采用 MJ, GJ 和 RJ 的独立样品集进行建模时, 模型仅仅对具有相同颜色的冬枣样品的 SSC 实现了较好的预测; 分别在 MJ 样品中加入 GJ 和 GJ-RJ 样品进行 MJ-GJ 和 MJ-GJ-RJ 两个混合样品集的定量模型的构建时, MJ-GJ 模型对 MJ 和 GJ 样品的 SSC 具有较好的预测效果, 而对 RJ 样品的预测误差较大; MJ-GJ-RJ 模型对三种颜色的冬枣 SSC 均有较好的预测结果; 采用 Vis-NIR 光谱对不同颜色冬枣的 SSC 进行分析时, 当校正集样品与测试集样品颜色属性相似进行模型构建时, 模型具有更好的通用性和分析精度。采用 MCUVE 方法对模型进行变量优选后, 模型的各项评价指标均得到改善。

References

- [1] CHEN Chong, LUO Hua-ping, SUO Yu-ting, et al(陈 冲, 罗华平, 索玉婷, 等). Agriculture and Technology(农业与技术), 2019, 39(5): 1.
- [2] LI Yong, WEI Yi-min, WANG Feng(李 勇, 魏益民, 王 锋). Journal of Nuclear Agriculture(核农学报), 2005, (3): 236.
- [3] YAO Xin-miao, LU Shu-wen, XIE Tie-min, et al(姚鑫淼, 卢淑雯, 解铁民, 等). Corn Science(玉米科学), 2013, 21(4): 153.
- [4] CAO Yong, JIAO Xue-jun, JIANG Jin, et al(曹 勇, 焦学军, 姜 劲, 等). Acta Optica Sinica(光学学报), 2018, 38(3): 369.
- [5] Tian Xi, Fan Shuxiang, Li Jiangbo, et al. Infrared Physics &. Technology, 2019, 102: 979.
- [6] ZHANG Li-juan, XIA Qi-le, CHEN Jian-bing, et al(张丽娟, 夏其乐, 陈剑兵, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2020, 40(7): 2246.
- [7] LI Zhuang-zhuang, WU Qiang-shui, HUANG Sha. Science Technology and Engineering(科学技术与工程), 2020, 20(10): 4061.
- [8] HAO Yong, WU Wen-hui, SHANG Qing-yuan, et al(郝 勇, 吴文辉, 商庆园, 等). Acta Optica Sinica(光学学报), 2019, 39(9): 0930004.
- [9] Mao Zhiyi, Shan Ruifeng, Wang Jiajun, et al. Pubmed, 2014, 128: 711.
- [10] ZHANG He-dong, WU Jing-zhu, HAN Ping, et al(张鹤冬, 吴静珠, 韩 平, 等). Journal of Food Safety and Quality Inspection(食品安全质量检测学报), 2019, 10(1): 209.
- [11] Parichat Theanjumpol, Kumpon Wongzeewasakun, Nadthawat Muenmanee, et al. Elsevier B. V., 2019, 153: 257.

Research on Construction of Visible-Near Infrared Spectroscopy Analysis Model for Soluble Solid Content in Different Colors of Jujube

HAO Yong¹, DU Jiao-jun¹, ZHANG Shu-min², WANG Qi-ming¹

1. School of Mechatronics & Vehicle Engineering, East China Jiaotong University, Nanchang 330013, China

2. Nanchang Customs Technology Center, Nanchang 330013, China

Abstract The quality of jujube is susceptible to factors such as the environment, causing changes in its post-harvest redness index, leading to large differences in fruit color, which affects the analysis accuracy of its soluble solids content (SSC) detection model. Visible-near infrared spectroscopy (Vis-NIRs) combined with spectral preprocessing methods including Norris-Williams smoothing (NWS), continuous wavelet derivative (CWD), multiplicative scattering correction (MSC), standard normal variate (SNV) and NWS-MSC were used to build the partial least squares (PLS), quantitative analysis models of the SSC of jujube, with different colors (red and green-MJ, green-GJ and red-RJ). Five independent sample sets, including MJ, GJ, RJ, MJ-GJ and MJ-GJ-RJ, were used to establish the quantitative analysis models of SSC for jujube, and test set samples MJ-GJ-RJ were used for model evaluation. The correlation coefficient of calibration set (R_c) and the root mean square error of cross-validation (RMSECV) were used to evaluate model accuracy. The correlation coefficients of prediction (R_p) and the root mean square error for prediction (RMSEP) were used to evaluate model prediction accuracy. The research results showed that when the independent sample sets of MJ, GJ and RJ were used for modeling, the models only achieved a better prediction for the SSC of jujube samples with the same color, respectively. When adding GJ and GJ-RJ samples to the MJ samples to construct the quantitative model of the two mixed sample sets, including MJ-GJ and MJ-GJ-RJ. The MJ-GJ model had better prediction results of SSC for MJ and GJ jujube samples, the model's RMSECV, R_c , RMSEP, and R_p were 1.108, 0.698, 0.980, 0.724 and 1.108, 0.698, 0.983, 0.822, respectively, but the effect of RJ samples was relatively larger, the model's RMSECV, R_c , RMSEP, R_p were 1.108, 0.698, 1.928, 0.597. The MJ-GJ-RJ model obtained good prediction results of SSC for the three colors jujube: for the SSC model of MJ, the RMSECV, R_c , RMSEP, R_p of the MJ-GJ-RJ model were 1.158, 0.796, 1.077, 0.668; for the SSC model of GJ, the model's RMSECV, R_c , RMSEP, R_p were 1.158, 0.796, 0.881, 0.861; for the SSC model of RJ, the model's RMSECV, R_c , RMSEP, R_p were 1.158, 0.796, 1.140, 0.841. After using the Monte Carlo uninformative variable elimination (MCUVE) method to optimize the variables of the MJ-GJ-RJ model further, the R_c and R_p were increased from 0.796 and 0.864 to 0.884 and 0.922, respectively. The RMSECV and RMSEP were reduced from 1.158 and 0.946 to 0.886 and 0.721, respectively. The model has better analysis accuracy. When the SSC of different color jujube was analyzed using near-infrared spectroscopy, similar sample set properties for calibration and prediction or modeling variables are required to construct universality models.

Keywords Visible-near infrared spectroscopy; Partial least squares; Jujube; Soluble solids content; Monte Carlo uninformative variable elimination

(Received Oct. 28, 2020; accepted Mar. 2, 2021)