

基于现场通用模型的3种蜜炙饮片水分属性动态在线监测方法研究

张 瀚^{1,2}, 王文哲^{1,2}, 胡小艳^{1,2}, 王 静^{1,2}, 韩燕雨^{1,2}, 王晓萌^{1,2}, 张晓梦^{1,2},
郭新雨^{1,2}, 郇星月^{1,2}, 赵 静^{1,2}, 李 楠^{1,2}, 王逸飞^{3*}, 吴志生^{1,2*}

(1. 北京中医药大学中药学院, 北京 102488; 2. 中药制药与新药开发教育部工程研究中心, 北京 102488;
3. 北京中医药大学房山医院, 北京 102499)

摘要: 针对中药制造过程原料单元蜜炙饮片水分含量关键质量属性离线静态检测具有滞后性及破坏性等问题, 以蜜炙款冬、蜜炙黄芪、蜜炙甘草为载体, 采用烘干法测定水分含量作为参考值, 采用运动载物台模拟实际现场传输带生产过程样品运动过程, 采集近红外 (near infrared, NIR) 光谱, 结合机器学习方法, 建立多品种蜜炙饮片水分含量 NIR 现场动态检测模型。结果表明, 采用二阶导数法对光谱进行预处理, 决策树数量 (ntree)、随机特征数量 (max feature)、生成叶节点的最小样本数 (node size) 分别选择 46、76、8 时, 建立的水分含量定量分析模型效果最优。模型的预测决定系数 (the prediction coefficient of determination, R^2_{pre}) 及预测均方根误差 (root mean square error of prediction, RMSEP) 分别为 0.903 2 和 0.330 2。本研究建立的多品种蜜炙饮片水分含量 NIR 通用定量模型具有良好的预测性能, 可实现同时对蜜炙款冬、蜜炙黄芪、蜜炙甘草水分含量的快速检测, 为中药制造过程原料单元蜜炙饮片的水分测定提供方法。

关键词: 蜜炙饮片; 快速检测; 通用模型; 水分关键质量属性; 现场动态检测

中图分类号: R284 文献标识码: A 文章编号: 0513-4870(2023)10-2890-10

Research on dynamic on-line monitoring method of moisture attribute in three honey-processed Chinese herbal slice based on *in-situ* general model

ZHANG Han^{1,2}, WANG Wen-zhe^{1,2}, HU Xiao-yan^{1,2}, WANG Jing^{1,2}, HAN Yan-yu^{1,2},
WANG Xiao-meng^{1,2}, ZHANG Xiao-meng^{1,2}, GUO Xin-yu^{1,2}, HUAN Xing-yue^{1,2},
ZHAO Jing^{1,2}, LI Nan^{1,2}, WANG Yi-fei^{3*}, WU Zhi-sheng^{1,2*}

(1. School of Chinese Materia Medica, Beijing University of Chinese Medicine, Beijing 102488, China; 2. Engineering Research Center of Chinese Medicine Production and New Drug Development, Beijing 102488, China; 3. Fangshan Hospital, Beijing University of Chinese Medicine, Beijing 102499, China)

Abstract: Aiming at the hysteresis and destructiveness of off-line static detection of critical quality attribute of the moisture content of the raw material unit of the traditional Chinese medicine manufacturing process, honey-processed *Tussilago farfara*, honey-processed *Astragalus* and honey-processed *Glycyrrhiza uralensis* were used as the research carriers, and the drying method was used to measure the moisture content as a reference value. The moving stage was used to simulate the movement process of samples on the conveyor belt in the actual on-site

收稿日期: 2022-12-21; 修回日期: 2023-02-28.

基金项目: 国家自然科学基金优秀青年科学基金项目 (82022073); 国家自然科学基金项目 (82274110); 广东省重点领域研发计划项目 (2020B1111120002); 中央高校基本科研业务费北京中医药大学揭榜挂帅项目 (2022-JYB-JBZR-019, 2022-JYB-JBZR-018); 国家医学攻关产教融合创新平台-中药智能制造工程平台项目 (90010062820031).

*通讯作者 Tel: 86-10-53912101, E-mail: wzs@bucm.edu.cn; wenifread@163.com

DOI: 10.16438/j.0513-4870.2022-1383

production process, and near-infrared (NIR) spectra were collected, combined with machine learning, to establish NIR on-site dynamic detection model of moisture content in multi-variety honey-processed Chinese herbal slice. The results show that the second derivative method is used to preprocess the spectrum. The number of decision trees (ntree), the number of random features (max feature), and the minimum number of samples for generating leaf nodes (node size) are selected: 46, 76, and 8, respectively. The quantitative analysis model of moisture content has the best effect. The prediction coefficient of determination (the prediction coefficient of determination, R^2_{pre}) and the root mean square error of prediction (root mean square error of prediction, RMSEP) of the model were 0.903 2 and 0.330 2, respectively. The NIR quantitative model for the moisture content of multi-variety honey-processed Chinese herbal slice established in this study has good predictive performance, and can achieve rapid, accurate and non-destructive quantitative analysis of the moisture content of honey-processed *Tussilago farfara*, honey-processed *Astragalus* and honey-processed *Glycyrrhiza uralensis* at the same time, and provides a method for determining the moisture content of honey-processed Chinese herbal slice of the raw material unit of the traditional Chinese medicine manufacturing process.

Key words: honey-processed Chinese herbal slice; rapid detection; generic model; moisture critical quality attribute; on-site dynamic monitoring

中药制造质量可控性是制约中药现代化发展的关键问题之一。如何实现中药制造过程在线实时监测或现场检测、保证产品质量稳定均一,对于实现中药产业化具有重大现实意义^[1]。中药制造原料单元是中药制造过程的首要单元。水分含量是控制中药制造产品稳定均一的关键质量属性^[2]。水分含量过高会引起中药材及其提取物和成品霉变、虫蛀,严重影响药效,甚至对患者造成危害^[3]。蜜炙法是中药原料炮制过程中较为常见的方法。蜜炙饮片炮制过程其主观判断的成分较多,时间、火候、加蜜量等变量没有具体参数,导致不同批次的炮制品水分含量存在较大差异^[4]。因此,蜜炙饮片水分含量的有效控制对保证中药原料质量起着重要作用,相比饮片使用时采用烘箱干燥法等静态滞后检测方法耗时较长、破坏性较大的缺点,亟待建立一种可现场、通用、快速动态检测蜜炙饮片水分含量的方法。

近红外 (near infrared, NIR) 光谱法具有快速、无损、可进行动态现场检测的优势,现已广泛应用于食品^[5,6]、中药^[7-10]等各个领域原料及产品的水分预测。近年来,研究人员通过NIR技术结合化学计量学等分析方法,针对饮片的水分含量预测^[11]做了大量研究工作, Lv等^[12]将西洋参粉碎过筛后采集NIR光谱,采用偏最小二乘 (partial least squares, PLS) 方法,建立了西洋参的水分含量NIR定量模型。Shi等^[13]将黄柏、知母粉碎过筛后采集NIR光谱,分别建立了黄柏、知母的水分含量NIR定量模型。但这些研究主要针对单个品种饮片水分含量的离线静态检测,且光谱采集需对饮片进行粉碎过筛等预处理,具有一定的破坏性,无法实现实际生产过程中的现场动态检测,缺少针对多品种蜜炙饮片的水分含量现场动态检测方法研究。

本研究为解决中药制造过程原料单元中蜜炙饮片

多品种水分含量的在线实时监测问题,以炙甘草、炙款冬、炙黄芪3种蜜炙饮片为载体,采用运动载物台模拟实际现场传输带生产过程样品运动过程,采用NIR光谱技术结合集成学习,建立了3种蜜炙饮片的水分定量现场动态检测通用模型,以提升NIR定量模型的泛化性,并与PLS模型进行性能比较,为实现中药制造过程原料质量的稳定均一可控,提供一种现场、快速、无损动态多品种蜜炙饮片检测水分含量的方法。

材料与amp;方法

仪器 DA7440近红外在线分析仪(珀金埃尔默股份有限公司); Unscrambler X10.4光谱分析软件(挪威CAMO公司); MATLAB光谱分析软件(version 7.0, MathWorks公司); 电热鼓风干燥箱(DHG-9070A,上海一恒科学仪器有限公司)。

材料 28批蜜炙甘草饮片、21批蜜炙款冬饮片、25批蜜炙黄芪饮片的具体批号信息见附表1。所有蜜炙饮片由北京市房山区房山中医院提供。

样品NIR光谱的采集 每批样品取样3次,每次随机抽选饮片样本20g,采集3次光谱,将样品平铺于旋转式运动载物台的装样盘中,通过运动载物台旋转来模拟实际生产过程中传送带运动过程,进行NIR扫描,采集载物台运动1周后的光谱(20s),计算平均值作为取样样品光谱。采用DA7440近红外在线分析仪进行光谱采集,波长范围:950~1650nm,扫描次数:30次,分辨率:3.125nm。

样品水分值的测定 精密称量3g样品,将扫描NIR后的3种蜜炙饮片按照《中华人民共和国药典》0832水分测定法第二法(烘干法)的要求测定样品的

水分值。

校正集和验证集样品的选择 采用 Unscramble 软件剔除离群值, 基于 MATLAB 软件, 采用 K-S (Kennard-Stone) 样本划分方法, 在样本中根据欧氏距离以 3:1 比例划分校正集与预测集。校正集用于建立模型, 预测集用于验证模型。

光谱预处理方法考察 在 NIR 采集过程中, 光谱易受背景噪音、环境温湿度、样品自身物理变化等因素的干扰, 导致基线漂移, 因此在建立 PLS 定量模型前, 需采用合理的光谱预处理方法对样品的原始吸收光谱进行预处理, 以减小混杂因素对模型的干扰。选择原始光谱、Savitzky-Golay 平滑法 (Savitzky-Golay smoothing, SG)、标准正则变换 (standard normalization variate, SNV)、归一化 (normalization)、标准化 (autoscaling)、多元散射校正 (multiple scatter correction, MSC)、一阶导数法 (first derivative, 1D)、二阶导数法 (second derivative, 2D) 对样品的原始吸收光谱进行预处理后建立 PLS 水分含量定量模型, 以交叉验证均方根误差 (root mean square error of cross-validation, RMSECV) 为模型性能评价指标, 考察不同预处理方法对模型的影响。

光谱波段筛选方法考察 NIR 吸收重叠较严重, 导致光谱中含有大量冗余信息。如在建立 PLS 定量模型时进行全谱计算, 不仅工作量大, 且会严重降低所建模型的预测精度。因此需采用合理的光谱波段筛选方法对样品的原始光谱进行波段筛选, 以提高 PLS 定量模型的模型预测精度。在最佳预处理方法的基础上, 采用变量重要性投影 (variable importance in the projection, VIP)、移动窗口偏最小二乘法 (moving windows partial least squares, MWPLS)、竞争-自适应重加权抽样 (competitive adaptive reweighted sampling, CARS)、逐步回归法 (stepwise regression, SR)、无信息变量消除法 (uninformative variable elimination, UVE) 5 种波段筛选方法建立 PLS 模型。分别以预测决定系数 (the prediction coefficient of determination, R^2_{pre})、预测均方根误差 (root mean square error of prediction, RMSEP)、训练决定系数 (the calibration coefficient of determination, R^2_{cal})、校正均方根误差 (root mean

square error of calibration, RMSEC) 为评价指标, 综合评价筛选出建模的最佳波段。

基于随机森林方法建立多品种蜜炙饮片水分含量定量模型 随机森林^[14]是一种基于集成学习的机器学习方法, 使用 Python 中的 sklearn 工具包中的 RandomForestRegressor 程序建立随机森林回归模型, 并对蜜炙饮片预测集水分含量进行预测。选择原始光谱、SNV、normalization、MSC、1D、2D 对样品的原始吸收光谱进行预处理后建立水分含量定量模型, 以 R^2_{pre} 、RMSEP 为模型性能评价指标, 考察不同预处理方法对模型的影响。

在建立随机森林模型前, 首先对随机森林的参数进行优化, 随机森林最重要的 3 个参数分别为: ntree、max feature、node size。选择不同的预处理方法对 3 种蜜炙饮片的原始光谱进行处理, ntree 选择范围为 1~301, 步长为 5, max feature 选择范围为 1~141, 步长为 5, node size 选择范围为 1~15, 步长为 1, 以袋外误差 (OOB-err) 作为评价指标, 考察不同预处理建模的最优模型参数。

结果与讨论

1 蜜炙饮片样品水分参考值

3 种蜜炙饮片样品水分含量烘箱法测定结果如表 1 所示。蜜炙款冬获得共 62 个样品, 水分含量范围为 6.76%~10.50%, 其中位值和均值分别为 9.15%、8.86%, RSD 为 11.23%。蜜炙黄芪获得共 70 个样品, 水分含量范围为 5.46%~13.52%, 其中位值和均值分别为 8.31%、8.43%, RSD 为 19.19%。蜜炙甘草获得共 82 个样品, 水分含量范围为 6.87%~10.65%, 其中位值和均值分别为 9.41%、9.25%, RSD 为 8.33%。除蜜炙黄芪中的 19030120 批次获得的 3 个样品外, 其余样品均符合《中华人民共和国药典》中饮片所含水分含量不得大于 13% 的规定。3 种蜜炙饮片的水分含量中位数和均值分别为 9.01%、8.87%, RSD 为 13.66%, 说明蜜炙饮片的水分含量波动较不稳定, 采用烘干法测定水分, 耗时较长, 而传统近红外水分含量定量预测模型通常只对单一品种建立专用模型, 因此, 亟需建立一种能现场、通用、快速测定多品种的蜜炙饮片水分含量的方法。

Table 1 Moisture content reference value of honey-processed Chinese herbal slice samples

Sample	Batch	Moisture content	Batch	Moisture content	Batch	Moisture content		
Honey-processed <i>Tussilago farfara</i>	200201	1	8.28	200901	1	7.85		
		2	8.11		2	8.15		
		3	8.48		3	8.33		
	200216	1	7.80	201001	1	9.85		
		2	8.11		2	9.86		
		3	8.08		3	8.00		
						2012140	1	7.01
							2	7.34
							3	7.45
					19022304	1	7.63	
						2	7.80	
						3	8.00	

Continued										
Sample	Batch	Moisture content	Batch	Moisture content	Batch	Moisture content	Batch	Moisture content		
Honey-processed <i>Astragalus</i>	200401	1	10.20	210101	1	10.22	19030229	1	9.67	
		2	10.13		2	9.69		2	10.05	
		3	9.67		3	9.85		3	9.79	
	200501	1	9.96	210105	1	9.14	19030330	1	9.87	
		2	9.37		2	9.71		2	9.79	
		3	10.5		3	9.56		3	9.68	
	200601	1	7.81	210301	1	9.54	20051402	1	10.12	
		2	7.96		2	9.8		2	9.75	
		3	8.02		3	9.31		3	10.07	
	200701	1	9.15	2001025	1	8.35	20200030	1	8.48	
		2	9.21		2	8.35		2	8.06	
		3	9.47		3	8.16		3	8.32	
	200801	1	8.70	2012139	1	6.76	20201214	1	9.45	
		2	8.25		2	7.00		2	9.21	
		3	8.28		3	7.01		3	9.49	
	20072002	1	8.07	200902	1	6.57	210101	1	7.57	
		2	8.68		2	7.14		2	7.62	
		3	9.23		3	6.54		3	7.41	
	20080132	1	9.22	200601	1	8.36	20210221	1	8.88	
		2	8.95		2	8.80		2	8.82	
		3	9.46		3	7.26		3	8.58	
	20210219	1	8.32	19030120	1	13.52	200401	1	8.84	
		2	8.07		2	13.50		2	8.79	
					3	13.19		3	8.87	
	1911024	1	8.11	201201	1	8.83	200701	1	11.88	
		2	8.02		2	8.55		2	11.71	
		3	8.15		3	8.27		3	11.47	
	200801	1	8.14	200501b	1	9.96	18051201	1	7.89	
	2103045	1	6.85		2	9.47		2	9.11	
		2	8.09					3	8.50	
		3	7.82	200901	1	9.17	200201a	1	7.87	
	210105	1	8.58		2	8.94		2	7.29	
		2	8.29		3	8.58				
		3	8.31	200201b	1	5.84	200410	1	5.46	
	201001	1	6.31		2	5.72		2	7.29	
		2	6.61		3	5.86		3	7.25	
		3	6.50	200501a	1	7.69	2102050	1	8.47	
	2103110	1	8.55		2	7.69		2	9.36	
		2	7.84		3	7.92		3	8.17	
		3	8.97							
	Honey-processed <i>Glycyrrhiza uralensis</i>	20210424	1	9.86	20100119	1	10.16	210101A	1	9.42
		2	10.26		2	9.27		2	9.40	
	3	9.42		3	9.01		3	9.66		
21010356	1	10.65	200216	1	9.90	201201B	1	9.37		
	2	10.53		2	9.69		2	9.36		
	3	10.16		3	9.82		3	9.66		
2103136	1	8.32	210105B	1	9.19	210101B	1	9.56		
	2	7.41		2	8.99		2	9.69		
				3	9.08		3	9.72		
210301	1	9.60	201001A	1	9.28	200101B	1	7.04		
	2	9.34		2	9.33		2	6.97		
	3	9.26		3	8.92		3	6.87		
210401	1	9.82	20071502	1	9.85	200601	1	9.07		
	2	9.31		2	9.47		2	8.97		
	3	8.88		3	9.62		3	9.24		
200501	1	9.80	201201C	1	9.04	200301	1	9.73		
	2	9.94		2	8.69		2	9.35		
	3	9.84		3	8.55		3	9.43		
2002044	1	8.15	21010255	1	9.02	201001B	1	9.77		
	2	8.52		2	9.22		2	9.62		

Continued

Sample	Batch	Moisture content	Batch	Moisture content	Batch	Moisture content	
		3	3	9.52	3	9.47	
	2103126	1	1911019	1	200901B	1	9.98
		2		2		2	9.38
		3		3		3	9.48
	200901A	1	200401	1	200101C	1	9.08
		2		2		2	8.63
		3		3		3	8.65
	210105A	1					10.04
		2					10.18
		3					9.83

2 蜜炙饮片 NIR 原始光谱图分析

蜜炙饮片 NIR 原始光谱图见图 1, 3 种蜜炙饮片中的水分子在 1 400~1 550 nm 均有显著吸收, 不同水分含量的蜜炙饮片样品 NIR 光谱差异非常明显, 在近红外区, 水分子有一些特征性较强的吸收带, 而其他分子的倍频与合频吸收相对较弱, 因此, NIR 光谱可用于蜜炙饮片中水分含量测定。

3 基于 PLS 的单品种蜜炙饮片水分含量检测方法建立

3.1 样本集划分结果 基于 MATLAB 软件, 采用 K-S 法, 在蜜炙款冬的 62 个样本中根据其水分含量的分布范围选择 41 个有代表性的样品组成校正集, 其余 21 个样本为验证集, 在蜜炙黄芪的 70 个样本中根据其水分含量的分布范围选择 56 个有代表性的样品组成校正集, 其余 14 个样本为验证集, 在蜜炙甘草的 82 个样本中根据其水分含量的分布范围选择 54 个有代表性的样品组成校正集, 其余 28 个样本为验证集。校正集用于建立模型, 验证集用于验证模型。

3.2 单品种蜜炙饮片光谱预处理方法考察 分别使用原光谱及预处理后的光谱数据, 结合 PLS 对蜜炙款冬、蜜炙甘草、蜜炙黄芪样品进行水分含量建模分析, 光谱预处理方法考察结果见表 2。

结果表明, 采用 SG9+2D 方法所建的蜜炙款冬水分含量 PLS 模型, R^2_{cal} 为 0.921 5, RMSEC 和 RMSECV

分别为 0.276 2 和 0.326 5, 性能偏差比 (validation set to standard error of prediction, RPD) 为 3.020 4, 所得结果较其他方法理想。因此, 采用 SG9+2D 建立蜜炙款冬 PLS 水分模型。

采用 1D 方法所建的蜜炙黄芪水分含量 PLS 模型, R^2_{cal} 为 0.897 1, RMSEC 和 RMSECV 分别为 0.502 6 和 0.576 7, RPD 为 2.718 3, 所得结果较其他方法理想。因此, 采用 1D 建立蜜炙黄芪 PLS 水分模型。

采用 2D 方法所建的蜜炙甘草水分含量 PLS 模型, R^2_{cal} 为 0.777 3, RMSEC 和 RMSECV 分别为 0.343 6 和 0.409 7, RPD 为 1.776 9, 所得结果较其他方法理想。因此, 采用 2D 建立蜜炙甘草水分含量 PLS 模型。

3.3 单品种蜜炙饮片光谱波段筛选方法考察 在最佳预处理方法的基础上, 采用 VIP、MWPLS、CARS、SR、UVE 进行波段筛选。通过 5 种波段筛选方法建立单品种蜜炙饮片水分含量 PLS 定量模型, 其性能参数如表 3 所示。蜜炙款冬水分含量定量模型采用 SR 变量筛选法后, 模型性能有所提高, R^2_{cal} 为 0.914 5, R^2_{pre} 为 0.920 4, RMSEC 为 0.258 0, RMSEP 为 0.327 3, RPD 为 5.087 1, 故采用 SR 变量筛选法, 选用波段 950~970、1 155~1 170、1 265~1 290、1 350、1 370~1 385、1 410~1 425、1 495~1 510 nm 建模。蜜炙黄芪水分含量定量模型采用 UVE 变量筛选法后, 模型性能有所提高, R^2_{cal} 为 0.887 8, R^2_{pre} 为 0.890 0, RMSEC 为 0.502 3,

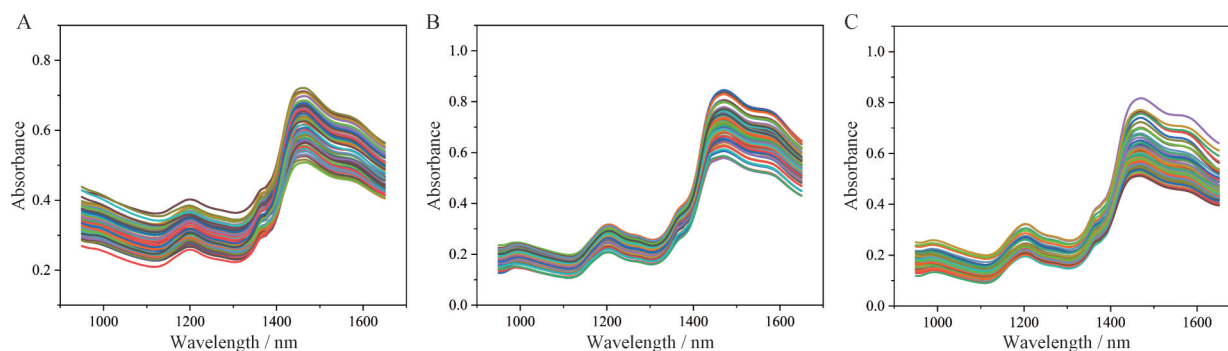


Figure 1 The near infrared (NIR) original spectrum of honey-processed Chinese herbal slices. A: Honey-processed *Tussilago farfara*; B: Honey-processed *Astragalus*; C: Honey-processed *Glycyrrhiza uralensis*

Table 2 Performance indexes of quantitative model with different spectral pretreatment methods. LVs: Level of potential variables; R^2_{cal} : The calibration coefficient of determination; RMSEC: Root mean square error of calibration; RMSECV: Root mean square error of cross-validation; RPD: Validation set to standard error of prediction; RAW: Raw spectrum; SNV: Standard normalization variate; SG: Savitzky-Golay smoothing; 1D: First derivative; 2D: Second derivative; MSC: Multiple scatter correction

Sample	Pretreatment	LVs	R^2_{cal}	RMSEC	RMSECV	RPD
Honey-processed <i>Tussilago farfara</i>	RAW	3	0.840 2	0.395 5	0.439 7	2.250 4
	SNV	2	0.846 9	0.385 8	0.409 2	2.410 2
	Normalization	3	0.864 3	0.363 3	0.397 0	2.484 4
	SG9	3	0.863 7	0.364 0	0.407 0	2.423 2
	SG9+SNV	3	0.873 3	0.351 0	0.380 7	2.590 3
	SG9+1D	4	0.907 2	0.300 4	0.341 2	2.890 5
	SG9+2D	3	0.921 5	0.276 2	0.326 5	3.020 4
	1D	3	0.884 6	0.335 0	0.356 4	2.767 3
	2D	2	0.900 1	0.311 7	0.334 8	2.945 7
Honey-processed <i>Astragalus</i>	RAW	5	0.863 1	0.579 8	0.645 8	2.426 3
	SNV	5	0.878 1	0.547 1	0.621 1	2.522 9
	Normalization	5	0.886 2	0.528 6	0.596 9	2.625 3
	SG9	5	0.864 0	0.577 8	0.644 0	2.433 4
	SG9+SNV	5	0.877 8	0.547 8	0.621 5	2.521 4
	SG9+1D	5	0.889 7	0.520 5	0.624 2	2.510 8
	SG9+2D	5	0.889 7	0.520 5	0.624 2	2.510 8
	1D	5	0.897 1	0.502 6	0.576 7	2.718 3
	2D	5	0.890 6	0.518 3	0.616 5	2.541 8
Honey-processed <i>Glycyrrhiza uralensis</i>	RAW	6	0.618 9	0.449 4	0.511 7	1.422 7
	Normalization	4	0.572 4	0.476 1	0.517 7	1.406 2
	SNV	6	0.699 4	0.399 2	0.450 9	1.614 9
	SG11	6	0.616 6	0.450 8	0.513 9	1.416 6
	SG9	6	0.617 8	0.450 1	0.512 9	1.419 5
	SG9+1D	4	0.671 0	0.417 6	0.477 3	1.525 5
	SG9+2D	4	0.729 6	0.378 6	0.431 0	1.689 2
	1D	4	0.672 1	0.416 9	0.479 9	1.517 1
	2D	4	0.777 3	0.343 6	0.409 7	1.776 9
Three kinds of honey-processed Chinese herbal slice	RAW	3	0.592 2	0.771 7	0.793 6	1.522 8
	Normalization	3	0.662 4	0.702 2	0.733 8	1.647 0
	SNV	4	0.759 3	0.592 9	0.624 4	1.935 5
	SG11	5	0.765 8	0.584 8	0.619 5	1.950 7
	1D	5	0.811 2	0.525 1	0.553 0	2.185 5
	2D	5	0.814 5	0.520 5	0.548 6	2.203 0
	MSC	4	0.748 2	0.606 4	0.635 1	1.902 8

RMSEP为0.566 7, RPD为6.822 0, 故采用UVE变量筛选法选用波段965~970、995~1 625 nm进行建模。蜜炙甘草水分含量定量模型采用CARS变量筛选法后, 模型性能有所提高, R^2_{cal} 为0.839 4, R^2_{pre} 为0.707 6, RMSEC为0.286 6, RMSEP为0.398 9, RPD为2.612 4, 故采用CARS变量筛选法, 选用波段1 065~1 070、1 090~1 100、1 170、1 270~1 285、1 425、1 465 nm进行建模。

3.4 单品种蜜炙饮片水分含量定量模型性能评价 基于Unscramble软件, 采用SNV+2D光谱预处理方法, 潜变量因子数选择3, 结合PLS方法, 建立蜜炙款冬校正集样本水分含量的近红外定量分析模型。水分预测模型的 R^2_{cal} 为0.914 5, R^2_{pre} 为0.914 5, RMSEC为0.258 0, RMSEP为0.327 3, RPD为5.087 1, 其近红外预测值与水分参考值相关性如图2A所示。将剩余21个验

证集样品的近红外图谱输入定量分析模型, 预测其水分含量, 并与所测水分值进行比较, 平均相对误差为2.73%。

采用1D光谱预处理方法, 潜变量因子数选择5, 结合PLS方法, 建立蜜炙黄芪校正集样本水分含量的近红外定量分析模型。水分预测模型的 R^2_{cal} 为0.887 8, R^2_{pre} 为0.890 0, RMSEC为0.502 3, RMSEP为0.566 7, RPD为6.822 1, 其近红外预测值与水分参考值相关性如图2B所示。将剩余14个验证集样品的近红外图谱输入定量分析模型, 预测其水分含量, 并与所测水分值进行比较, 平均相对误差为5.34%。

采用2D光谱预处理方法, 潜变量因子数选择4, 结合PLS方法, 建立蜜炙甘草校正集样本水分含量的近红外定量分析模型。水分预测模型的 R^2_{cal} 为0.839 4, R^2_{pre} 为0.707 6, RMSEC为0.286 6, RMSEP为0.398 9,

Table 3 Performance indexes of quantitative model with different spectral band filtering methods. BIAS: Rate of deviation; R^2_{pre} : The prediction coefficient of determination; RMSEP: Root mean square error of prediction; BIASp: Rate of deviation of prediction; VIP: Variable importance in the projection; MWPLS: Moving windows partial least squares; SR: Stepwise regression; CARS: Competitive adaptive reweighted sampling; UVE: Uniformative variable elimination; Cal: Calibration; Pre: Prediction

Sample	Band filtering	Cal				Pre			
		R^2_{cal}	RMSEC	RMSECV	BIAS	R^2_{pre}	RMSEP	BIASp	RPD
Honey-processed <i>Tussilago farfara</i>	RAW	0.912 9	0.260 5	0.335 8	-0.004 2	0.916 0	0.336 3	-0.020 6	4.889 3
	VIP	0.918 3	0.252 3	0.306 8	-0.006 7	0.908 6	0.350 9	-0.046 7	4.718 7
	MWPLS	0.912 9	0.260 5	0.335 8	-0.004 2	0.916 0	0.336 3	-0.020 6	4.889 3
	SR	0.914 5	0.258 0	0.316 0	-0.001 6	0.920 4	0.327 3	-0.055 2	5.087 1
	CARS	0.935 2	0.224 7	0.274 7	-0.007 8	0.844 4	0.457 7	-0.016 3	3.587 9
	UVE	0.883 6	0.301 3	0.344 5	0.001 0	0.921 9	0.305 1	-0.003 2	5.061 8
Honey-processed <i>Astragalus</i>	RAW	0.884 1	0.510 2	0.598 3	0.007 0	0.887 7	0.572 6	-0.241 2	6.768 7
	VIP	0.867 3	0.545 9	0.660 5	-0.005 9	0.834 1	0.696 1	-0.318 4	5.678 8
	MWPLS	0.884 1	0.510 2	0.598 3	0.007 0	0.887 7	0.572 6	-0.241 2	6.768 7
	SR	0.867 5	0.545 6	0.630 4	-0.001 8	0.841 4	0.680 7	-0.293 3	5.722 8
	CARS	0.905 8	0.460 0	0.519 2	-0.002 1	0.876 2	0.601 4	-0.220 8	6.284 5
	UVE	0.887 7	0.502 3	0.603 2	-0.000 8	0.890 0	0.566 8	-0.236 0	6.822 1
Honey-processed <i>Glycyrrhiza uralensis</i>	RAW	0.821 0	0.302 6	0.378 9	-0.001 0	0.650 9	0.435 8	-0.044 4	2.383 8
	VIP	0.810 0	0.311 7	0.377 9	-0.004 9	0.661 7	0.429 0	-0.028 7	2.414 4
	MWPLS	0.821 0	0.302 6	0.378 9	-0.001 0	0.650 9	0.435 8	-0.044 4	2.383 8
	SR	0.641 8	0.428 1	0.471 2	0.002 9	0.465 8	0.539 2	-0.066 9	1.931 8
	CARS	0.839 4	0.286 6	0.317 8	0.000 1	0.707 6	0.398 9	-0.050 9	2.612 4
	UVE	0.819 4	0.304 0	0.378 4	-0.000 2	0.641 7	0.441 5	-0.041 6	2.351 2
Three kinds of honey-processed Chinese herbal slice	RAW	0.780 7	0.584 1	0.614 4	-0.003 0	0.840 4	0.585 9	-0.045 8	4.361 0
	VIP	0.765 3	0.604 3	0.632 5	-0.003 0	0.833 1	0.606 2	-0.081 5	4.316 2
	UVE	0.780 9	0.584 0	0.614 2	-0.003 1	0.841 5	0.585 8	-0.049 1	4.380 7
	MWPLS	0.780 7	0.584 1	0.614 4	-0.003 0	0.840 4	0.585 9	-0.045 8	4.361 0
	CARS	0.826 8	0.519 1	0.557 1	-0.001 3	0.861 0	0.520 7	-0.007 5	4.646 8
	SR	0.805 8	0.549 7	0.572 5	-0.001 7	0.840 2	0.551 5	-0.005 4	4.333 0

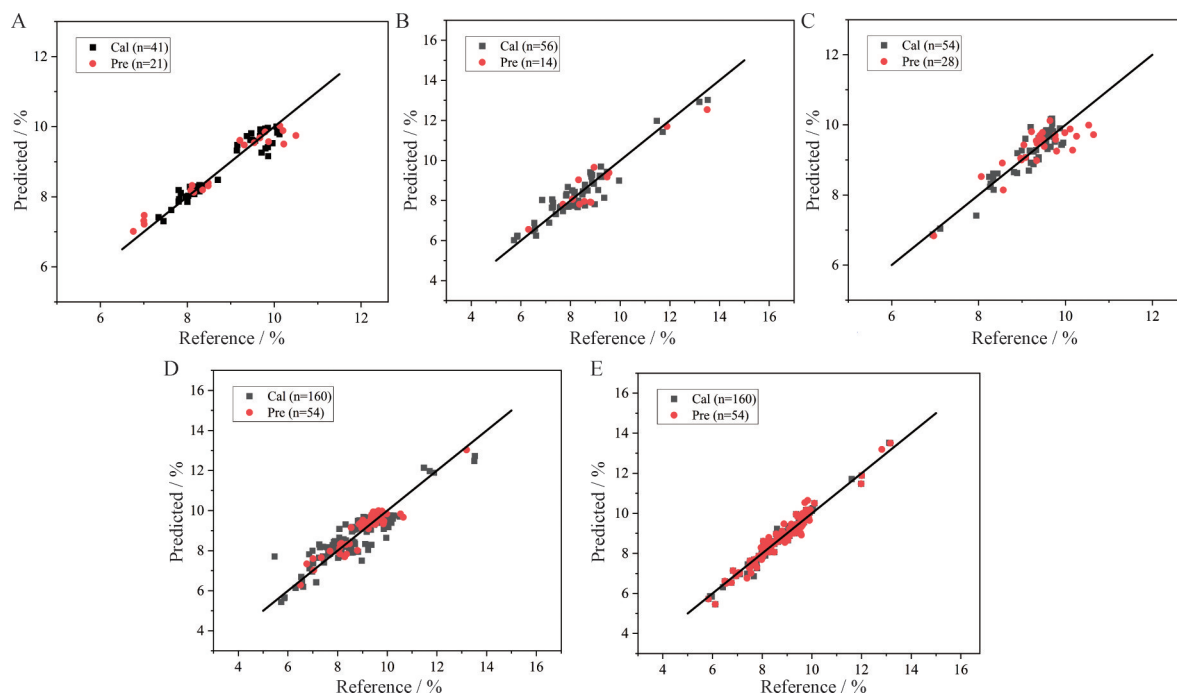


Figure 2 NIR prediction and moisture related reference graph. A: Honey-processed *Tussilago farfara*; B: Honey-processed *Astragalus*; C: Honey-processed *Glycyrrhiza uralensis*; D: Partial least squares (PLS) model of the three kind of honey-processed Chinese herbal slice; E: Random forest model of the three kind of honey-processed Chinese herbal slice

RPD为2.612 4, 其近红外预测值与水分参考值相关性如图2C所示。将剩余28个验证集样品的近红外图谱

输入定量分析模型, 预测其水分含量, 并与所测水分值进行比较, 平均相对误差为3.39%。

4 基于PLS的多品种蜜炙饮片水分含量检测方法建立

基于MATLAB软件,采用K-S法,在蜜炙饮片的214个样本中根据其水分含量的分布范围选择160个有代表性的样品组成校正集,其余54个样本为验证集。校正集用于建立模型,验证集用于验证模型。

分别使用原光谱及预处理后的光谱数据,结合PLS对蜜炙饮片样品进行水分含量建模分析,光谱预处理方法考察结果见表2。结果表明,采用2D方法所建的多品种蜜炙饮片水分含量PLS模型, R^2_{cal} 为0.814 5, RMSEC和RMSECV分别为0.520 5和0.548 6, RPD为2.203 0, 所得结果较其他方法理想。因此,采用2D建立多品种蜜炙饮片水分含量PLS模型。

在最佳预处理方法的基础上,采用VIP、MWPLS、CARS、SR、UVE变量筛选,5种波段筛选方法,建立PLS模型其性能参数如表3所示。分别以RMSEC、RMSECV、RMSEP为评价指标,综合评价筛选出建模的最佳波段。模型采用CARS变量筛选法后,模型性能有所提高, R^2_{cal} 为0.826 8, R^2_{pre} 为0.861 0, RMSEC为0.519 1, RMSEP为0.520 7, RPD为4.646 8, 故采用CARS变量筛选法,选用波段1 060~1 065、1 135、1 160~1 165、1 190~1 195、1 225、1 375、1 455、1 520 nm进行建模。

基于The Unscramble软件,采用2D光谱预处理方法,潜变量因子数选择6,结合PLS方法,建立多品种蜜炙饮片校正集样本水分含量的近红外定量分析模型。水分预测模型的 R^2_{cal} 为0.826 8, R^2_{pre} 为0.861 0, RMSEC为0.519 1, RMSEP为0.520 7, RPD为4.646 8, 其近红外预测值与水分参考值相关性如图2D所示。将剩余54个验证集样品的近红外光谱输入定量分析模型,预测其水分含量,并与所测水分值进行比较,平均相对误差为3.26%。

5 基于随机森林的多品种蜜炙饮片水分含量检测方法建立

采用与“基于PLS的多品种蜜炙饮片水分含量检测方法建立”项相同的训练集及预测集,进行建模预测,最优模型参数结果见图3A~C。由图3A可见,随着ntree的增加,不同预处理建立的模型OOB-err均呈现下降趋势,原光谱、归一化、SNV、1D、2D、MSC预处理方法的最优ntree分别为56、166、71、46、46、56。由图3B可见,随着max feature的增加,不同预处理建立的模型OOB-err均呈现下降趋势,原光谱、归一化、SNV、1D、2D、MSC预处理方法的最优max feature分别为41、51、51、141、76、81。由图3C可见,随着node size的增加,不同预处理建立的模型OOB-err无明显变

化,原光谱、归一化、SNV、1D、2D、MSC预处理方法的最优ntree分别为13、12、1、9、8、15。不同预处理建模结果见图3D。由图3D可见,采用2D方法建立的随机森林水分含量定量模型 R^2_{pre} 达到了0.903 0, RMSEP为0.330 2, 所得结果较其他预处理方法较为理想,因此使用2D方法建立随机森林模型。

采用2D光谱预处理方法,ntree选择46,max feature选择76,node size选择8,结合随机森林方法,建立多品种蜜炙饮片校正集样本水分含量的近红外定量分析模型。水分预测模型的 R^2_{cal} 为0.965 9, R^2_{pre} 为0.903 0, RMSEC为0.230 2, RMSEP为0.330 2, 其近红外预测值与水分参考值相关性如图2E所示。将剩余54个验证集样品的近红外光谱输入定量分析模型,预测其水分含量,并与所测水分值进行比较,平均相对误差为2.89%。

6 单品种与多品种蜜炙饮片水分含量定量模型性能比较

单品种与多品种蜜炙饮片水分含量定量模型预测值与水分参考值相关性见图2,由图2A~C可知单品种PLS模型均呈现不同程度的偏离情况,模型效果较不理想,由图2D可见,将3种蜜炙饮片合并建立PLS模型后,离散程度加大,证明在引入不同品种的样本后,模型性能下降,PLS模型的传递性较差。由图2E可见,相比于PLS建模,随机森林建立的蜜炙饮片水分含量定量模型建模数据散点图更集中于45°线附近,建模效果明显优于PLS方法建立的定量模型。

单品种与多品种蜜炙饮片水分含量通用定量模型性能比较结果见表4,结果表明,基于PLS建立的多品种蜜炙饮片水分含量通用定量模型与蜜炙甘草、蜜炙黄芪水分含量定量模型相比,模型性能有所提升,与蜜炙款冬水分含量定量模型相比, R^2_{cal} 由0.914 5降低到0.826 8, R^2_{pre} 由0.920 4降低到0.861 0, RMSECV由0.258 0提升到0.519 1, RMSEP由0.327 3提升到0.520 7, 说明在对多种蜜炙饮片进行水分含量预测时,PLS模型的传递效果较差。而采用随机森林方法建立的多品种蜜炙饮片水分含量通用定量模型 R^2_{cal} 达到了0.965 9, R^2_{pre} 也提升至0.903 0, RMSEP降低至0.330 2, 与采用PLS方法建立的多品种蜜炙饮片定量模型相比,模型性能大大提升,并且可以达到了蜜炙款冬独立模型的效果。由模型性能比较结果可知,基于随机森林的水分含量定量模型预测结果较为理想,达到了蜜炙款冬独立模型的效果,说明基于随机森林方法建立的蜜炙饮片水分含量定量模型可以同时用于蜜炙款冬、蜜炙黄芪、蜜炙甘草水分含量的快速、准确测定。

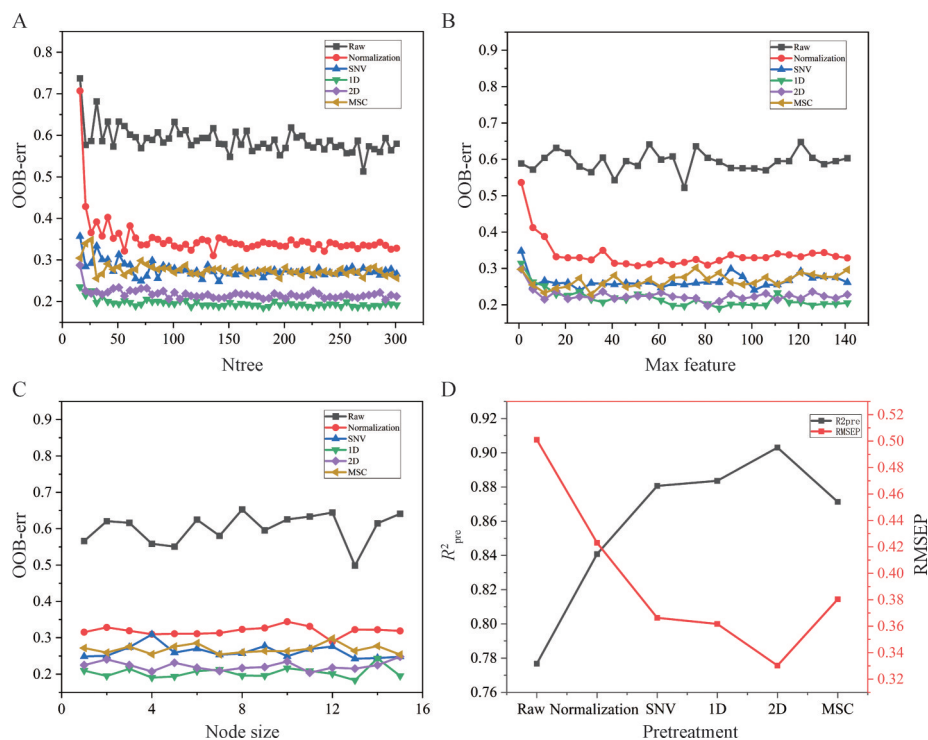


Figure 3 Influences of parameters on Random Forests model. A: Ntree; B: Max feature; C: Node size; D: Spectral pretreatment methods. OOB-err: Out-of-bag error

Table 4 Performance indexes of quantitative model with different algorithms

Model	Algorithm	Cal		Pre	
		R^2_{cal}	RMSEC	R^2_{pre}	RMSEP
Honey-processed <i>Tussilago farfara</i>	PLS	0.914 5	0.258 0	0.920 4	0.327 3
Honey-processed <i>Astragalus</i>	PLS	0.887 7	0.502 3	0.890 0	0.566 8
Honey-processed <i>Glycyrrhiza uralensis</i>	PLS	0.839 4	0.286 6	0.707 6	0.398 9
Generic model	PLS	0.826 8	0.519 1	0.861 0	0.520 7
	Random Forest	0.965 9	0.230 2	0.903 0	0.330 2

结论

本研究从解决多品种蜜炙饮片水分含量现场快速检测的问题入手,采用运动载物台模拟实际生产样品运动过程,采集近红外光谱,结合PLS与随机森林两种建模方法建立了多品种蜜炙饮片水分含量NIR定量分析模型,以RMSEC、RMSEP、 R^2_{pre} 为评价指标,比较模型性能。结果表明,随机森林建立的定量分析模型性能较优, R^2_{cal} 为0.965 9, R^2_{pre} 为0.903 0, RMSEC为0.230 2, RMSEP为0.330 2,平均相对误差为2.89%。所建模型具有良好的预测性能,与单品种蜜炙饮片水分含量独立模型相比,预测能力大于蜜炙黄芪、蜜炙甘草独立模型,可达到蜜炙款冬水分含量独立模型的效果,可实现同时对多品种蜜炙饮片水分含量进行现场检测,对中药制造过程原料单元水分含量的现场检测提供了方法,具有重要意义。

作者贡献: 张瀚、王逸飞负责研究思路设计、分析数据和文章撰写;张瀚、王文哲、胡小艳、王静负责实验实施;韩燕

雨、王晓萌、张晓梦、郭新雨、郇星月、赵静负责论文修改;李楠、吴志生负责思路指导和对文章的知识性内容进行审阅并提供研究经费。

利益冲突: 所有作者均声明没有利益冲突。

References

- [1] Wu ZS, Qiao YJ. Manufacturing Measurement of Chinese Materia Medica (中药制造测量学) [M]. Beijing: Science Press, 2021: 1-10.
- [2] Wei YN, Huang XG, Zeng JQ, et al. Critical quality attribute assessment of big brand traditional Chinese medicine: study on NIR field detection method of chemical properties of moisture in Tongren Niu Huang Qingxin Pills [J]. China J Chin Mater Med (中国中药杂志), 2021, 46: 1592-1597.
- [3] Huang XG. Research on Near-infrared Intelligent Detection of Mixing Uniformity and Moisture in Traditional Chinese Medicine Manufacturing Towards Group Standards (朝向团体标准的中药制造混合均匀度与水分智能检测研究) [D]. Beijing: Beijing

- University of Chinese Medicine, 2020.
- [4] Lu ZJ. Study on the Commercial Specification Grades and Chemical Difference Before and After Baked with Honey of Farfarae Flos (款冬花蜜炙商品规格等级及蜜炙前后的化学差异研究) [D]. Taiyuan: Shanxi University, 2020.
- [5] Golea CM, Codină GG, Oroian M. Prediction of wheat flours composition using Fourier transform infrared spectrometry (FT-IR) [J]. Food Control, 2023, 143: 109318.
- [6] de Moraes IA, Cruz-Tirado LJP, Barbin DF. Online measurement of carambola (*Averrhoa carambola* L.) physicochemical properties and estimation of maturity stages using a portable NIR spectrometer [J]. Sci Hortic, 2022, 304: 111263.
- [7] Shi X, Gan X, Wang X, et al. Rapid detection of *Ganoderma lucidum* spore powder adulterated with dyed starch by NIR spectroscopy and chemometrics [J]. LWT, 2022, 167: 113829.
- [8] Bai G, Hou YY, Ding GY, et al. A smart near-infrared spectroscopy evaluation system for quality management of Chinese medicinal materials based on quality markers [J]. Acta Pharm Sin (药学报), 2019, 54: 197-203.
- [9] Zhang P, Niu M, Tan P, et al. Construction and application of effect-related near-infrared spectroscopy for traditional Chinese medicines - exemplification by rhubarb dispensing granules [J]. Acta Pharm Sin (药学报), 2019, 54: 2162-2168.
- [10] Li XY, Wu ZS, Feng X, et al. Quality-by-Design: multivariate model for multicomponent quantification in refining process of honey [J]. Pharmacogn Mag, 2017, 13: 193.
- [11] Ying ZX, Du WF, Jiang KL, et al. Application of near-infrared spectroscopy in rapid quality control of Chinese material medica and decoction pieces [J]. China J Tradit Chin Med Pharm (中华中医药杂志), 2019, 34: 2184-2188.
- [12] Lv YX, Zuo CF, Chen WJ, et al. Rapid determination of moisture content in decoction slices of American ginseng by NIRS [J]. Prac J Med Pharm (实用医药杂志), 2021, 38: 253-256, 289.
- [13] Shi B, Xie HY, Lei JY, et al. Application of near-infrared spectroscopy in determination of moisture content in Anemarrhena Rhizome and Bark of Chinese Corktree [J]. Chin Arch Tradit Chin Med, 2017, 35: 669-672.
- [14] Breiman L. Random forests [J]. Machine Learning, 2001, 45: 5-32.