基于广义路径追踪算法建立桂枝茯苓胶囊和天舒胶囊中间体水分的近红外光谱通用定量模型

张永超[1, 2]，徐芳芳[1, 2*]，李执栋[1, 2]，李秀梅[2]，吴　云[1, 2]，刘洪波[1, 2]，王振中[1, 2]，张　欣[1, 2*]

1. 中药制药过程控制与智能制造技术全国重点实验室，江苏 连云港　222001
2. 江苏康缘药业股份有限公司，江苏 连云港　222001

摘　要：目的　以桂枝茯苓胶囊（Guizhi Fuling Capsules，GFC）和天舒胶囊（Tianshu Capsule，TC）为研究对象，将近红外光谱（near-infrared spectroscopy，NIRS）技术与机器学习算法结合，建立快速检测 2 种制剂中间体水分的方法。**方法**　采集 GFC 总混颗粒和 TC 总混颗粒的 NIRS，考察不同的预处理方法、变量筛选方法及算法对模型的影响，筛选最佳建模条件，并对 2 种中间体建立 1 个水分 NIRS 通用定量模型。**结果**　对同一中间体建立定量模型时，广义路径追踪（generalized path seeker，GPS）算法均优于偏最小二乘（partial least square，PLS）算法；GPS 通用模型与 PLS 通用模型相比，预测性能更高，验证集相对偏差（relative standard errors of prediction，RSEP）由 3.17%降至 3.03%，性能偏差比（ratio of performance to deviation，RPD）由 4.83 升至 5.05，可用于水分的预测，且与独立模型的预测性能相差不大。**结论**　GPS 算法结合 NIRS 技术建立的通用定量模型，可快速、准确地检测 2 种制剂中间体的水分。

关键词：桂枝茯苓胶囊；天舒胶囊；中间体；广义路径追踪算法；偏最小二乘算法；近红外光谱；机器学习算法；水分；通用模型；验证集相对偏差；性能偏差比

中图分类号：R283.6　　**文献标志码：**A　　**文章编号：**0253 - 2670(2023)22 - 7436 - 09

DOI: 10.7501/j.issn.0253-2670.2023.22.020

Based on generalized path seeker algorithm to establish near infrared universal quantitative model of moisture content in intermediates of Guizhi Fuling Capsules and Tianshu Capsules

ZHANG Yong-chao[1, 2], XU Fang-fang[1, 2], LI Zhi-dong[1, 2], LI Xiu-mei[2], WU Yun[1, 2], LIU Hong-bo[1, 2], WANG Zhen-zhong[1, 2], ZHANG Xin[1, 2]

1. National Key Laboratory on Technologies for Chinese Medicine Pharmaceutical Process Control and Intelligent Manufacture, Lianyungang 222001, China
2. Jangsu Kanion Pharmaceutical Co., Ltd., Lianyungang 222001, China

Abstract: Objective　Taking Guizhi Fuling Capsules (GFC, 桂枝茯苓胶囊) and Tianshu Capsules (TC, 天舒胶囊) as research objects, a rapid method for detecting the moisture content of two preparation intermediates was established by combining near-infrared spectroscopy (NIRS) technology with machine learning algorithms. **Methods**　The NIRS of GFC total mixed particles and TC total mixed particles were collected. The effects of different preprocessing methods, variable screening methods and algorithms on the model were investigated. The optimal modeling conditions were selected to establish a universal NIRS quantitative model for moisture content of two intermediates. **Results**　The generalized path seeker (GPS) algorithm was superior to the partial least squares (PLS) algorithm in establishing quantitative models for the same intermediate. Compared with the PLS universal model, the GPS universal model had higher predictive performance, with the relative standard errors of prediction (RSEP) decreasing from 3.17% to 3.03%, and the ratio of performance to deviation (RPD) increasing from 4.83 to 5.05. The GPS universal model could be used to predict the moisture content of intermediates, and there was little difference in prediction accuracy between GPS and that of

收稿日期：2023-04-06

基金项目：连云港市重大技术攻关“揭榜挂帅”项目：中药口服固体制剂智能化连续制造关键技术研究（CGJBGS2101）

作者简介：张永超，硕士，研究方向为中药制药过程新技术。E-mail: zyc020896@163.com

***通信作者：**徐芳芳，博士，研究方向为中药制药过程新技术。E-mail: 879164331@qq.com

张　欣，博士，研究方向为中药制药过程新技术。E-mail: zxtcm@126.com

the independent models. **Conclusion** The universal quantitative model established by GPS algorithm combined with NIRS technology could quickly and accurately determine the moisture content of two preparation intermediates.

Key words: Guizhi Fuling Capsules; Tianshu Capsules; intermediate; generalized path seeker; partial least square; near infrared spectrum; machine learning algorithms; moisture; universal model; relative standard errors of prediction; ratio of performance to deviation

水分是中药制剂中间体质量评价的重要指标，中间体的水分含量会影响多种关键质量属性，例如流动性、溶化性和崩解时间等，最终会影响药物的稳定性[1-3]。常规水分检测方法存在检测时间长、分析效率低、样品被破坏等缺点，同一样品不能再次检测，数据可追溯性较差。

近红外光谱（near infrared spectroscopy，NIRS）主要由 C-H、N-H、O-H 和 S-H 等基团基频振动的倍频和合频组成。将 NIRS 与化学计量法结合，能够快速检测化学成分含量及物理性质指标，目前已实现了对制剂中间体的水分[4]、粒径[5]和成分含量[6]等关键质量属性的快速检测。但是，多数研究均是对 1 种中间体进行分析，最终只能实现快速检测 1 种中间体的相关指标。NIRS 通用模型是指针对 1 个共有指标建立 1 个 NIRS 模型，可以分析 2 种或 2 种以上的样本，通用性强，稳健性更高，相比单一样本 NIRS 模型，能够节约更多成本。NIRS 通用模型在食品和农产品检测中应用较多[7]，在中药领域应用较少，仅有部分研究者针对不同的中药材建立了水分[8]、成分含量[9]等共有指标的通用模型，说明对不同样本建立通用模型具有一定的可行性。然而建立 NIRS 通用模型的算法较为单一，多数研究均基于常规的偏最小二乘（partial least square，PLS）算法建立模型，采用广义路径追踪（generalized path seeker，GPS）算法结合 NIRS 技术的应用未见报道。

GPS 算法是一种高度多样化的正则化回归，是 Jerome H. Friedman 于 2008 年发明的，主要用于处理连续或二元数据，并产生若干路径的回归或逻辑回归模型，其性能优于多数其他类型的回归模型。本研究以桂枝茯苓胶囊（Guizhi Fuling Capsules，GFC）和天舒胶囊（Tianshu Capsule，TC）为研究对象，尝试将 GPS 算法与 NIRS 技术结合，建立一个快速检测 2 种中间体水分的通用模型。

1 仪器与材料

1.1 仪器

Antaris II 型傅里叶近红外变换光谱仪，配有积分球漫反射采样系统、Result 光谱采集软件，美国 Thermo 公司；XY-105MW 型快速水分测定仪，常州市幸运电子设备有限公司；ME104E 型电子天平，梅特勒-托利多仪器（上海）有限公司。

1.2 材料

63 批桂枝茯苓胶囊总混颗粒（GFCKL），批号为 220801～220817、220901～220914、221001、221002、221201～221211、230101～230108、230201～230211；60 批天舒胶囊总混颗粒（TCKL），批号为 220801～220811、220901～220908、221001～221006、221101～221112、230101～230110、230201～230213，均由江苏康缘药业股份有限公司提供。

2 方法与结果

2.1 NIRS 采集

取 6 g 左右样品，置于配备的样品杯里，轻轻压实，采用积分球漫反射方式采集 NIRS。扫描范围为 10 000～4000 cm^{-1}，分辨率为 8 cm^{-1}，2 倍增益，扫描次数 64 次，以空气为背景，每小时扫描 1 次背景。每个样品扫描 3 次，平均值用于分析。

2.2 水分参考值测定

精密称取 2.0 g 待测样品，均匀平铺于水分测定仪样品盘上，在 105 ℃下加热 10 min，根据仪器读数即得。每个样品测量 3 次，平均值用于分析。

2.3 NIRS 预处理方法

NIRS 质量会受到各种因素影响，例如环境温湿度、仪器状态和颗粒粒度等，最终获得的 NIRS 会存在噪声信号、基线漂移等现象。为了消除无关信息，提高模型的稳健性，通常在建模前对 NIRS 进行合适的预处理。常见的预处理方法有矢量归一化法，标准正态变量变换法（standard normal variate transformation，SNV）、多元散射校正（multiplicative scatter correction，MSC）、导数法（一阶求导、二阶求导）、卷积平滑法（Savitzky-Golay，S-G）、基线校正和去趋势法等。矢量归一化法能增强光谱差异，校正由光程或样品稀释等导致的光谱变化；MSC 和 SNV 可以消除样品颗粒分布不均带来的干扰；导数法可以消除基线漂移和背景干扰；卷积平滑法能够有效去除噪声[10]。

2.4 GPS 算法[11]

GPS 算法以正态多元回归的形式建立高质量的

线性模型，它利用广谱的弹性系数建立多个候选线性模型，初始模型没有预测变量，之后在每一步中添加 1 个新变量或者更新现有变量的 1 个系数，建立若干个步数不同的路径模型，并自动筛选最优线性模型，从速度和覆盖率 2 个方面显著提升正则化回归。其主要优势之一是能够有效处理具有大量预测因子和相对较少观测值的数据矩阵，并能很好地处理高度相关的预测因子。相对传统回归，GPS 模型性能会更好、更稳定，能够应对大数据高纬度降维的挑战，但也有一些局限性，该算法不能自动发现非线性因素、预测因子之间的交互作用等。

不同于传统回归模型，GPS 算法使用弹性惩罚函数族作为数学工具来实施不同的变量选择策略，弹性惩罚函数族是由弹性的实数参数来定义。弹性可以设置为 0 和 2（包括 0 和 2）之间的任何实数，并在数学上对得到的路径解施加具有不同稀疏度的变量选择策略。无论选择的弹性的实际值是多少，任何路径最终都将达到（至少在理论上）完整预测集合中的最优解，关键的区别在于路径如何到达该点，以及在各种变量中引入或调整系数的力度有多大。同时，由稀疏策略产生的路径可能由于路径迭代、调整率以及其他因素的限制而过早终止。本研究中 GPS 模型的弹性惩罚函数族设置为 0.0、1.0、1.1、2.0，路径迭代参数设置为迭代速度为 1，学习率为 0.001，以均方误差（mean square error，MSE）为评价指标自动筛选最优系数路径模型。

2.5 数据处理与评价方法

采用 Unscrambler 11.0（Camo Analytics AS，Norway）软件进行主成分分析（principal component analysis，PCA）、NIRS 预处理及 PLS 模型建立，采用 SPM 8.3（Salford Systems，USA）软件建立 GPS 模型。本研究以样本水分为因变量，以对应的 NIRS 值为自变量，分别采用 PLS 算法与 GPS 算法建立通用定量模型。以校正集相关系数（correction set correlation coefficient，R_{cal}）、验证集相关系数（verification set correlation coefficient，R_{pre}），校正均方根误差（root mean square errors of calibration，RMSEC）、交叉验证均方根误差（root mean square errors of cross validation，RMSECV），验证均方根误差（root mean square errors of prediction，RMSEP）、验证集相对偏差（relative standard errors of prediction，RSEP）和性能偏差比（ratio of performance to deviation，RPD）为指标评价模型性能。R_{cal}、R_{pre} 越大，模型相关性越高，RMSEC 和 RMSECV 越小且较接近时，校正模型性能越高；RMSEP 较小、RPD 较大时，模型预测性能较高；当 RPD＞3 时，表示模型预测精度高[12]。本研究采用留一交叉验证法，以残余方差为评价指标确定 PLS 模型的主成分数[13]。本研究以交叉验证的 MSE 为评价指标确定最优路径的 GPS 模型。

2.6 NIRS 差异性分析

对先收集的 111 个样品的 NIRS 进行 PCA，前 2 个主成分可以解释 90%的光谱信息，主成分得分图见图 1。2 种中间体分布较集中，没有明显聚集成 2 类，提示 2 种中间体的 NIRS 相似度较高，推测对 NIRS 相似度较高的中间体建立 1 个通用定量模型具有一定的可行性。

2.7 PLS 算法建模

2.7.1 样本划分 对先收集的 111 个样品，包括 57 批 GFCKL 和 54 批 TCKL，采用随机抽样法，按照 4∶1 划分校正集与验证集，划分结果见表 1。验证集中参考值范围包含于校正集中，表明该划分较为合理。

2.7.2 光谱预处理方法的选择 中间体的原始 NIRS 见图 2。由图 2-B 可知，2 种中间体的 NIRS 较为相

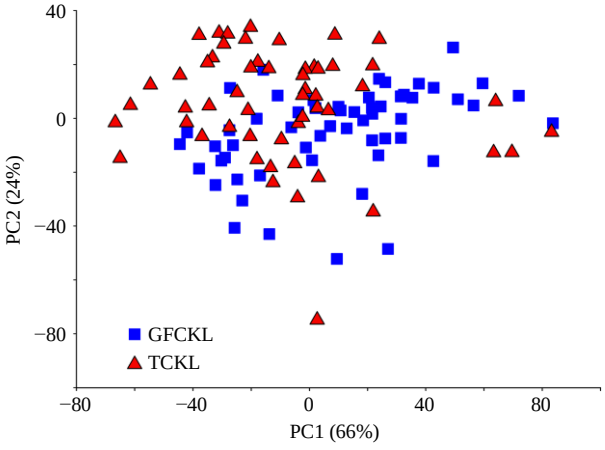


图 1 111 批样品的主成分得分图

Fig. 1 PCA scores of samples in 111 batches

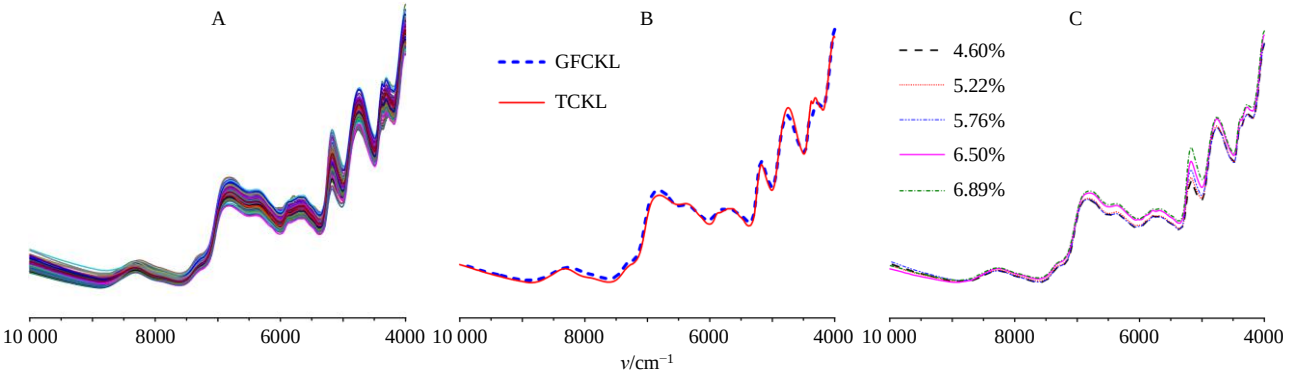
表 1 样品校正集与验证集水分参考值范围

Table 1 Moisture reference values of sample calibration and validation set

中间体	校正集			验证集		
	样本数	参考值/%	平均值/%	样本数	参考值/%	平均值/%
GFCKL	46	4.35～7.28	5.84	11	4.70～6.68	5.80
TCKL	43	3.37～7.14	5.25	11	3.91～6.58	5.18
总样品	89	3.37～7.28	5.58	22	3.91～7.14	5.41

似，在 5000 cm^{-1} 和 7000 cm^{-1} 附近均有较强的信号吸收，与 O-H 的伸缩振动相符[14-15]。图 2-C 为不同水分样本的 NIRS 变化图，随着水分含量的增加，吸光度在 5000～7000 cm^{-1} 呈现增长趋势。

本研究考察了以下预处理方法：SNV、MSC、基线校正、归一化、去趋势、一阶求导结合 SG 平滑法（S-G 1^{st}）、SNV＋S-G 1^{st}、MSC＋S-G 1^{st} 和归一化＋S-G 1^{st}，不同预处理方法对模型性能的影响见表 2。以 RPD 与 RSEP 为评价标准，综合其他指标优选最佳预处理方法。GFCKL 模型中，采用归一化结合一阶求导 SG 平滑法预处理方法最佳，RPD 为 3.83，RSEP 为 3.14%；TCKL 模型中，采用基线



A-111 个样品的原始 NIRS　B-2 种中间体的平均光谱　C-不同水分样本的光谱图

A-raw NIRS of 111 samples　B-average spectra of two intermediates　C-spectra of samples with different moisture contents

图 2　中间体的原始 NIRS 图

Fig. 2　Raw NIRS of intermediates

表 2　不同预处理方法对 PLS 模型的影响

Table 2　Effects of different pretreatment methods on PLS model

项目	预处理方法	主成分数	校正集			验证集			
			R_{cal}	RMSEC/%	RMSECV/%	R_{pre}	RMSEP/%	RSEP/%	RPD
GFCKL 模型	无预处理	3	0.954	0.254	0.280	0.963	0.205	3.51	3.42
	SNV	3	0.966	0.219	0.250	0.961	0.188	3.23	3.73
	MSC	3	0.966	0.219	0.251	0.962	0.188	3.23	3.73
	基线校正	2	0.951	0.262	0.280	0.956	0.224	3.83	3.13
	归一化	2	0.954	0.252	0.269	0.963	0.215	3.69	3.26
	去趋势	4	0.966	0.220	0.257	0.960	0.194	3.33	3.61
	S-G 1^{st}	2	0.956	0.247	0.264	0.949	0.224	3.84	3.13
	SNV＋S-G 1^{st}	4	0.984	0.148	0.207	0.959	0.191	3.27	3.67
	MSC＋S-G 1^{st}	4	0.985	0.148	0.208	0.959	0.191	3.27	3.67
	归一化＋S-G 1^{st}	2	0.963	0.229	0.249	0.975	0.183	3.14	3.83
TCKL 模型	无预处理	5	0.967	0.212	0.246	0.971	0.174	3.33	4.29
	SNV	3	0.960	0.232	0.264	0.954	0.214	4.10	3.49
	MSC	3	0.960	0.232	0.264	0.954	0.214	4.10	3.49
	基线校正	5	0.966	0.216	0.253	0.971	0.174	3.33	4.30
	归一化	5	0.972	0.195	0.229	0.959	0.205	3.92	3.65
	去趋势	3	0.964	0.221	0.252	0.965	0.193	3.69	3.87
	S-G 1^{st}	5	0.978	0.170	0.223	0.971	0.182	3.49	4.10
	SNV＋S-G 1^{st}	4	0.976	0.180	0.229	0.962	0.199	3.81	3.76
	MSC＋S-G 1^{st}	4	0.976	0.180	0.229	0.962	0.199	3.80	3.77
	归一化＋S-G 1^{st}	4	0.974	0.189	0.235	0.961	0.201	3.85	3.72

续表 2

项目	预处理方法	主成分数	校正集			验证集			
			R_{cal}	RMSEC/%	RMSECV/%	R_{pre}	RMSEP/%	RSEP/%	RPD
通用模型	无预处理	6	0.960	0.242	0.263	0.979	0.194	3.54	4.32
	SNV	4	0.960	0.242	0.264	0.982	0.173	3.17	4.83
	MSC	5	0.967	0.220	0.249	0.978	0.187	3.42	4.48
	基线校正	7	0.968	0.216	0.255	0.975	0.196	3.58	4.28
	归一化	6	0.966	0.223	0.252	0.979	0.185	3.38	4.53
	去趋势	5	0.966	0.221	0.247	0.980	0.187	3.41	4.48
	S-G 1^{st}	5	0.969	0.213	0.242	0.981	0.200	3.66	4.18
	SNV＋S-G 1^{st}	5	0.974	0.195	0.228	0.983	0.183	3.34	4.58
	MSC＋S-G 1^{st}	5	0.974	0.194	0.228	0.983	0.183	3.34	4.58
	归一化＋S-G 1^{st}	4	0.968	0.217	0.248	0.981	0.186	3.40	4.50

校正预处理后建模性能最佳，RPD 为 4.30，RSEP 为 3.33%；通用模型中，采用 SNV 预处理后模型性能最佳，RPD 为 4.83，RSEP 为 3.17%。

2.7.3 特征变量筛选 筛选特征变量可以剔除无关信息，提高模型性能。本研究在上述筛选出的最佳预处理方法基础上进一步筛选特征变量。主要考察了以下变量筛选方法：间隔偏最小二乘法（interval PLS，iPLS），组合间隔偏最小二乘法（synergy interval PLS，siPLS）和移动窗口偏最小二乘法（moving window PLS，mwPLS）。

iPLS[16]是将全光谱划分成若干个子区间，然后在每个子区间进行建模。本研究是将光谱划分成 20 个区间，以 RMSECV 为评价指标，选出最佳光谱区间。siPLS[17]是将全光谱划分成若干个子区间后，再将子区间任意组合进行建模。本研究是将全光谱划分成 20 个区间，以组合数为 4，以 RMSECV 为评价指标，选出最佳光谱区间。mwPLS[18]是从整个光谱的第 1 个波长点开始移动，沿波长变化方向截取选定窗口宽度的区间，建立一系列的 PLS 模型。本研究是以全波长的 10%（155 个波数）为窗口，以 RMSECV 为评价指标，选出最佳光谱区间。

本研究采用上述方法筛选变量后建模，结果见表 3。GFCKL 模型中，采用 mwPLS 法筛选变量后模型的性能提升最多，最佳建模区间为 4 157.77～5 230.00 cm^{-1}，在 5170 cm^{-1} 处的强吸收峰是水分子伸缩振动和弯曲震动的组合频谱带，包含上述区间

表 3 不同变量筛选方法对 PLS 模型的影响

Table 3 Effects of different variable screening methods on PLS model

项目	方法	光谱区间/cm^{-1}	主成分数	校正集			验证集			
				R_{cal}	RMSEC/%	RMSECV/%	R_{pre}	RMSEP/%	RSEP/%	RPD
GFCKL 模型	全光谱	3 999.64～10 001.03	2	0.963	0.229	0.249	0.975	0.183	3.14	3.83
	iPLS	4 601.32～4 898.31	5	0.983	0.155	0.196	0.947	0.215	3.69	3.26
	siPLS	4 300.48～4 597.46	4	0.983	0.155	0.205	0.963	0.182	3.12	3.86
		5 203.00～5 800.83								
		9 113.93～9 407.06								
	mwPLS	4 157.77～5 230.00	4	0.982	0.158	0.199	0.975	0.150	2.57	4.68
TCKL 模型	全光谱	3 999.64～10 001.03	5	0.966	0.216	0.253	0.971	0.174	3.33	4.30
	iPLS	4 902.16～5 199.15	2	0.924	0.308	0.328	0.935	0.255	4.88	2.93
	siPLS	6 707.21～7 004.19	7	0.976	0.174	0.277	0.985	0.126	2.40	5.95
		7 308.89～7 605.87								
		8 211.41～8 508.39								
		9 410.92～9 704.04								
	mwPLS	4 007.35～4 979.30	6	0.968	0.201	0.258	0.968	0.188	3.59	3.98

续表 3

项目	方法	光谱区间/cm^{-1}	主成分数	校正集			验证集			
				R_{cal}	RMSEC/%	RMSECV/%	R_{pre}	RMSEP/%	RSEP/%	RPD
通用模型	全光谱	3 999.64～10 001.03	4	0.960	0.242	0.264	0.982	0.173	3.17	4.83
	iPLS	4 902.16～5 199.15	3	0.954	0.259	0.279	0.970	0.232	4.23	3.61
	siPLS	4 300.48～4 597.46	5	0.962	0.234	0.256	0.976	0.204	3.72	4.11
		8 211.41～9 110.08								
	mwPLS	4 126.92～4 805.74	5	0.959	0.243	0.263	0.970	0.214	3.90	3.92

内；TCKL 模型和通用模型，经不同方法筛选变量后，RSEP 均变大，RPD 均变小，模型预测性能均降低，最佳建模区间均为 3 999.64～10 001.03 cm^{-1}。

2.8 GPS 算法建模

2.8.1 样本划分 方法和结果同“2.7.1”项。

2.8.2 光谱预处理方法的选择 考察不同的预处理方法对模型的影响，所用方法同“2.7.2”项，结果见表 4。GFCKL 模型中，MSC 结合一阶求导 SG 平滑法预处理方法最佳，RPD 为 6.69，RSEP 为 1.80%；TCKL 模型中，采用基线校正预处理光谱最佳，RPD 为 4.84，RSEP 为 2.96%；通用模型中，采用 SNV 结合一阶求导 SG 平滑法预处理光谱最佳，RPD 为 5.05，RSEP 为 3.03%。

2.8.3 特征变量筛选 基于上述筛选的最佳预处理方法，进一步筛选特征变量。按照变量重要性排序，通过软件自动剔除最不重要的变量，重新建模。以交叉验证的 MSE 和决定系数（coefficient of determination，R^2）为评价指标优选最佳模型，筛选变量过程见图 3。GFCKL 模型中，随着变量个数减少，R^2 呈现增大趋势，MSE 呈现减小趋势，当变量个数减少至 11 时，R^2 达到最大值 0.983，MSE 最小为 0.012，认为此时的模型最优；TCKL 模型中，当

表 4 不同预处理方法对 GPS 模型的影响

Table 4 Effects of different pretreatment methods on GPS model

项目	预处理方法	校正集			验证集			
		R_{cal}	RMSEC/%	RMSECV/%	R_{pre}	RMSEP/%	RSEP/%	RPD
GFCKL 模型	无预处理	0.969	0.208	0.248	0.961	0.203	3.48	3.45
	SNV	0.985	0.143	0.199	0.978	0.139	2.39	5.04
	MSC	0.987	0.137	0.194	0.978	0.142	2.43	4.94
	基线校正	0.974	0.191	0.248	0.972	0.178	3.04	3.95
	归一化	0.986	0.143	0.197	0.974	0.155	2.66	4.52
	去趋势	0.994	0.096	0.163	0.985	0.115	1.96	6.13
	S-G 1st	0.990	0.119	0.184	0.987	0.119	2.04	5.89
	SNV＋S-G 1st	0.995	0.084	0.152	0.985	0.116	1.99	6.03
	MSC＋S-G 1st	0.990	0.119	0.158	0.988	0.105	1.80	6.69
	归一化＋S-G 1st	0.995	0.083	0.118	0.987	0.110	1.88	6.40
TCKL 模型	无预处理	0.974	0.183	0.249	0.971	0.172	3.30	4.34
	SNV	0.981	0.155	0.212	0.976	0.161	3.08	4.65
	MSC	0.981	0.158	0.211	0.976	0.162	3.10	4.62
	基线校正	0.975	0.180	0.233	0.977	0.155	2.96	4.84
	归一化	0.964	0.214	0.272	0.958	0.225	4.30	3.33
	去趋势	0.979	0.164	0.215	0.976	0.163	3.11	4.60
	S-G 1st	0.982	0.151	0.225	0.969	0.179	3.42	4.19
	SNV＋S-G 1st	0.976	0.175	0.221	0.968	0.184	3.53	4.06
	MSC＋S-G 1st	0.981	0.156	0.223	0.976	0.171	3.27	4.38
	归一化＋S-G 1st	0.979	0.164	0.253	0.968	0.188	3.59	3.99

续表 4

项目	预处理方法	校正集			验证集			
		R_{cal}	RMSEC/%	RMSECV/%	R_{pre}	RMSEP/%	RSEP/%	RPD
通用模型	无预处理	0.968	0.215	0.249	0.981	0.179	3.27	4.68
	SNV	0.975	0.192	0.227	0.979	0.182	3.33	4.59
	MSC	0.972	0.202	0.232	0.978	0.185	3.38	4.53
	基线校正	0.965	0.225	0.257	0.983	0.169	3.09	4.96
	归一化	0.972	0.203	0.246	0.979	0.183	3.35	4.57
	去趋势	0.976	0.188	0.230	0.980	0.180	3.28	4.66
	S-G 1st	0.983	0.158	0.245	0.989	0.136	2.49	6.13
	SNV＋S-G 1st	0.980	0.169	0.215	0.982	0.166	3.03	5.05
	MSC＋S-G 1st	0.987	0.139	0.239	0.989	0.141	2.58	5.92
	归一化＋S-G 1st	0.982	0.164	0.237	0.991	0.129	2.36	6.49

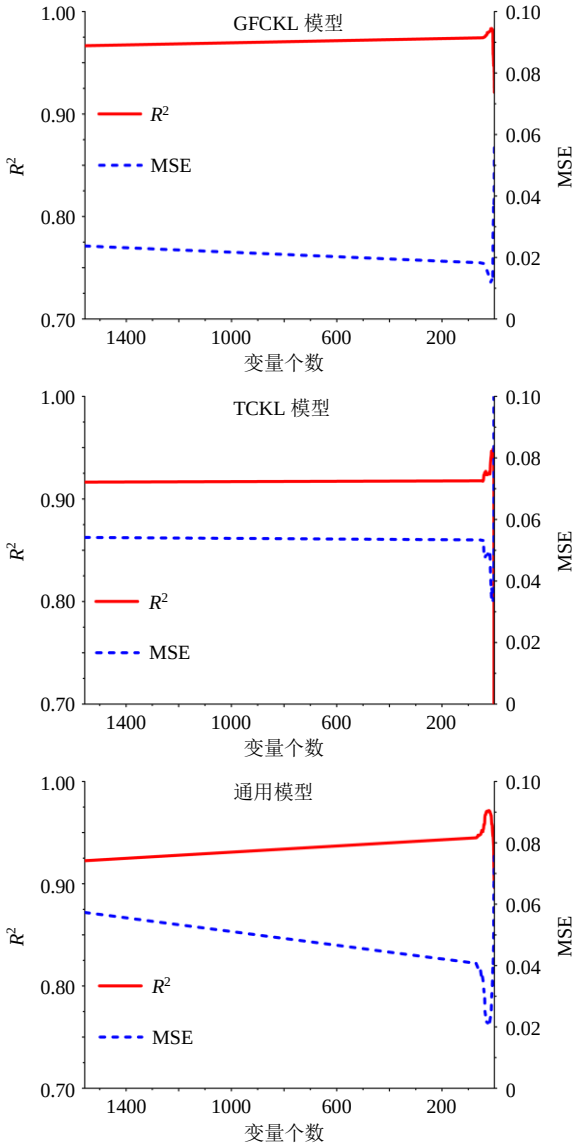


图 3　变量筛选过程中模型性能变化趋势

Fig. 3　Trend of model performance during variable screening

变量个数减少至 6 时，R^2 达到最大值 0.949，MSE 最小为 0.033；通用模型中，当变量个数减少至 22 时，R^2 达到最大值 0.971，MSE 为最小值 0.021。

由表 5 知，GFCKL 模型和通用模型经变量筛选后，RSEP 变大，RPD 变小，模型预测性能降低，最佳建模区间均为 3 999.64～10 001.03 cm^{-1}；TCKL 模型经变量筛选后，模型性能提升，因此，最佳建模波数为 5 129.72、5 546.27、8 161.27、8 450.54、8 469.83、9 403.20 cm^{-1}。在 5170、5350 cm^{-1} 处的强吸收峰是水分子伸缩振动和弯曲震动的组合频谱带，筛选出的波数 5 129.72、5 546.27 cm^{-1} 在此吸收峰附近；在 8310 cm^{-1} 附近存在较弱的吸收峰，筛选出的波数 8161.27、8 450.54、8 469.83 cm^{-1} 在此吸收峰附近。

2.9　PLS 算法模型与 GPS 算法模型比较

分别采用 2 种算法建立的模型如表 6 所示。对同一中间体建立定量模型时，GPS 算法模型均优于 PLS 算法模型，可能是因为 GPS 算法更擅长高纬度降维，面对较多维度的光谱数据更具有优势，认为 GPS 算法为最佳建模算法。

2.10　定量模型建立

采用上述筛选的最佳算法建立定量模型，结果见表 7 和图 4。各模型的 R_{cal}、R_{pre} 接近于 1，说明参考值与预测值相关性较高；RMSEC、RMSECV、RMSEP 较小，RPD 大于 3，RSEP 小于 5%，说明独立模型和通用模型的预测性能均较高，均可用于预测水分。

2.11　外部验证

将后收集的 12 批样品作为外部验证样本导入 GPS 模型中，包括 6 批 GFCKL 和 6 批 TCKL，预

表 5　不同变量筛选方法对 GPS 模型的影响

Table 5　Effects of different variable screening methods on GPS model

项目	方法	变量个数/个	校正集			验证集			
			R_{cal}	RMSEC/%	RMSECV/%	R_{pre}	RMSEP/%	RSEP/%	RPD
GFCKL 模型	全光谱	1557	0.990	0.119	0.158	0.988	0.105	1.80	6.69
	变量重要性法	11	0.995	0.085	0.109	0.986	0.116	1.99	6.05
TCKL 模型	全光谱	1557	0.975	0.180	0.233	0.977	0.155	2.96	4.84
	变量重要性法	6	0.979	0.163	0.182	0.985	0.128	2.44	5.86
通用模型	全光谱	1557	0.980	0.169	0.215	0.982	0.166	3.03	5.05
	变量重要性法	22	0.992	0.108	0.145	0.982	0.183	3.34	4.58

表 6　2 种算法模型比较

Table 6　Comparison of two algorithm models

项目	算法	校正集			验证集			
		R_{cal}	RMSEC/%	RMSECV/%	R_{pre}	RMSEP/%	RSEP/%	RPD
GFCKL 模型	PLS	0.982	0.158	0.199	0.975	0.150	2.57	4.68
	GPS	0.990	0.119	0.158	0.988	0.105	1.80	6.69
TCKL 模型	PLS	0.966	0.216	0.253	0.971	0.174	3.33	4.30
	GPS	0.979	0.163	0.182	0.985	0.128	2.44	5.86
通用模型	PLS	0.960	0.242	0.264	0.982	0.173	3.17	4.83
	GPS	0.980	0.169	0.215	0.982	0.166	3.03	5.05

表 7　最佳 GPS 模型的评价参数

Table 7　Evaluation parameters of best GPS model

项目	预处理方法方法	变量个数/个	校正集			验证集			
			R_{cal}	RMSEC/%	RMSECV/%	R_{pre}	RMSEP/%	RSEP/%	RPD
GFCKL 模型	MSC＋SG 1st	1557	0.990	0.119	0.158	0.988	0.105	1.80	6.69
TCKL 模型	基线校正	6	0.979	0.163	0.182	0.985	0.128	2.44	5.86
通用模型	SNV＋SG 1st	1557	0.980	0.169	0.215	0.982	0.166	3.03	5.05

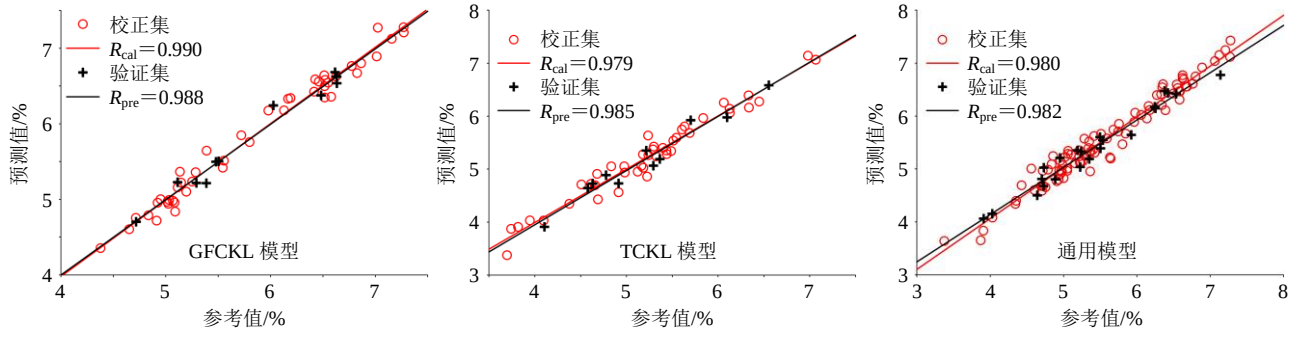


图 4　参考值与预测值的相关性

Fig. 4　Correlation between reference value and predicted value

测中间体的水分，并与参考值进行比较，结果见表 8。独立模型和通用模型的平均相对预测误差（average relative prediction error，ARPE）均小于 5%，说明 2 种模型的预测准确性较高。同时，独立模型和通用模型对同一中间体的 ARPE 差值小于 2%，说明 2 种模型预测性能相差较小，通用模型可以代替独立模型快速预测 2 种中间体水分。

3　讨论

本研究以 GFC 和 TC 为研究对象，考察了不同预处理方法和不同变量筛选方法对模型的影响，并采用 GPS 和 PLS 2 种算法建立了中间体水分的 NIRS 通用定量模型。结果表明：（1）GPS 算法比

表 8 独立模型与通用模型 ARPE 比较

Table 8 Comparison of APRE between independent model and universal model

样品	ARPE/%		
	独立模型	通用模型	差值
GFCKL	4.18	2.38	1.80
TCKL	3.51	3.80	0.29

PLS 算法表现更优，对相同中间体建立模型时，GPS 算法模型的预测性能更高；（2）采用 GPS 算法建立通用模型与独立模型时，2 种模型预测性能相差较小，均可用于预测 2 种中间体的水分。

对不同品种中间体能够成功建立通用模型，推测有以下原因：（1）本研究中 2 种制剂中间体的 NIRS 相似度较高，这可能是建立通用模型的前提条件；若再纳入更多制剂中间体再建模，能否可行还有待探索。（2）使用全光谱建模可能是关键，本研究中的 2 种算法模型，均是采用全光谱建模效果最佳。分析认为全光谱信息丰富，不会造成关键信息丢失，可能更利于通用模型的建立。（3）NIRS 对水分子的吸收较为明显，一般在 5000 cm^{-1} 和 7000 cm^{-1} 附近存在较强的吸收峰，使得光谱信息包含较多的水分信息，因此，NIRS 与水分的关联性较强，利于对不同含水量的样本建立通用模型。本研究中 GPS 算法模型均优于 PLS 算法模型，可能是因为 GPS 算法能够自动建立多个线性模型，且自动优选最佳模型，能更好应对高维度的光谱数据。由于本研究样本量较少，后续将纳入更多样本对 2 种算法再验证与比较。相比独立模型，通用模型在模型建立、维护、更新等方面等会节省较多成本。目前，在制药领域，近红外通用模型研究较少，通用的深层次机理还需要进一步探索。本研究首次尝试将 GPS 算法与 NIRS 技术结合，成功建立了快速检测 2 个不同品种中间体水分的通用模型，模型的准确性优于常用的 PLS 模型，提示在建模研究中，可以采用多种算法提高模型的预测性能，为 NIRS 技术在定量模型研究方面提供新思路。

利益冲突 所有作者均声明不存在利益冲突

参考文献

[1] Faulhammer E, Llusa M, Radeke C, *et al*. The effects of material attributes on capsule fill weight and weight variability in dosator nozzle machines [J]. *Int J Pharm*, 2014, 471(1/2): 332-338.

[2] 汪盛华, 秦春娟, 安双凤, 等. 水提干法制粒的中药配方颗粒溶化性与粉体物理属性相关性研究 [J]. 中草药, 2023, 54(5): 1439-1448.

[3] 夏春燕, 徐冰, 徐芳芳, 等. 天舒片素片崩解时间实时放行检验研究 [J]. 中国中药杂志, 2020, 45(2): 250-258.

[4] 李民, 张春辉, 刘春兰, 等. 近红外光谱法测定骨龙胶囊中间体粉末中水分 [J]. 现代药物与临床, 2019, 34(8): 2280-2282.

[5] 张永超, 徐芳芳, 张欣, 等. 腰痹通胶囊 4 种中间体粒径的近红外光谱通用定量模型研究 [J]. 中草药, 2021, 52(1): 55-64.

[6] 宋侨, 胡俊杰, 白玉, 等. 马应龙麝香痔疮膏中间体中煅炉甘石与冰片近红外含量模型建立 [J]. 药学研究, 2020, 39(1): 16-21.

[7] 李明, 韩东海, 鲁丁强, 等. 近红外光谱通用模型在农产品及食品检测中的研究进展 [J]. 光谱学与光谱分析, 2022, 42(11): 3355-3360.

[8] 马卉, 冯雪静, 陈明, 等. 近红外光谱结合化学计量学快速测定蓝芩口服液原药材水分含量 [J]. 中国现代应用药学, 2021, 38(23): 2932-2939.

[9] 张丝雨. 基于近红外光谱技术的一清胶囊原药材质量控制研究 [D]. 杭州: 浙江大学, 2020.

[10] 褚小立, 袁洪福, 陆婉珍. 近红外分析中光谱预处理及波长选择方法进展与应用 [J]. 化学进展, 2004, 16(4): 528-542.

[11] Friedman J H. Fast sparse regression and classification [J]. *Int J Forecast*, 2012, 28(3): 722-738.

[12] 张娜, 徐冰, 贾帅芸, 等. 丹参提取过程多源信息融合建模方法研究 [J]. 中草药, 2018, 49(6): 1304-1310.

[13] 刘燕德, 黎丽莎, 李斌, 等. 多品种苹果可溶性固形物近红外无损检测通用模型研究 [J]. 华中农业大学学报: 自然科学版, 2022, 41(2): 237-244.

[14] Rantanen J, Antikainen O, Mannermaa J P, *et al*. Use of the near-infrared reflectance method for measurement of moisture content during granulation [J]. *Pharm Dev Technol*, 2000, 5(2): 209-217.

[15] Ma L J, Peng Y F, Pei Y L, *et al*. Systematic discovery about NIR spectral assignment from chemical structural property to natural chemical compounds [J]. *Sci Rep*, 2019, 9(1): 9503.

[16] 吴静珠, 石瑞杰, 陈岩, 等. 食用油油酸的近红外特征谱区优选 [J]. 中国粮油学报, 2015, 30(2): 118-121.

[17] 徐芳芳, 杜慧, 张欣, 等. 在线中红外光谱监测热毒宁注射液金银花与青蒿醇沉过程 7 种指标成分研究 [J]. 中草药, 2021, 52(10): 2909-2917.

[18] 刘秋安, 徐芳芳, 张欣, 等. 基于近红外光谱技术和分类与回归树算法建立天舒片崩解时间预测模型 [J]. 中草药, 2021, 52(16): 4837-4843.

[责任编辑 郑礼胜]