近红外光谱技术应用于中草药口服液在线质量控制的化学计量学建模

李晶晶[1]，周昭露[1]，黄生权[2]，鲁亮[2]，黄延盛[2]，田淑华[3]，FALOLA Akinola A[3]，李璇[3]，李杰[1]，舒逸聪[1]，王学重[1]

（[1]华南理工大学化学与化工学院，广东 广州 510640；[2]无限极（中国）有限公司，广东 广州 510623；[3]晶格码（青岛）智能科技有限公司，山东 青岛 266109 ）

摘要：以一种中草药口服液中多糖含量、可溶性固形物含量及 pH 为对象，利用近红外光谱技术对其进行实时在线检测，重点研究了化学计量学建模方法。建模过程采用了一种系统工程的方法，综合考虑了光谱预处理、特征波段选择、数据选择和分组、近红外原始光谱和化学分析参考值的误差分析，并把工具箱集成在一起。还考察了光谱采集的重复性、采样点取样条件，并对实验室测定的参考值进行了重复性及测量误差的考察。通过对比最终选择的是偏最小二乘法（PLS）结合遗传算法自动选波段。除此之外，以模型的交叉验证均方根误差（RMSECV）、预测均方根误差（RMSEP）、质量指标化学参考值与预测值间的相对误差为指标筛选最佳建模参数。建模结果显示，多糖模型预测值与化学参考值间的相对误差可控制在10%内，可溶性固形物误差在 5%内，pH 误差在 4%内。生产使用表明，所建立的模型可有效应用于在线、快速对质量指标做出准确的预测。

关键词：近红外光谱；偏最小二乘；中草药；质量控制；变量选择

中图分类号：TQ 460.6 $^{+3}$ 文献标志码：A 文章编号：1000-6613（2018）05-1923-10

DOI：10.16085/j.issn.1000-6613.2017-1413

Development of chemometric models based on near infrared spectroscopy for on-line quality control of a Chinese herbal liquid tonic

Li Jingjing1, Zhou Zhaolu1, Huang Shengquan2, Lu Liang2, Huang Yansheng2, Tian Shuhua3, FALOLA Akinola A3, Li Xuan3, Li Jie1, Shu Yicong1, Wang Xuezhong1

([1]School of Chemistry and Chemical Engineering, South China University of Technology, Guangzhou 510640, Guangdong, China;[2]Infinitus(China), Guangzhou 510623, Guangdong, China;[3]Pharmavision(Qingdao)Intelligent Technology Ltd., Qingdao 266109, Shandong, China)

Abstract: Chemometric model development methods were investigated for the application of the near infrared spectroscopy (NIR) to on-line real-time monitoring of the content of polysaccharides, the content of soluble solids as well as pH of a liquid tonic. A system engineering approach was employed in building chemometric models via integration of spectra preprocessing, feature wave number selection, grouping and data selection for training and validation, as well as examination of the raw NIR measurements and laboratory analytical reference values. The final selected methods for data pre-processing, feature wave number selection and model building were the first order derivative, genetic algorithm, and PLS. The model performance was evaluated by RMSEC, RMSEV, and RMSEP through cross validation and assessed using test data as well as real data from the industrial process. It was shown that the relative error between the predicted value and the reference value could be controlled within 10% for the content of polysaccharide. The relative errors for predicting soluble

solid and pH were controlled within 5% and 4%, respectively. The models were successfully used in commercial applications for timely estimation of the product quality related properties.

Key words :near infrared spectroscopy ;partial least squares ;Chinese herbal medicine ;quality control; variable selection

以中药材为基础的中药和营养保健品是中华民族的瑰宝，在国际市场上也越来越得到认可。但是根据2007年的一组统计数据在国际中成药市场上日本占80%份额，韩国占10%，而我国仅占5%。国外的企业甚至从我国低价购入原料药，制成制剂后以高价返销中国市场[1]。我国中药生产企业急需提高产品制造的水平，才能在国际市场上取得更大的份额。利用过程分析技术对产品质量进行在线检测和调控是提高中药制造水平的有效手段之一。本文所研究的某中药口服液是从多种天然植物中提取免疫调节剂——复合多糖，再配以多种补益类及清热解毒、药食两用的天然植物等精制而成。能有效调节、增强人体免疫力，提高机体抵抗力、防止疾病的侵袭，增强并改善体质。该中药口服液的主要功效成分为多糖，因此，控制多糖含量是保障口服液质量的关键。目前测定多糖含量的方法主要有3种，即色谱法、化学法以及近红外(NIR)光谱法[2]。因色谱法测量多糖含量需要有对应的标准物质，且分析时间长，所以色谱法更多用于分析多糖的结构和组成；化学法虽然是一种离线分析手段，耗时且操作繁杂，但其测量结果较准确，因此，化学法的测量结果常常作为近红外光谱分析的参考值。近年来数据挖掘及软仪表技术对于生产过程中的在线质量监测得到越来越广泛的应用[3-4]，其中近红外光谱分析技术是一种有效简便的分析方法，因其具有快速、高效、实时在线、操作简单、无样品预处理、对样品无破坏等优势，在食品药品的定性定量分析中受到越来越广泛的应用。在食品方面,KAWANO等[5-6]运用近红外光谱技术测量了桃子和柑橘中的糖含量，而且报道了一种水果自动分类方法；何勇等[7]利用近红外结合主元分析和神经网络鉴别苹果品种；WU等[8]研究了近红外在米酒发酵过程中对质量参数的检测。王学重等[9-11]运用近红外技术对医药的结晶过程进行监控。在中药研究方面，战皓等[12]运用近红外技术对中药进行了定性分析和含量测定；吴功煌等[13]将近红外模式识别技术应用于中药材及中成药的真伪鉴别、中药材的产地和来源鉴别；刘冰等 [14]应用近红外对黄芪精口服液中的黄范多糖和黄芪甲苷进行了建模预测。

本文将从数据前处理、预处理方法的选择、选波段方法的选择几个方面进行评估，且避免了传统手动选波段方法的粗糙，使用遗传算法等自动选波段方法既避免了全波段建模引入过多冗余信息又规避了单变量波段选择对未知数据预测性能差的问题。建立中草药口服液多糖含量、可溶性固形物含量、pH模型,并结合生产工艺流程将近红外模型实时在线应用于实际生产过程中，这对于中草药口服液在线质量控制具有十分重要的意义。

1 工艺流程及建模策略

1.1 工艺流程

本研究中所用数据均来自于某企业某中药口服液的生产过程，成品的生产包括浓缩和配制两个过程，其工艺流程分别见图1、图2，本研究中所讨论的模型为配制终点的模型。图1中浓缩过程有两套并行设备，近红外探头固定于浓缩终点的缓存罐底部，当一批口服液浓缩至终点后便进入缓存罐进行搅拌缓存，待搅拌完毕进行光谱采集，同时在固定探头处放出采集光谱的样品进行质量指标的离线测检，近红外模型的应用可对浓缩终点进行判定。图2所示工艺流程中是将浓缩液冷却后入库，待使用时与溶解后的辅料进行混合配制，经过离心去除固体颗粒后经管道打入暂存罐，并进行后续的罐装等操作。此过程中近红外探头安装在离心机与缓存罐之间的管道上，与浓缩过程相同，在采集近红外光谱的同时采集样品进行质量指标的

离线检测。但值得注意的是在采集近红外光谱及取样的过程中，应控制管道中的样品流速，原因有两点：一是流速过快会导致管道震动、样品中产生较多气泡，这对光谱的采集会造成很大噪音干扰；二是流速过快时无法保证采集光谱的样品和用于质量指标离线检测的样品为同一样品，所以在采集光谱时应注意控制管道中样品的流速并将取样点设置在环境较稳定的位置。

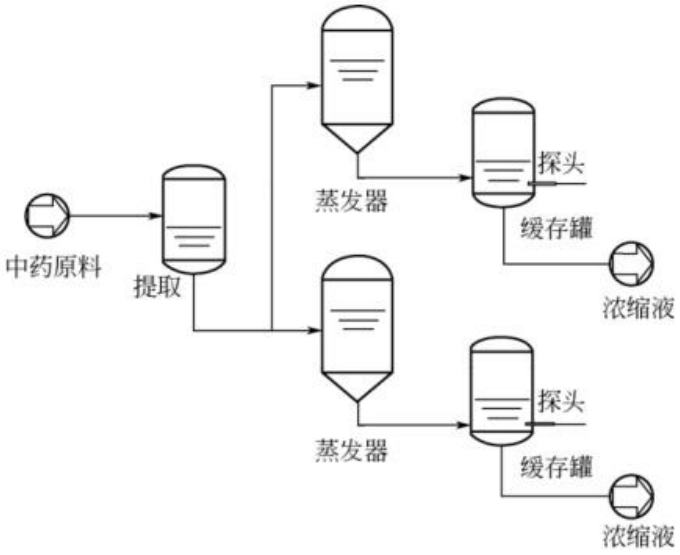


图 1 浓缩工艺流程

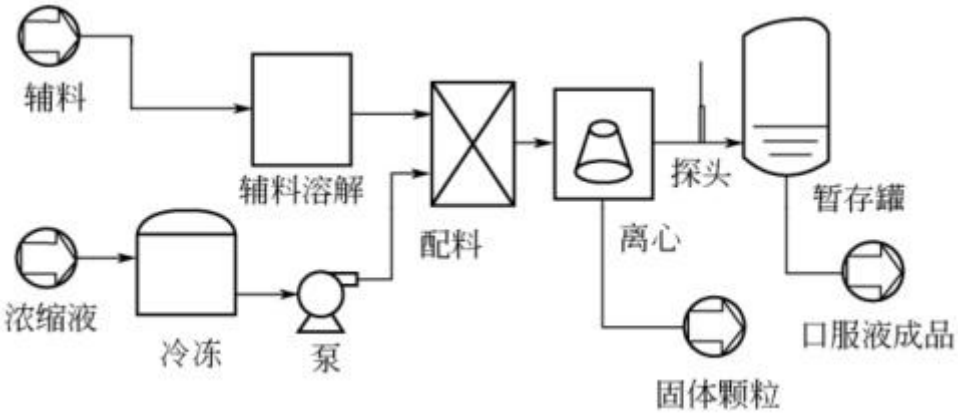


图 2 某中药口服液配制工艺流程

当模型在应用时，工人可在电脑端点击采集光谱，由探头采入的光谱便传输到电脑中作为模型的输入，30s 内即可得到对应的质量指标预测值，比起离线检测 1~2 天的时间，大大缩短了检测时间，且对样品无破坏性，无需预处理，无需耗材，经济环保，为口服液生产过程实时在线质量控制提供了技术支持。

1.2 建模流程

本研究的建模流程见图 3，整个流程包括 3 部分，即数据的前处理、建模及模型的评估。

数据前处理：对于在线采集的近红外光谱，每 个样品的光谱均采集了 3 次并取平均值，目的是减少仪器或环境所造成的误差，但若有 1 组光谱明显 区别于另外 2 组，则剔除这一组并将剩余 2 组光谱取平均作为建模光谱；若 3 组原始光谱相互间差异 均很大（肉眼可轻易区分），则将光谱剔除并重新采集光谱直至取到可信的光谱。为保证建模数据的准 确性，每组样品设置两组平行样，计算平行样间化学值的相对误差，若相对误差大于 5%，则剔除此 样品，重新取样再测量。将合格的样品近红外光谱数据及质量指标化学值数据使用 sample set partitioning based on joint x-y distance (SPXY)[15-17]方法进行校正集和验证集的划分。

常用的建模方法有人工神经网络（ANN）支持向量机（SVM）主成分回归（PCR）及独立主元回归（ICR）[18]等。ZHAO 等[19-20]在 ICR 的基础上，将全谱划分为若干波段，在各波段上独立建模

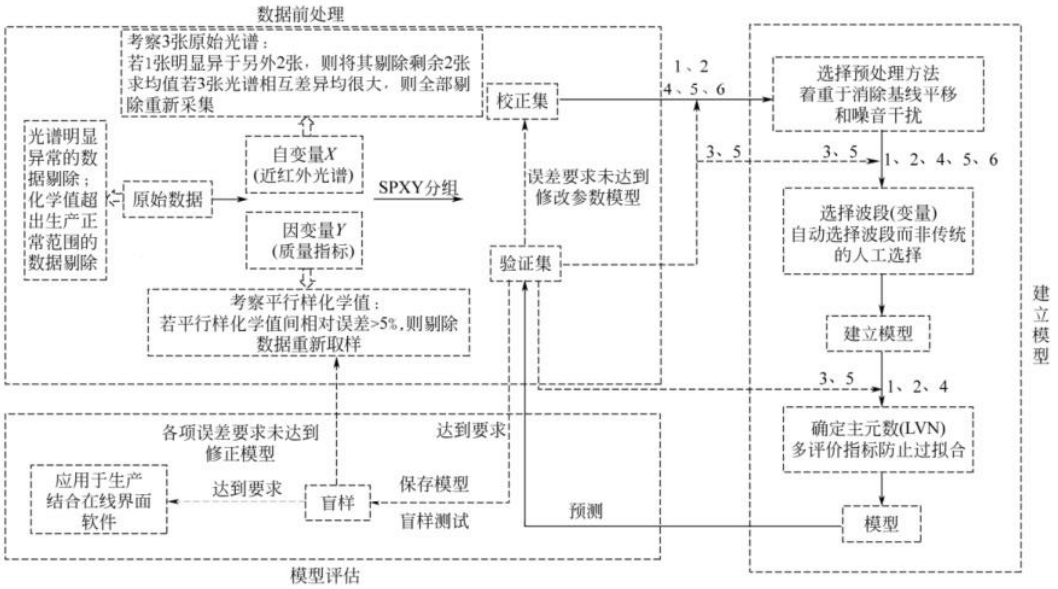


图 3 建模流程图 1—误差均方根（RMSEC）；2—交叉验证均方根（RMSECV）；3—预测均方根（RMSEP）；4—RMSECV 随 LVN 变化趋势； 5—预测值与标准值间相对误差；6—建模数据主元分析

可对混合物中各化学成分进行预测分析，又在此工作的基础上提取各波段间的共同信息和特征信息，分析两种信息对物质质量检测的累计影响。本文在 前期工作中对各种线性和非线性建模方法，包括 ANN、SVM、PCR、ICR 以及偏最小二乘（PLS）等做了充分的尝试，发现对于当前的数据，偏最小二乘结合遗传算法（GA）自动选波段提取光谱的主要信息建立的模型最好。本文利用 GA 在全谱中自动（非人工选择）提取含有主要信息的波段的方法不同于上面提到的手动方法[19-20]。

建模时使用校正集对预处理方法进行选择，常用的预处理方法有一阶求导（1 st derivative）、二阶求导（2 st derivative）、Detrend、Baseline、SNV[21]、MSC[22]及各方法之间的组合等，评价指标为误差均方根（RMSEC）、交叉验证均方根（RMSECV）、预测均方根（RMSEP）[23]，选出最佳预处理方法后进行建模波段的选择，常用选波段的方法有 GA[10-24,25]、间隔偏最小二乘法(iPLS)[26]、移动窗口偏最小二乘法(mwPLS)[27-28]、随机蛙(Random Frog)[29]、竞争自适应重加权采样（CARS)[30]、无信息变量消除法（UVE）[31]、变量投影重要性（VIP）[32-33]等，评价指标为 RMSEC、RMSECV 及 RMSEP，确定最佳选波段方法及主元数后将模型输出保存。

模型的评估：除了上文提到的 RMSEC、RMSECV 和 RMSEP 这 3 个评价指标，还要查看 RMSECV 和主元数 (LVN) 关系图，选择主元数至关重要，既不能遗漏重要信息，又不能引入过多无用信息，若主元数过多会导致过拟合。原则是当主元数达到某一值后，RMSECV、RMSEC、RMSEP 变大或保持不变，若不能保证三者均为此趋势，则要进行主元数的权衡，这需要一定的建模经验。当以上评估条件达到后，使用模型对未知数据进行预测，查看未知数据的 RMSEP 能否满足要求，同时计算化学值与预测值间的相对误差，若相对误差满足生产中的控制要求，则模型可以使用，若未达到生产要求，则对模型进行修正直至满足要求。

2 数据采集及前处理

2.1 实验部分

2.1.1 仪器、试剂及样品

主要仪器：Antaris MX 傅里叶近红外在线光谱分析仪,美国 Thermo Fisher Scientific 公司;UV2600 紫外-可见分光光度计,上海天美;ME2002

电子天平，梅特勒- 托利多；LXJ-IIB 大容量多管离 心机，上海安亭；DGG-9240BD 电热恒温干燥箱，上海森信；RX-5000a 数字折光仪，ATAGO。

药品和试剂：无水乙醇、丙酮、乙醚、蒽酮、葡 萄糖，均为分析纯，广州化学试剂厂；蒸馏水。

口服液样品：2015 年 11 月—2016 年 11 月于某 药企某中药口服液生产线浓缩终点在线采集的不同月份不同批次的口服液样品，用于多糖建模的样品 224 批，用于可溶性固形物建模的样品 235 批，用于 pH 建模的样品 185 批，后又于 2017 年 1 月在同一生产线同一浓缩暂存罐中采集口服液样品 20 批。

建模环境：Matlab (R2014a)及 Matlab 自带的 PLS Toolbox_821。

2.1.2 实验方法

多糖总糖含量的测定：糖含量测定参考《食品中还原糖的测定》GB/T 5009.7—2008 第二法硫酸-蒽酮法进行测定，每个样品测定两组平行样，建模数据为两次测量的平均值。可溶性固形物含量的测定：折光计法，每个样品测定两组平行样，建模数据为两次测量的平均值。pH 的测定：酸度计进行测定，每个样品测定两组平行样，建模数据为两次测量的平均值。

2.2 近红外光谱的采集

近红外探头固定于口服液浓缩终点的暂存罐底部，可进行实时在线扫描检测，在采集光谱的同时采集该样品进行 3 种质量指标（多糖含量、可溶性固形物含量、pH）的离线化学检测。为对 3 种模型进行应用价值的考察，于 2017 年 1 月在同一生产线同一暂存罐中对同一产品以相同方式获得 20 批口服液的光谱数据及 3 种质量指标的化学值，为方便说明，此 20 批数据称之为盲样。

光谱采集条件为：近红外透反射模式，以空气为参比，采集波长为 1000～2500nm，扫描次数为 32，分辨率为 8cm^{-1}（全谱共 1557 个变量），吸光度数据格式为 SPA，每个样品重复扫描 3 次，3 次扫描的平均值作为该样品的近红外光谱数据。图 4 为多糖模型 224 组样品近红外光谱图，图 5 为可溶性固形物模型 235 组样品近红外光谱图，图 6 为 pH 模型 185 组样品近红外光谱图，图 7 为 20 组盲样近红外光谱图。光谱图中横轴为波数 4000～10000cm^{-1}间的 1557 个波数，纵轴为 1557 个波数所对应的吸光度；由图 4～图 7 可知，变量 1～200 即波数 4000～4770cm^{-1}间的光谱噪音较大，且该段主要为水峰，在建立模型时手动将此波段剔除；另由图 4~图 7 可知光谱基线漂移较严重，在选择预处理方法时应注意基线漂移的消除。

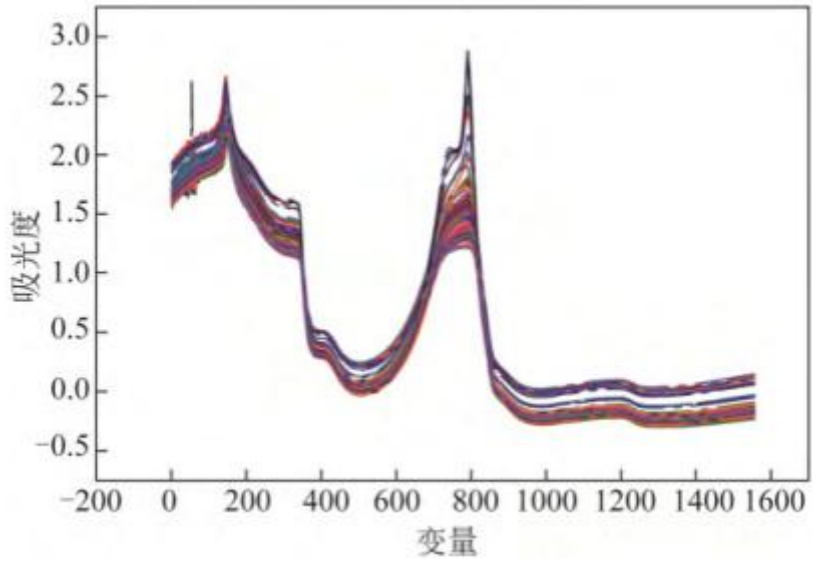


图 4 多糖模型建模样品近红外光谱

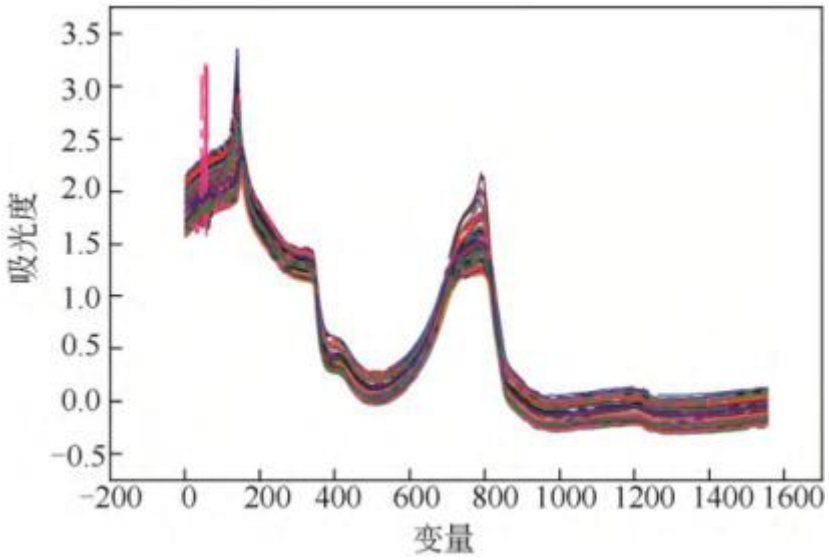


图 5 可溶性固形物模型建模样品近红外光谱

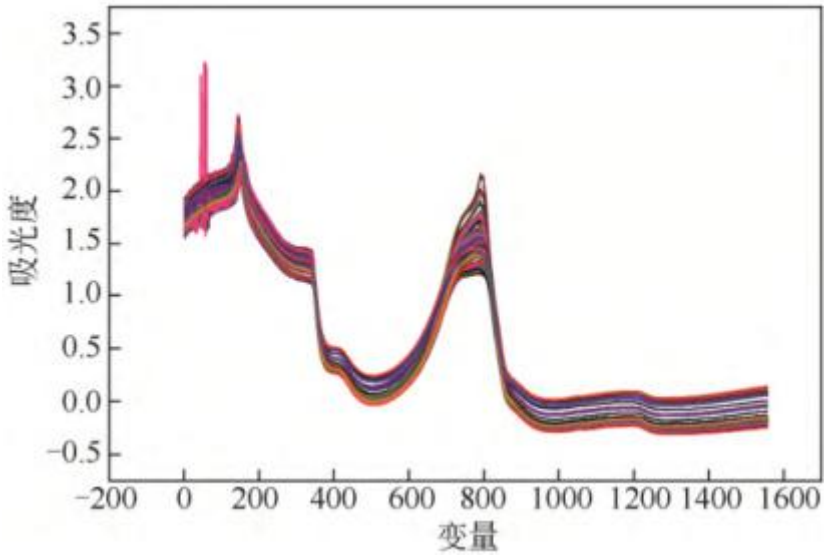


图 6 pH 模型建模样品近红外光谱

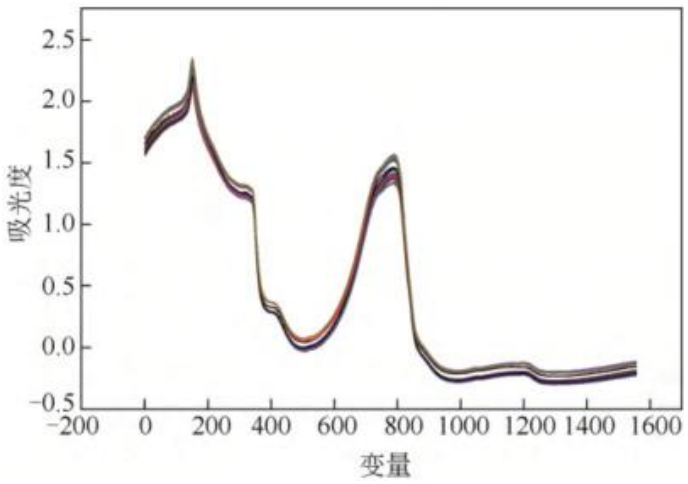


图 7 盲样近红外光谱

2.3 校正集和验证集的划分

224 组多糖含量的范围为 435.65~701.09mg/100mL,235 组可溶性固形物含量的范围为:14.8%~17.3%,185 组 pH 的范围为 4.26~4.65。为保证验证集的数据范围在校正集的数据范围内，将多糖的最大值和最小值样本取出作为校正集，剩余样本的多糖数据及相应光谱使用 SPXY 方法进行数据划分，分组后多糖模型的校正集为 180 组，验证集为 44 组。使用相同方法对可溶性固形物和 pH 数据进行分组，可溶性固形物模型的校正集为 200 组，验证集为 35 组；pH 模型的校正集为 160 组，验证集为 25 组。

多糖模型的校正集、验证集及盲样的多糖含量 范围、平均值及标准偏差见表 1。

表 1 多糖建模数据化学值统计

模型	校正集	验证集	盲样
多糖			
样品数	180	44	20
浓度	435.65~701.09	448.50~600.32	482.66~571.30
$mg/100ml^{-1}$			
平均值	534.15	517.67	518.01
标准偏差	51.92	34.78	26.22
可溶性固形物			
样品数	200	35	20
质量分数范围/%	14.80~17.30	14.82~17.00	14.92~16.47
平均值	15.615	15.769	15.524
标准偏差	0.424	0.575	0.379
pH			
样品数	160	25	20
范围	4.31~4.65	4.26~4.63	4.35~4.65
平均值	4.500	4.492	4.540
标准偏差	0.0689	0.0978	0.0819

表 1 数据表明，多糖含量范围较广，且口服液 样品多糖含量较低，这对口服液中的多糖建模提出 了挑战。另外，验证集及盲样的质量指标范围均应 在校正集的数据范围内，这对模型的应用范围产生 了一定的限制。

3 建立模型

3.1 光谱预处理方法的选择

在采集光谱过程中,工厂环境噪音、机器噪音、样品不均匀导致的光散射等因素导致光谱出现高频随机噪音、信号本底过强、基线漂移严重等现象,这将严重影响模型的准确性，所以需对光谱进行预处理。光谱预处理方法主要有一阶求导（1^{st} derivative）、二阶求导（ 2^{st} derivative）、Detrend、Baseline、SNV、MSC 及各方法之间的组合等，表 2 为采用不同预处理方法建模的部分结果，综合考虑 RMSEC、RMSECV、及 R^2CV ，最佳预处理方法为 Detrend（linear）。图 8 为多糖模型近红外光谱经预处理后的光谱图，由图 8 可知，经过预处理的近红外光谱基线平移和漂移问题明显改善，这对建模有利。

表 2 不同预处理方法对多糖模型的影响

预处理方法	LVN	RMSEC	RMSECV	R 2 C	R 2 CV
去趋势 1	13	24.922	36.351	0.76824	0.53033
去趋势 2	6	36.643	39.174	0.49899	0.43003
多元散射校正 1	9	33.978	38.322	0.56918	0.45779
多元散射校正 2	1	51.777	52.162	0.00011	0.01478
去基线 1	8	31.454	40.747	0.63083	0.40409
去基线 2	2	39.084	41.745	0.43002	0.35286
一阶求导，w=3	3	36.838	41.045	0.49368	0.37771
一阶求导，w=5	3	37.707	40.894	0.46947	0.37995
一阶求导，w=7	3	38.241	40.977	0.45433	0.37665
一阶求导，w=9	3	38.615	41.072	0.44362	0.37329
一阶求导，w=11	3	38.870	41.151	0.43622	0.37059
一阶求导，w=13	3	39.050	41.209	0.43099	0.36860
一阶求导，w=15	3	39.179	41.248	0.42725	0.36729
标准正交变量	9	34.202	38.349	0.56314	0.45669
无预处理	10	32.903	36.797	0.59603	0.49964

该中药口服液的有效成分是多糖，其结构复 杂，近红外光谱的重叠峰严重，若采用全谱建模必然会引入不必要的信息从而影响模型准确度，传统 的手动选波段方法较粗糙，采用自动选择波段方法筛选有效变量剔除冗余信息能够在一定程度上提高 模型的准确度和精密度。

3.2 有效建模波段的选择

本研究使用的变量选择方法有遗传算法(GA)、间隔偏最小二乘法(iPLS)、移动窗口偏最小二乘法(mwPLS)、随机蛙（Random Frog）竞争自适 应重加权采样（CARS）、无信息变量消除法(UVE)、变量投影重要性(VIP)，其中 Random Frog、CARS、UVE 及 VIP 算法为单变量选择，故又考察了 GA、iPLS、mwPLS 算法中每一段波段的宽度（称之为窗口宽度）对建模结果的影响，考察的窗口宽度为 1~50 及 100、150、200 和 250。表 3 为不同波段选择方法建模的部分结果，图 9 为各算法 RMSECV 随主元数（LVN）的变化趋势，其中箭头所指即为该算法的最佳主元数。综合考虑 RMSEC、RMSECV、R^2C、R^2CV 及 RMSECV 随主元数

（LVN）的变化趋势，最佳选波段方法为窗口宽度为 6 的 GA 算法，共选择变量 336 个。算法 Random Frog、CARS 和 UVE，虽然对于训练数据有较好的训练结果，但是对于验证数据的预测却不如另外几种方法，即模型的泛化能力不够好。使用 GA 算法窗口宽度为 6 的

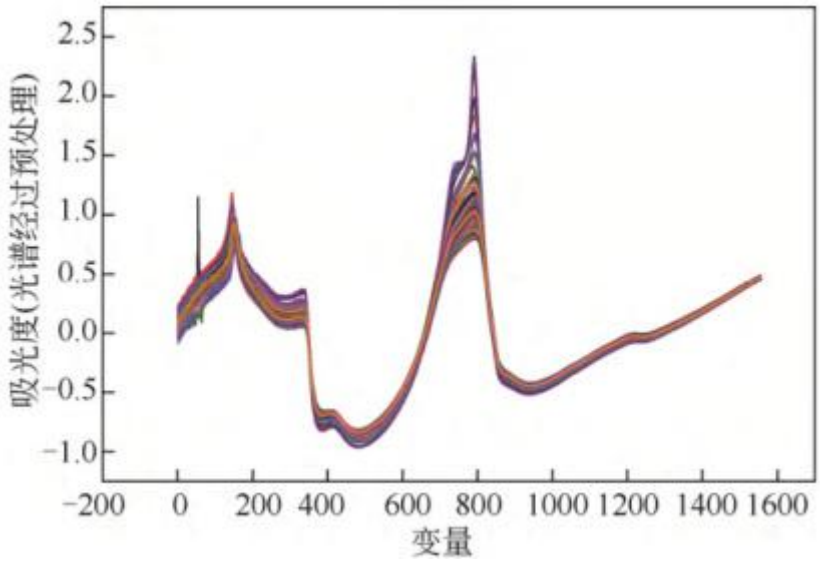


图 8 用于多糖建模样品预处理后的近红外光谱

表 3 不同变量选择方法建模结果

方法	选择变量数	窗口宽度	LVN	校正结果		交叉验证		验证结果
				R^2c	RMSEC	R^2CV	RMSECV	RMSEP
遗传算法	336	6	14	0.83466	21.055	0.69825	28.590	23.760
遗传算法	486	9	13	0.81036	22.544	0.65488	30.654	24.408
间隔偏最小二乘	400	50	15	0.85274	19.871	0.63743	31.798	27.315
间隔偏最小二乘	150	25	10	0.68215	29.193	0.56098	34.599	24.580
移动窗口偏最小二乘	541	50	5	0.44594	38.543	0.44906	38.452	24.492
移动窗口偏最小二乘	510	25	5	0.43917	38.777	0.39152	40.435	24.211
随机蛙	358	1	12	0.78291	24.126	0.63034	31.747	24.603
竞争自适应重加权采样	125	1	14	0.81336	22.370	0.66927	30.225	41.997
无信息变量消除法	808	1	12	0.77874	24.356	0.65168	30.739	28.073
无预处理	1357	1357	13	0.76824	24.922	0.53034	36.351	35.041

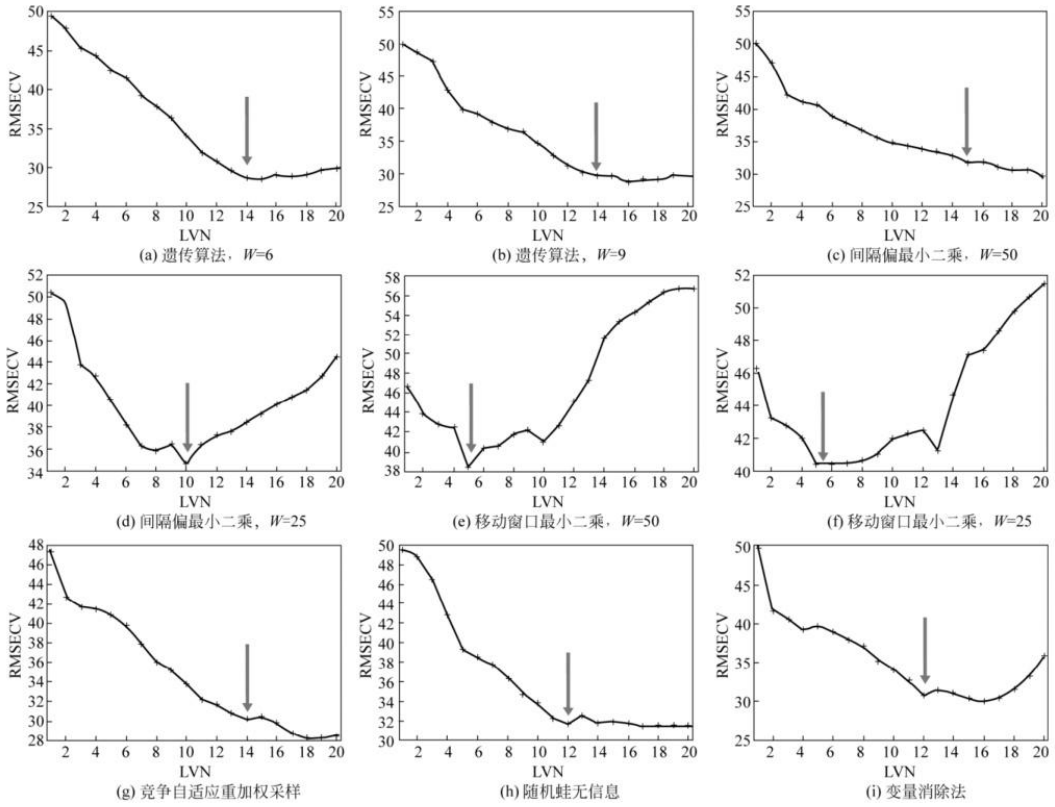


图 9 不同算法 RMSECV 随 LVN 变化趋势

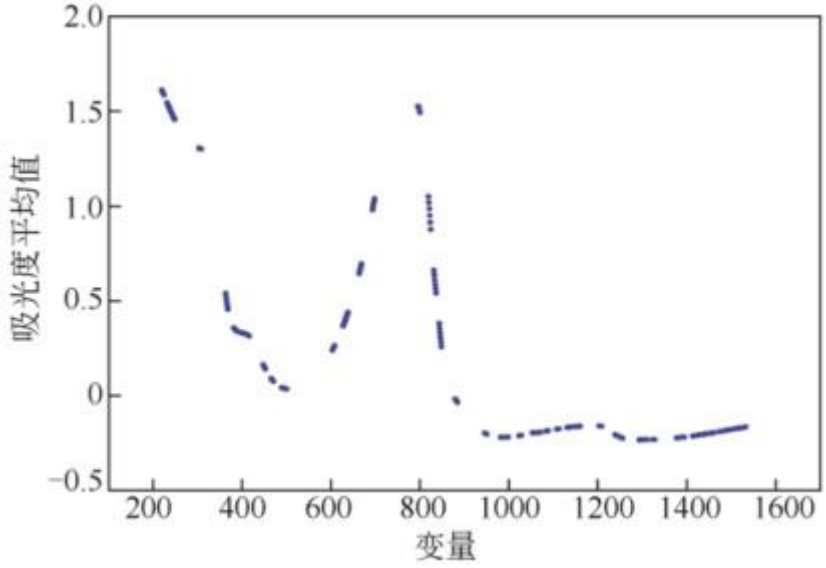


图 10 遗传算法选择波段结果

波段选择结果见图 10,由图 10 可知,该方法的结果并非传统手动波段选择方法中所选的一段或几段的组合,而是在整个波谱范围内都有变量的选择,这也从结果上验证了对于无特征吸收峰物质的近红外建模的特殊性。

3.3 建模结果与讨论

综合 3.1 节及 3.2 节所述，综合 RMSEC、

RMSECV、RMSEP、R^2C、R^2CV 等评价指标，多糖模型的最佳预处理方法为 Detrend（linear），最佳选波段方法为窗口宽度为 6 的遗传算法，所选变量为 336 个，根据交叉验证结果选择主元数 14。所选择的

预处理方法 Detrend（linear）较好的解决了近红外光谱基线漂移严重的问题，遗传算法选波段既避免了全波段建模引入过多冗余信息又规避了单变量波段选择对未知数据预测性能差的问题[34]。

3.3.1 多糖模型对验证集的预测结果与分析

使用上述参数建立模型并对 44 组多糖验证集进行预测，验证集样品中总糖含量化学值与预测值间相对误差见表 4,验证集的预测结果关系见图 11。由图 11 可见化学值与预测值之间线性关系良好,且由表 4 可见,44 组验证数据中仅有一组预测误差超出 10%，且为 10.742%，可近似认为所有数据的相对误差均控制在 10%以内,这在同领域中复合多糖的近红外预测研究及实际应用中是首次突破的。式(1)为预测值与化学值之间相对误差的计算公式。

表 4 验证集样品化学值、预测值间相对误差

序号	相对误差/%	序号	相对误差/%	序号	相对误差/%	序号	相对误差/%
1	3.777	12	2.189	23	-3.317	34	0.771
2	3.135	13	2.750	24	4.103	35	-1.652
3	-0.172	14	3.272	25	-6.003	36	-1.364
4	1.049	15	3.722	26	3.080	37	0.763
5	10.742	16	-3.170	27	7.019	38	2.807
6	-4.985	17	1.524	28	5.456	39	-2.336
7	5.506	18	-5.061	29	2.722	40	2.378
8	-4.516	19	-4.243	30	-0.851	41	0.181
9	-1.837	20	5.209	31	2.933	42	-1.910
10	-7.034	21	5.016	32	-1.344	43	-0.509
11	3.753	22	-0.128	33	-8.784	44	0.497

表 5 盲样样品化学值与预测值间相对误差

序号	相对误差/%	序号	相对误差/%	序号	相对误差/%	序号	相对误差/%
1	-4.734	6	3.194	11	-6.918	16	2.129
2	6.537	7	3.250	12	-8.086	17	0.458
3	-2.968	8	7.496	13	8.736	18	4.210
4	-4.633	9	0.331	14	-6.512	19	-2.937
5	-2.784	10	-1.440	15	-12.288	20	0.181

3.3.2 多糖模型的实际应用价值考察

为了考察模型在该口服液生产过程中的应用价值，于 2017 年 1 月在同一生产线同一浓缩暂存罐中采集口服液样品 20 批作为盲样，使用上文所述的模型进行多糖的预测，20 组盲样的化学值与预测值间相对误差见表 5。由表 5 中数据可知仅有一组数据的相对误差超出 10%，将该组样品的 3 组原始光谱及平均光谱输入模型中进行预测，图 12 为该样品 3 组原始光谱及一组平均光谱在校正集样品中的T^2-Q统计图，由其T^2-Q统计图可看到第 3 个原始光谱有很大概率是异常光谱，将该原始光谱剔除后用剩余两组光谱取平均值建模后的预测误差在 10%以内。这对模型的应用是有价值的，可在模型的应用界面设置显示T^2-Q统计图，离群的数据有 95%的概率为异常数据，则需多加注意。同时在应用过程中，要注意光谱采集的稳定，控制噪音及其他干扰，提高光谱稳定性，这对模型的应用是有价值的[35]。

4 口服液中可溶性固形物及 pH 的预测

将建立多糖模型的方法应用在该口服液中可溶性固形物含量及 pH 的预测中，对于可溶性固形物模型，可溶性固形物质量分数范围为 14.8%～17.3%，模型最佳预处理方法为 Detrend（linear），遗传算法（窗口宽度为 3）自动选波段，所选波段为变量 201～1557 间的 336 个变量，使用留一交叉验证，所选主元数为 10。校正集、验证集及盲样的可溶性固形物含量化学值与预测值间的相对误差见图 13。对于 pH 模型，pH 范围为 4.26～4.65，模型最佳预处理方法为一阶求导（窗口宽度为 7），遗传算法（窗口宽度为 2）自动选波段，所选波段为变量 201～1557 间的 346 个变量，使用留一交叉验证，所选主元数为 6。校正集、验证集及盲样的 pH 化学值与预测值间的相对误差见图 14。由图 13、图 14 可见，可溶性固形物质量分数的预测中，无论是校正集、验证集还是盲样，预测值与化学值之间的相对误差均控制在 5%内，对于 pH 模型，所有误差均可控制在 4%内，说明相应模型可对可溶性固形物质量分数及 pH 进行准确的预测，可为口服液生产过程中可溶性固形物质量分数及 pH 的控制提供可靠的技术支持。

5 结语及展望

NIR 光谱技术具有测量速度快、准确性高、仪器简单方便的特点，在食品药品、石油化工及生物等领域得到了广泛的应用。但是由于中药成分复杂，与化学药相比很多成分没有明显特征吸收峰，这对于应用 NIR 进行化学计量学建模提出了严峻的挑战。本文采用系统工程的方法，结合口服液生产流程，对数据前处理、预处理、选波段方法进行了评估，并对光谱的采集条件、取样点、重复性、质量指标化学值的重复性进行考察，应用近红外光谱技术结合偏最小二乘法实现了对口服液中的 3 种质量指标（多糖含量、可溶性固形物质量分数、pH）的实时在线检测。在多种预处理方法及波段选择方法中找到最佳参数，所建立的模型对多糖含量、可溶性固形物质量分数及 pH 进行预测，预测结果与化学值间的相对误差分别控制在 10%、5%及 4%内，可准确地对质量指标进行预测。除此之外，该方法对样品无破坏，经济环保，检测时间大大缩短，为口服液生产过程质量监控提供了有力支持，是中草药口服液指标控制尤其是多糖控制技术上的一大突破，并将该技术成功应用于实际生产过程中，真正实现了实时在线监控。

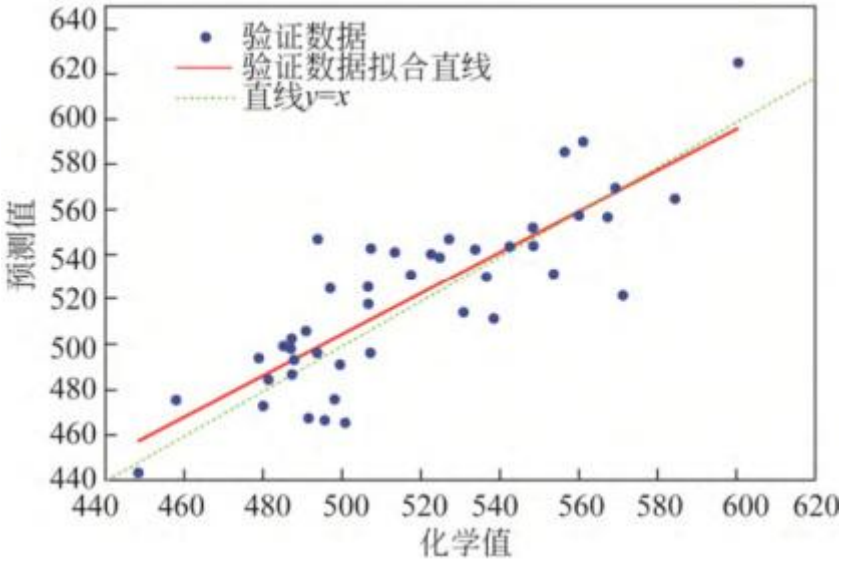


图 11 验证集样品预测结果相关关系

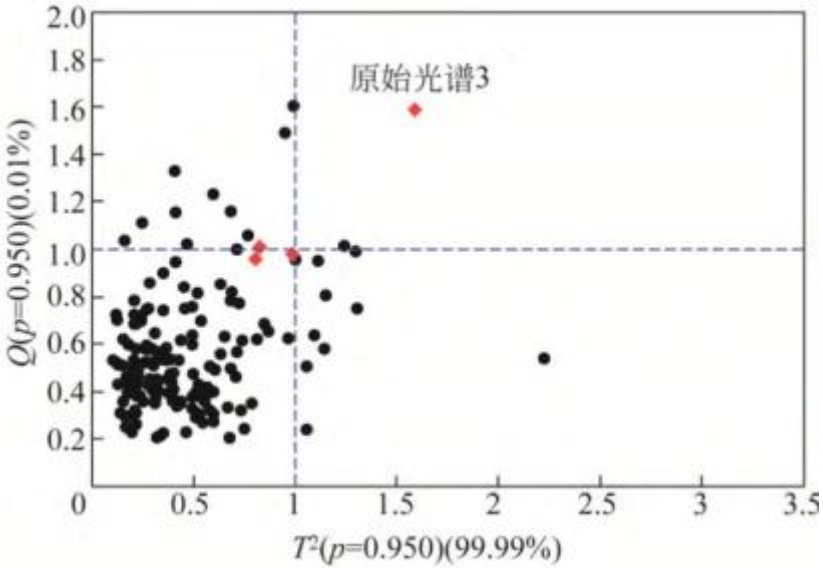


图 12 校正数据光谱 T^2-Q 统计图

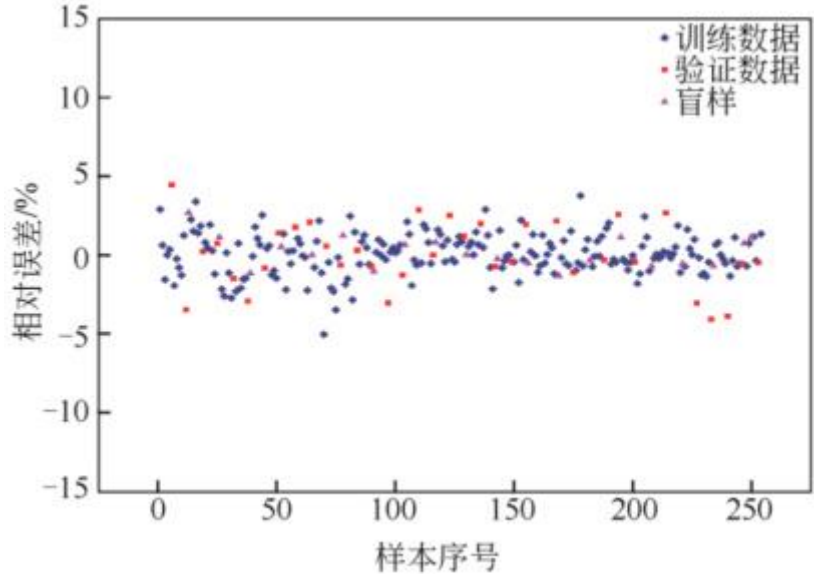


图 13 可溶性固形物模型校正集、验证集、盲样预测值与化学值相对误差结果

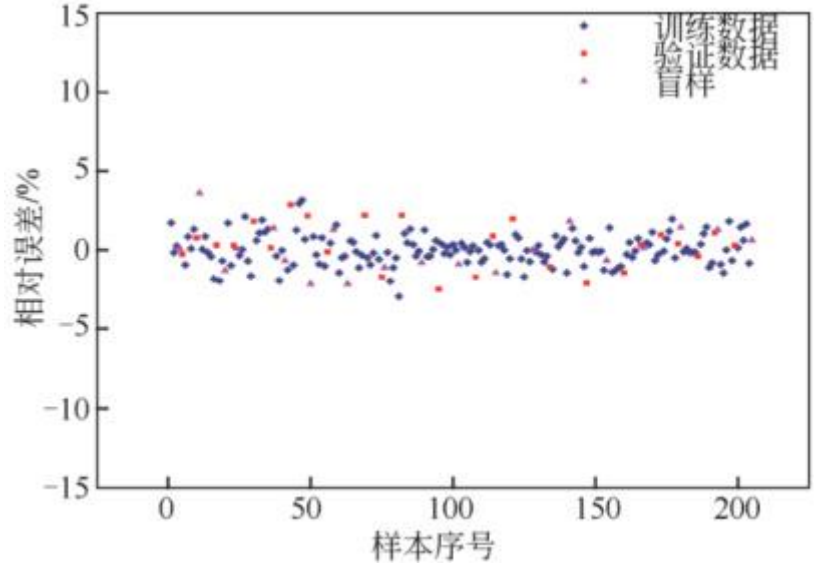


图 14 pH 模型校正集、验证集、盲样预测值与化学值相对误差结果

本文主要建立基于数据的模型，但是在模型的 实际应用过程中，随着生产条件的微小变化经过日 积月累，使用的模型是否继续可以对质量指标进行 准确预测，或者当模型无法长期满足生产时，模型 的更新周期是多久，这都需要更加深入的研究。此 外，通过近红外吸收与口服液中各种糖如单糖、二 糖的含量进行关联建模可对口服液中各种类型的糖 含量进行预测，这对口服液乃至中药行业药物生产 过程中的质量控制将具有十分重要的指导意义。

参考文献

[1] 周昭露, 李杰, 黄生权, 等. 近红外光谱技术在中药质量控制应用中的化学计量学建模: 综述和展望[J]. 化工进展, 2016, 35(6): 1627-1645.

ZHOU Z L, LI J, HUANG S Q, et al. Development of chemometric modelling in the application of NIR to the quality control of Chinese herbal medicine: literature review and future perspectives [J]. Chemical Industry and Engineering Progress, 2016, 35(6): 1627-1645.

[2] JAMROGIEWICZ M. Application of the near-infrared spectroscopy in the pharmaceutical technology[J]. Journal of Pharmaceutical & Biomedical Analysis, 2012, 66 (66): 1-10.

[3] WANG X Z. Data mining and knowledge discovery for process monitoring and control[M]. London: Springer, 1999.

[4] 李杰, 李传坤, 王春利, 等. 基于优化动态 PLS 的常压塔重整量化总缩点软测量研究[J]. 计算机与应用化学, 2016, 33(3): 27-29. LI J, LI C K, WANG C L, et al. The final boiling point soft-sensing of reforming feedstock on crude column based on optimized dynamic PLS [J]. Computers and Applied Chemistry, 2016, 33(3): 27-29.

[5] KAWANO S, FUJIWARA T, IWAMOTO M. Nondestructive determination of sugar content in plasma manufacture using near infrared (NIR) transmittance[J]. Engei Gakkai Zasshi, 1993, 62 (2): 465-470.

[6] KAWANO S, WATANABE H, IWAMOTO M. Determination of sugar content in intact peaches by near infrared spectroscopy with fiber optics in interactance mode[J]. Engei Gakkai Zasshi, 1992, 61 (2): 445-451.

[7] 何勇, 李晓丽, 邵咏妮. 基于主成分分析和神经网络的近红外光谱苹果品种鉴别方法研究[J]. 光谱学与光谱分析, 2006, 26 (5): 850-853.

HE Y, LI X L, SHAO Y N. Discrimination of varieties of apple using near infrared spectra based on principal component analysis and artificial neural network model[J]. Spectroscopy and Spectral Analysis, 2006, 26 (5): 850-853.

WU Z, XU E, WANG F, et al. Rapid determination of process variables of Chinese rice wine using FT-NIR spectroscopy and efficient wavelengths selection methods[J]. Food Analytical Methods, 2015, 8 (6): 1456-1467.

[9] MA C Y, WANG X Z. Simultaneous characterization of multiple properties of solid and liquid phases in crystallization processes using NIR[J]. Particuology, 2011, 9 (6): 589-597.

[10] LI R F, WANG X Z, ABEE B S B. Monitoring batch cooling crystallization using NIR: development of calibration models using genetic algorithm and PLS[J]. Particle & Particle Systems Characterization, 2008, 25 (4): 314-327.

[11] ABEBE S, WANG X Z, LI R, et al. The information content in NIR spectral data for slurry of organic crystals[J]. Powder Technology, 2008, 179 (3): 176-183.

[12] 战皓,柳梦婷,方婧,等. 近红外分析技术在中药鉴定和含量测定中的应用研究进展[J]. 中国实验方剂学杂志, 2015, 21 (12): 231-234.

ZHAN H, LIU M T, FANG J, et al. Application of near infrared analytical technique in identification and content determination of traditional Chinese medicine[J]. Chinese Journal of Experimental Traditional Medical Formulae, 2015, 21 (12): 231-234.

[13] 吴功煌, 新斯文, 乔延江. 近近红外光谱识别技术在中药质量控制中的应用研究进展[J]. 世界科学技术 (中医药现代化), 2010, 12 (2): 265-270.

WU G H, SHI X Y, QIAO Y J. The application and development of pattern recognition by near-infrared spectroscopy in quality control of traditional Chinese medicine [J]. World Science and Technology (Modernization of Traditional Chinese Medicine and Materia Medica), 2010, 12 (2): 265-270.

[14] 刘冰, 刘振尧, 朱乾华, 等. 近红外光谱法同时测定黄芪精口服液中黄民多糖和黄芪苗苷的含量[J]. 分析科学学报, 2011, 27(2): 195-198.

LIU B, LIU Z Y, ZHU Q H, et al. Simultaneous determination of estragalus polysaccharin and Astragaloside IV in Huangqiijing oral liquid by near infrared spectroscopy[J]. Journal of Analytical Science, 2011, 27 (2): 195-198.

[15] GALV O R K H, ARAUJO M C U, JOS G E, et al. A method for calibration and validation subset partitioning[J]. Talanta, 2005, 67 (4): 736.

[16] ZHANG C H, YUN Y H, FAN W, et al. Rapid analysis of polysaccharides contents in glycyrrhiza by near infrared spectroscopy and chemometrics[J]. International Journal of Biological Macromolecules, 2015, 79: 983.

[17] 刘爽悦, 李文龙, 翟斌海, 等. 基于近红外光谱的丹红注射液提取过程质量在线检测方法研究[J]. 中国中药杂志, 2013, 38(11): 1657-1662.

LIU S Y, LI W L, ZHAI H B, et al. On-line monitoring of extraction process of danhong injection based on near-infrared spectroscopy[J]. China Journal of Chinese Materia Medica, 2013, 38(11): 1657-1662.

[18] CHEN J, WANG X Z. A new approach to near-infrared spectral data

analysis using independent component analysis[J]. Journal of Chemical Information & Computer Sciences 2001 A1(4): 992-1001.

[19] ZHAO C, GAO F. Spectra calibration modeling and statistical analysis for cumulative quality interpretation and prediction[J]. AICHE Journal, 2012, 58 (2): 466-479.

[20] ZHAO C, GAO F, WANG F. Spectra data analysis and calibration modeling method using spectra subspace separation and multiblock independent component regression strategy[J]. AICHE Journal, 2011, 57 (5): 1202-1215.

[21] 保罗 戈培林. 化学计量学实用指南[M]. 北京: 科学出版社, 2012. GEMPERLINE Paul. Practical guide to chemometrics[M]. Beijing: Science Press, 2012.

[22] ISAKSSON T, NIES T. The Effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy[J]. Applied Spectroscopy, 1988, 42 (7): 1273-1284.

[23] ZHAO N, WU Z, CHENG Y, et al. MDL and RMSEP assessment of spectral pretreatments by adding different noises in calibration/validation datasets[J]. Spectrochimica Acta Part A: Molecular & Biomolecular Spectroscopy, 2016, 163: 20.

[24] JONES G, WILLETT P, GLEN R C, et al. Development and validation of a genetic algorithm for flexible docking[J]. Journal of Molecular Biology, 1997, 267 (3): 727-748.

[25] HAUPT R L, HAUPT S E. Practical genetic algorithms[M]. 2nd ed. New York: Wiley, 2004.

[26] NORGAARD L, SAUDLANDA, WAGNER J, et al. Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy[J]. Applied Spectroscopy, 2000, 54 (3): 413-419.

[27] JIANG J H, BERRY R J, SIESLER H W, et al. Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data[J]. Analytical Chemistry, 2002, 74 (14): 3555-3565.

[28] KASEMSUMRAN S, DU Y P, MURAYAMA K, et al. Near-infrared spectroscopic determination of human serum albumin γ-globulin and glucose in a control serum solution with searching combination moving window partial least-squares regression with applications to gene selection and disease classification[J]. Analytica Chimica Acta, 2004, 512 (2): 223-230.

[29] LI H D, XU Q S, LIANG Y Z. Random frog: An efficient reversible jump Markov Chain Monte Carlo-like approach for variable selection with applications to gene selection and disease classification[J]. Analytica Chimica Acta, 2012, 740: 20.

[30] LI H, LIANG Y, XU Q, et al. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration[J]. Analytica Chimica Acta, 2009, 648 (1): 77.

[31] AND V C, MASSARD D, NOORD O E D, et al. Elimination of uninformative variables for multivariate calibration[J]. Analytical Chemistry, 1996, 68 (21): 3851.

[32] OUSSAMA A, ELABADI F, PLATIKANOV S, et al. Detection of olive oil adulteration using FT-IR spectroscopy and PLS with variable importance of projection (VIP) scores[J]. Journal of the American Oil Chemists' Society, 2012, 89 (10): 1807-1812.

[33] LU B, CASTILLO I, CHIANG L et al. Industrial PLS model variable selection using moving window variable importance in projection[J]. Chemometrics & Intelligent Laboratory Systems, 2014, 135 (15): 90-109.

[34] TETKO I V, LIVINGSTONE D J, LUIK A I. Neural network studies. 1. Comparison of overfitting and overtraining[J]. Journal of Chemical Information & Computer Sciences, 1995, 35 (5): 826-833.

[35] 刘秀明, 王家俊, 李源栋, 等. GC/MS 指纹图谱结合 Hotelling T2 检验用于香精品质控制[J]. 烟草科技, 2015, 48 (7): 36-40. LIU X M, WANG J J, LI Y D, et al. Flavor quality control by combining GC/MS fingerprint with hotelling T2 test[J]. Tobacco Science Technology, 2015, 48 (7): 36-40.