

基于多建模流程共识的秸秆 NDF 和 ADF 含量 近红外速测特征波长选择策略

戴嘉伟¹ 杨增玲¹ 牛文娟² 葛辰俊¹ 马永正¹ 韩鲁佳¹

(1. 中国农业大学工学院, 北京 100083; 2. 华中农业大学工学院, 武汉 430070)

摘要: 中性洗涤纤维(NDF)和酸性洗涤纤维(ADF)含量是评价秸秆饲用品质的关键指标。为获得适用于低成本专用仪器的稳定特征谱区,本文提出一种基于多建模流程共识的波长选择策略。以 350 份秸秆样品为校正集,100 份独立样品为外部验证集,采集 10 000 ~ 4 000 cm⁻¹ 范围内的近红外光谱。将原始光谱、标准正态变量变换和 Savitzky-Golay 一阶导数分别与 CARS、MC-UVE、SPA、siPLS 和 VIP 组合,在 100 次 Monte Carlo 样本划分下重复建立 PLS 模型,并根据波长变量的跨流程重复选择特征识别共识波段。ADF 和 NDF 目标共识波段分别包含 449 和 434 个变量,占总变量的 14.96% 和 14.46%。固定共识波段模型在外部验证集上的 ADF 含量 RMSE、R²、RPD 和 Bias 分别为 13.54 g/kg、0.892、3.05 和 1.92 g/kg,NDF 对应值分别为 17.05 g/kg、0.890、3.03 和 -4.26 g/kg。共有波段主要反映多糖羟基、结合水和 C—H/C—O 合频等细胞壁基础信息,ADF 和 NDF 特异波段分别补充了纤维素/木质素富集区及半纤维素和结合水状态信息。该方法在显著减少变量的同时保留了稳定预测信息,可为小型近红外设备的候选波段和通道设计提供依据。

关键词: 秸秆; 低成本设备; 中性洗涤纤维; 酸性洗涤纤维; 近红外光谱; 特征波长

中图分类号: S816.17 文献标识码: A 文章编号: 1000-1298(2026)21-0037-12 OSID: 

Consensus-based Characteristic Wavelength Selection across Multiple Modeling Workflows for Rapid Near-infrared Determination of NDF and ADF Content in Crop Straw

Dai Jiawei¹ Yang Zengling¹ Niu Wenjuan² Ge Chenjun¹ Ma Yongzheng¹ Han Lujia¹

(1. College of Engineering, China Agricultural University, Beijing 100083, China

2. College of Engineering, Huazhong Agricultural University, Wuhan 430070, China)

Abstract: Neutral detergent fiber (NDF) and acid detergent fiber (ADF) are key indices of crop-straw feeding quality. To obtain stable spectral regions for low-cost instruments, a characteristic-wavelength selection strategy based on consensus across multiple modeling workflows was developed. A total of 350 straw samples were used for calibration, and 100 independent samples were used for external validation. Spectra from 10 000 cm⁻¹ to 4 000 cm⁻¹ were represented as raw, standard normal variate, or Savitzky-Golay first-derivative spectra and were paired with CARS, MC-UVE, SPA, siPLS, and VIP. Partial least-squares models were repeatedly constructed under 100 Monte Carlo splits, and recurrent selections across workflows were summarized as consensus bands. The ADF and NDF target-consensus sets contained 449 and 434 variables, accounting for 14.96% and 14.46% of all variables, respectively. On the independent validation set, fixed consensus-band models produced RMSE, R², RPD, and bias values of 13.54 g/kg, 0.892, 3.05, and 1.92 g/kg for ADF and 17.05 g/kg, 0.890, 3.03, and -4.26 g/kg for NDF. Shared bands mainly represented cell-wall information related to polysaccharide hydroxyl groups, bound water, and C—H/C—O combination bands. ADF-specific bands supplemented cellulose- and lignin-enriched responses, whereas NDF-specific bands reflected hemicellulose and bound-water status. The strategy retained stable predictive information with markedly fewer variables and provided

收稿日期: 2026-07-14 修回日期: 2026-07-18
基金项目: 国家重点研发计划项目(2022YFD2002101)
作者简介: 戴嘉伟(1998—),男,博士生,主要从事近红外光谱分析研究,E-mail: 1784674875@qq.com
通信作者: 杨增玲(1977—),女,教授,博士生导师,主要从事农业信息智能感知技术及装备研究,E-mail: yangzengling@cau.edu.cn

candidate regions for compact near-infrared instruments.

Key words: crop straw; low-cost instrument; neutral detergent fiber; acid detergent fiber; near-infrared spectroscopy; characteristic wavelengths

0 引言

中性洗涤纤维 (NDF) 和酸性洗涤纤维 (ADF) 是反刍动物饲料、牧草及木质纤维素类生物质品质评价的重要指标^[1-6]。传统洗涤纤维分析方法具有较高准确性,但检测过程依赖化学试剂处理、煮沸、过滤、干燥和称量等步骤,存在耗时长、操作环节多、试剂消耗量大和产生废液等问题,难以满足生物质原料快速分级、批量筛查和现场检测需求^[7]。

近红外光谱技术 (NIRS) 具有快速、无损、制样简单和易于在线集成等特点,已广泛应用于农产品、饲料和生物质成分的快速定量分析^[8-12]。在木质纤维素类样品中,近红外吸收主要来源于 O—H、C—H、N—H 等含氢基团倍频和合频跃迁,纤维素、半纤维素和木质素等组分在 1 000 ~ 2 500 nm 范围内形成宽带重叠响应^[13-18]。同时,秸秆样品的粒径、装填密度、水分状态和散射特性会影响漫反射光程,使光谱中同时包含化学组成信息和物理散射信息^[19-21]。因此,NDF 和 ADF 的近红外预测并非单一吸收峰对单一组分的直接响应,而是细胞壁多糖骨架、木质素结构、羟基环境、结合水状态和样品物理状态共同作用的综合表征^[22-25]。

已有研究表明,仪器类型、光谱范围和变量选择策略均会影响模型结构及特征波段分布^[26-30]。从化学组成来看,NDF 和 ADF 既具有共同的细胞壁来源,也具有不同的组分边界^[31-32]。其近红外信息可能共同分布于纤维素、半纤维素和木质素的多糖骨架区、羟基倍频区及 C—H/C—O 合频区,也可能因半纤维素参与程度、木质素含量和纤维素结构差异而形成目标特异响应。然而,当前变量筛选研究多基于单一预处理方法结合单种算法进行独立优化,所得特征波段往往随预处理方式或算法机制的改变而发生剧烈漂移,缺乏跨流程条件下的稳健性检验。这种“单一路径依赖”的筛选结果,不仅难以区分 NDF 与 ADF 在细胞壁光谱响应中的共性基础与个性差异,更因其对特定建模条件的过度适应而难以直接固化为低成本专用设备所需的固定波长通道,成为制约洗涤纤维指标近红外检测由实验室研究型平台向小型化、专用化仪器转化的关键瓶颈。

针对上述问题,本文以秸秆样品 NDF 和 ADF 含量检测为对象,构建一种面向低成本近红外设备选型的多流程建模共识波段识别方法。研究采用

Raw、SNV 和 Savitzky - Golay 一阶导数 3 种光谱表达方式,结合 CARS、MC - UVE、SPA、siPLS 和 VIP 5 种变量选择算法,在 100 次 Monte Carlo 样本划分条件下重复建立 ADF 和 NDF 的 PLS 模型,统计不同建模流程下各波长变量的重复选择频率,识别跨预处理、跨算法和跨样本划分稳定出现的共识波段,并结合独立外部验证评价其预测信息承载能力。同时,将共识波段模型与全谱及常用单一波长选择方法进行比较。研究目标在于提取 ADF 和 NDF 近红外预测中稳定、可解释且便于硬件转化的候选谱区,为设计秸秆洗涤纤维指标的低成本快速检测近红外设备提供依据。

1 材料与方法

1.1 样品来源与样本集构建

本研究共使用 450 份秸秆样品,其中 350 份作为校正集,100 份作为独立外部验证集。350 份校正集样品采自北京市、河北省、山西省、山东省和天津市,包含小麦秸秆 271 份、玉米秸秆 67 份和水稻秸秆 12 份;100 份独立外部验证集样品同样采自上述地区,包括小麦秸秆 31 份、玉米秸秆 45 份、水稻秸秆 20 份、棉花秸秆 2 份和油菜秸秆 2 份。

1.2 样本参考值测定

秸秆样品经干燥、粉碎和过筛后密封保存,测定前充分混匀。NDF 含量参照 ISO 16472:2006^[33]、ADF 含量参照 ISO 13906:2008^[34] 测定。结果以干基质量比表示,单位为 g/kg。

1.3 近红外光谱采集

近红外光谱采用傅里叶变换近红外光谱仪采集。采集前将样品充分混匀并装入样品杯,保持表面平整,以降低装样状态对漫反射光程的影响。光谱采集范围为 10 000 ~ 4 000 cm⁻¹,分辨率为 8 cm⁻¹,波数间隔为 2 cm⁻¹。每份样品重复扫描 3 次后取平均光谱作为建模光谱,采集过程中室内温度保持在 23 ~ 25℃。

1.4 光谱预处理方法

为构建可比较的光谱输入,本研究设置原始光谱 (Raw)、标准正态变量变换 (Standard normal variate,SNV) 和 Savitzky - Golay 一阶导数 (SGD1) 3 种光谱表达方式。Raw 不进行数学变换,用于保留原始吸收、散射和基线信息;SNV 以单条光谱为对象进行均值中心化和标准差归一化,用于降低粒

径、装样密度和漫反射光程差异引起的散射影响^[35-36]；SGD1 用于削弱低频基线漂移并增强重叠谱带的局部变化信息^[37-38]，其窗口长度设为 15 个波数点，多项式阶数设为 2，导数阶次设为 1。

1.5 特征波长选择与对照方法

在 Raw、SNV 和 SGD1 这 3 种光谱表达基础上，分别采用竞争性自适应重加权采样（CARS）、蒙特卡洛无信息变量消除（MC-UVE）、连续投影算法（SPA）、联合区间偏最小二乘（siPLS）和投影变量重要性（VIP）进行特征波长选择。上述方法分别从变量贡献、回归系数稳定性、共线性控制、连续区间信息和 PLS 投影贡献等角度提取目标相关变量。3 种光谱表达方式与 5 种变量选择算法组合形成 15 组建模流程。还选取了遗传算法（Genetic algorithm, GA）^[39]和随机森林（Random forest, RF）特征重要性^[40]对照。

1.6 模型的建立

本研究采用偏最小二乘法（PLS）建立 NDF 和 ADF 含量检测模型。近红外光谱变量数量多、相邻波长间共线性强，且吸收信息通常以宽带重叠形式存在。PLS 可将高维光谱矩阵压缩为少量潜变量，在保留目标变量相关信息的同时降低共线性影响。与非线性黑箱模型相比，PLS 模型结构相对稳定，潜变量、回归系数和变量贡献可用于解释波段与目标属性之间的关系，适合本研究对共识波段稳定性和光谱可解释性的分析需求。因此，对 NDF 和 ADF 含量分别建立单响应 PLS 模型，作为不同变量选择流程及共识波段模型的统一回归框架。

350 份校正样品用于 Monte Carlo 内部划分、光谱预处理、变量选择、共识波段识别和模型建立，在 100 次 Monte Carlo 划分中，350 份校正样品按 7:3 划分为 245 份和 105 份两部分。100 份独立外部验证样品不参与任何预处理参数确定、变量选择、共识阈值设定、潜变量数优化和模型筛选，仅用于固定模型的泛化能力评价。为避免信息泄漏，每次 Monte Carlo 划分中，光谱预处理参数、光谱变量标准化参数和目标变量标准化参数均由内部校正集计算，并应用于对应的内部预测集。变量选择算法仅在内部校正集内执行，筛选得到的波数变量用于重新建立 PLS 模型。PLS 潜变量数通过内部校正集交叉验证确定。

全谱基准模型采用 10 000 ~ 4 000 cm⁻¹ 范围内的全部 3 001 个 Raw 光谱变量，不进行特征波长选择。

模型性能采用决定系数 R^2 、均方根误差（RMSE）和剩余预测偏差（RPD）评价。独立外部验

证阶段同时计算偏差（Bias），用于表征预测结果的系统偏移。

1.7 共识波段识别与跨预处理支持评价

为识别 NDF 和 ADF 含量预测中稳定出现的光谱区段，本研究构建了跨光谱表达方式和跨变量选择算法的共识波段识别方法。对每个目标变量，分别记录 Raw、SNV 和 SGD1 条件下 5 种变量选择算法在 100 次 Monte Carlo 划分中得到的特征波数集合，并映射至完整波数轴，形成“波数变量 × 建模流程”的选择矩阵。

对于第 i 个波数变量，首先计算其在同一建模流程 100 次划分中的选择频率；随后统计其在 15 组流程中的平均选择频率、支持流程数、支持预处理数和支持算法数。其中，平均选择频率反映该变量被重复选择的程度，支持流程数反映该变量是否跨建模流程稳定出现，支持预处理数和支持算法数用于辅助判断该变量的稳定性是否依赖单一光谱表达方式或单一变量选择算法。共识强度由平均选择频率和支持流程比例共同确定，用于表征每个波数变量的跨流程稳定性。

由于近红外吸收多表现为宽带重叠响应，而变量选择算法在重复划分中可能产生少量孤立波点，本研究对共识强度曲线进行 31 点中心移动平滑。原始光谱变量间隔为 2 cm⁻¹，31 点窗口约对应 60 cm⁻¹ 的波数跨度，可降低随机划分或算法扰动导致的尖峰影响，同时保留连续吸收区间的主要轮廓。平滑处理仅用于增强共识强度曲线的谱区连续性，支持流程数仍依据原始选择矩阵统计。

平滑共识强度高于目标变量 85% 分位阈值，且支持流程数不少于 5 个的变量点定义为强共识点。相邻强共识点按光谱轴连续性合并为共识波段；若相邻强共识区间的间隔不超过 8 个变量点，则视为同一连续谱区，以避免单个或少量未达阈值变量将同一宽吸收带切断。若间隔超过 8 个变量点，则划分为不同共识波段。

NDF 和 ADF 分别进行目标共识波段识别。两个目标变量强共识区间的重叠部分定义为共同波段，用于表征 NDF 和 ADF 共享的细胞壁基础信息；仅在 ADF 中达到强共识标准的区间定义为 ADF 特异波段，仅在 NDF 中达到强共识标准的区间定义为 NDF 特异波段，用于分析二者因组分边界差异形成的目标特异光谱响应。

为评价各共识区段的跨预处理支持，对每个区段分别计算 Raw、SNV 和 SGD1 组内 5 种变量选择算法的区段平均选择频率；当某算法在该区段内的平均选择频率不低于 0.05 时，记为一种支持算法。

1.8 共识变量补回敏感性分析

为检验共识波段筛选是否遗漏仅在与已保留变量联合建模时发挥作用的互补光谱信息,本研究在Raw光谱条件下设置共识变量补回敏感性试验。ADF和NDF含量分别以共识变量作为基础变量集,对基础变量集之外的变量按照相应目标的平滑共识强度由高到低排序并逐步补回,使各阶段变量总数分别达到总变量数的25%、40%、60%、80%和100%。通过比较共识基础组、各级补回组和全谱组的预测性能,评价被排除变量中是否存在能够带来净预测收益的协同信息。

为区分补回变量的共识强度信息与单纯增加变量数量的影响,在25%、40%、60%和80%共4个中间阶段设置等变量数随机补回对照。对每个目标和每个阶段,从其余变量中随机补足至相同规模并重复100次,通过比较同一变量数下强度排序补回组与随机补回组的预测结果,判断补回效果是否来源于共识强度排序,而非变量数量增加本身。

2 结果与讨论

2.1 秸秆 NDF 和 ADF 含量及其近红外光谱响应

NDF和ADF含量的近红外预测基础来源于木质纤维素组分的化学变异及其在近红外区的复合吸收响应。图1a和图1b分别为校正集与独立外部验证集的样品种类及洗涤纤维含量分布。由图1b可

知,350份校正样品的NDF和ADF含量范围分别为500.49~846.67 g/kg和319.80~581.26 g/kg。100份独立外部验证样品对应范围分别为568.45~837.16 g/kg和370.76~552.00 g/kg。两个样本集覆盖了洗涤纤维指标的主要变异区间,为后续多建模流程下的稳定波段识别提供化学梯度基础。如图1c所示,校正集中NDF与ADF含量的Pearson相关系数 r 为0.835,说明二者共享较强的细胞壁基础信息。二者既相关又不完全等同,为后续区分共同波段和目标特异波段提供了化学依据。

如图1d所示,秸秆平均光谱在1000~2500 nm范围内呈现典型木质纤维素样品的宽带重叠特征。在1471、1724、1929、2104、2276、2332、2384、2493 nm附近存在较明显吸收带,主要与O—H、C—H等含氢基团的倍频和合频吸收有关。高低含量组差异光谱进一步表明,NDF和ADF的响应区间并不完全重合(图1e)。其中,1410~1460 nm附近主要对应纤维素和半纤维素羟基O—H伸缩振动的一倍频;1660~1780 nm区间主要与亚甲基、次甲基及芳香环C—H伸缩振动的一倍频有关;1880~1950 nm区间多与O—H伸缩和变形振动的合频相关,反映结合水及氢键网络状态;2160~2400 nm区间主要对应多糖骨架中C—H、C—O和C—C振动的合频响应。光谱响应表明,NDF和ADF的近红外信息是多糖骨架、木质素结构、羟基环境和结合水状

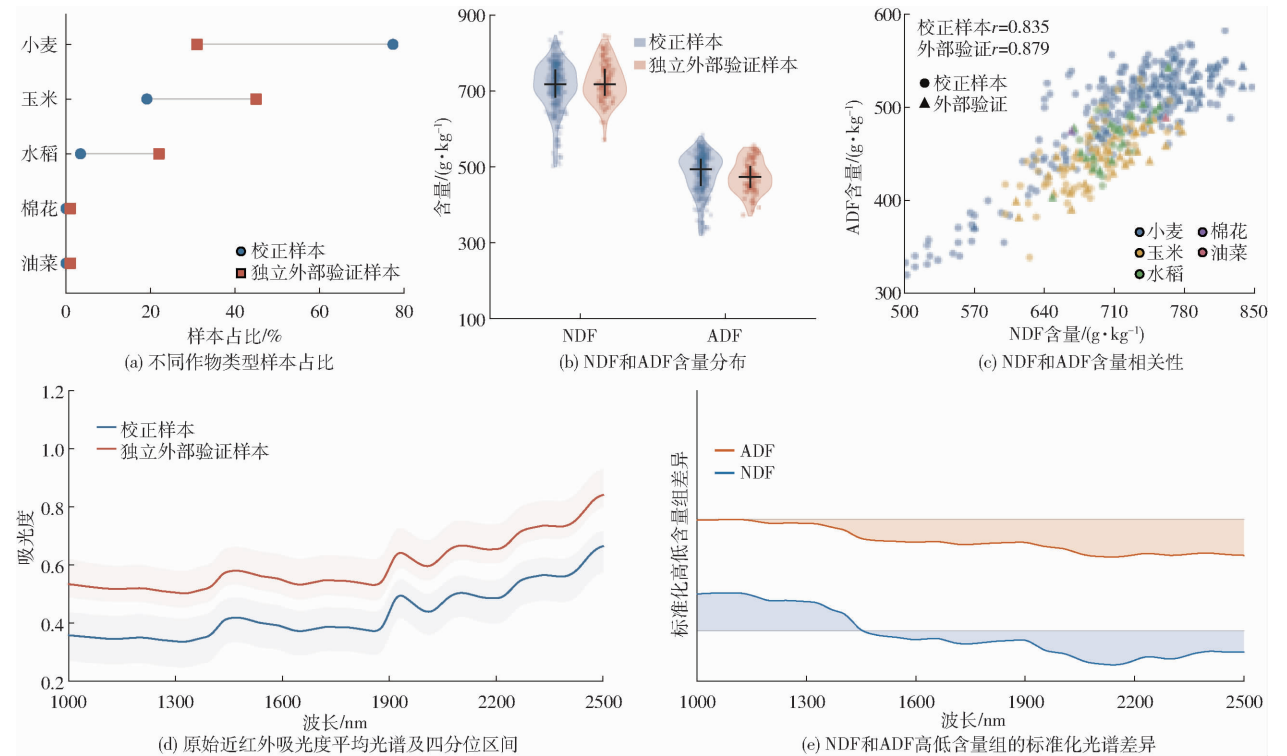


图1 校正样本与独立外部验证样本的组成、纤维化学特征及近红外光谱特征

Fig.1 Composition, detergent-fiber chemical characteristics, and near-infrared spectral characteristics of calibration and independent external validation samples

态共同作用的复合表征。平均光谱和高低含量组差异光谱表明两类指标存在响应差异,但仍难以判断哪些波段在不同预处理、不同变量选择算法和不同样本划分下稳定出现。因此,后文进一步基于多建模流程统计波长变量的重复选择情况,识别 ADF/NDF 含量检测中的稳定共识波段。

2.2 多建模流程及对照方法的预测与外部验证表现

15 组建模流程在 100 次 Monte Carlo 样本划分下重复建立 PLS 模型的内部测试结果见表 1,不同共识波段集合的信息保留能力及独立外部验证结果见表 2。如图 2 和表 1 所示,不同预处理方法和变量选择算法对 ADF/NDF 含量预测结果均存在明显差异。ADF 含量预测中,siPLS 和 VIP 类流程整体表现较稳定,其中 Raw - siPLS、SGD1 - siPLS、

SGD1 - VIP 的内部测试 RMSE 分别为 16.80、17.09、17.47 g/kg,说明保留连续区间信息和 PLS 投影贡献的变量选择方法更适合表征 ADF 的宽带组合吸收。NDF 含量预测中,Raw - VIP 和 Raw - siPLS 的 RMSE 分别为 25.45 g/kg 和 25.46 g/kg,均处于较低水平,而 Raw - SPA 的 RMSE 达到 36.55 g/kg。说明过度稀疏的孤立波点选择可能破坏 NDF 所依赖的连续谱区信息。不同预处理方式未表现出对两个指标均一致占优的规律。ADF 含量在 Raw、SNV、SGD1 条件下的变量选择流程平均 RMSE 分别为 18.51、18.66、18.82 g/kg。NDF 含量对应的平均 RMSE 分别为 28.97、28.18、28.48 g/kg。最优流程分布于不同预处理和算法组合中,说明单一建模流程输出的特征波段具有一定流程依赖性。

表 1 15 组变量选择流程的内部测试结果

Tab.1 Internal test results of 15 variable-selection workflows

指标	预处理	CARS		MC - UVE		SPA		siPLS		VIP	
		RMSE/	R^2	RMSE/	R^2	RMSE/	R^2	RMSE/	R^2	RMSE/	R^2
		(g·kg ⁻¹)		(g·kg ⁻¹)		(g·kg ⁻¹)		(g·kg ⁻¹)		(g·kg ⁻¹)	
ADF 含量	Raw	21.02	0.839	18.55	0.875	17.67	0.883	16.80	0.898	18.53	0.875
ADF 含量	SNV	21.21	0.837	18.24	0.879	18.38	0.877	17.73	0.886	17.73	0.886
ADF 含量	SGD1	22.37	0.818	19.29	0.865	17.87	0.884	17.09	0.894	17.47	0.889
NDF 含量	Raw	30.00	0.770	27.39	0.808	36.55	0.655	25.46	0.834	25.45	0.834
NDF 含量	SNV	32.00	0.738	28.50	0.791	27.04	0.812	26.70	0.817	26.65	0.818
NDF 含量	SGD1	31.91	0.739	28.85	0.787	27.48	0.807	27.11	0.812	27.06	0.813

注:表中数值为 100 次 Monte Carlo 校正集内部留出样本测试均值。

表 2 共识波段模型的信息保留能力与独立外部验证结果

Tab.2 Information retention and independent external validation of consensus-band models

指标	波段集合	变量数 (变量比例)	MC100			外部验证				
			RMSE/	R^2	RPD	主成 分数	RMSE/	R^2	RPD	Bias/
			(g·kg ⁻¹)				(g·kg ⁻¹)			(g·kg ⁻¹)
ADF 含量	共同波段	225(7.50%)	19.30 ± 1.27	—	—	—	—	—	—	—
ADF 含量	ADF 特异波段	224(7.46%)	19.91 ± 1.37	—	—	—	—	—	—	—
ADF 含量	ADF 目标共识波段	449(14.96%)	17.26 ± 1.19	0.892 ± 0.020	3.09 ± 0.28	17	13.54	0.892	3.05	1.92
NDF 含量	共同波段	225(7.50%)	25.39 ± 1.74	—	—	—	—	—	—	—
NDF 含量	NDF 特异波段	209(6.96%)	26.88 ± 1.74	—	—	—	—	—	—	—
NDF 含量	NDF 目标共识波段	434(14.46%)	25.45 ± 1.69	0.834 ± 0.035	2.51 ± 0.26	13	17.05	0.890	3.03	-4.26

注:表中列出 ADF 和 NDF 目标共识波段模型的固定外部验证结果,基准方法结果见表 3。

如表 2 所示,目标共识波段模型在大幅压缩变量数量的条件下仍保持较好的预测能力。ADF 目标共识波段包含 449 个变量,占总光谱变量的 14.96%,其内部测试 RMSE、 R^2 和 RPD 分别为 (17.26 ± 1.19) g/kg、0.892 ± 0.020 和 3.09 ± 0.28。NDF 目标共识波段包含 434 个变量,占总光谱变量的 14.46%,其内部测试 RMSE、 R^2 和 RPD 分别为 (25.45 ± 1.69) g/kg、0.834 ± 0.035 和 2.51 ± 0.26。

独立外部验证进一步检验了共识波段的预测信息承载能力。ADF 目标共识波段模型包含 449 个

变量,最佳潜变量数为 17,在独立外部验证样品上的 RMSE、 R^2 、RPD 和 Bias 分别为 13.54 g/kg、0.892、3.05 和 1.92 g/kg。NDF 目标共识波段模型包含 434 个变量,最佳潜变量数为 13,对应指标分别为 17.05 g/kg、0.890、3.03 和 -4.26 g/kg。

在 350 份校正样品中,全谱 Raw - PLS 与共识波段模型的 RMSECV 接近 (ADF: 16.31 g/kg 和 16.90 g/kg; NDF: 24.37 g/kg 和 24.42 g/kg),独立外部验证结果见表 3。共识模型的变量数由 3001 个分别减少至 449、434 个。

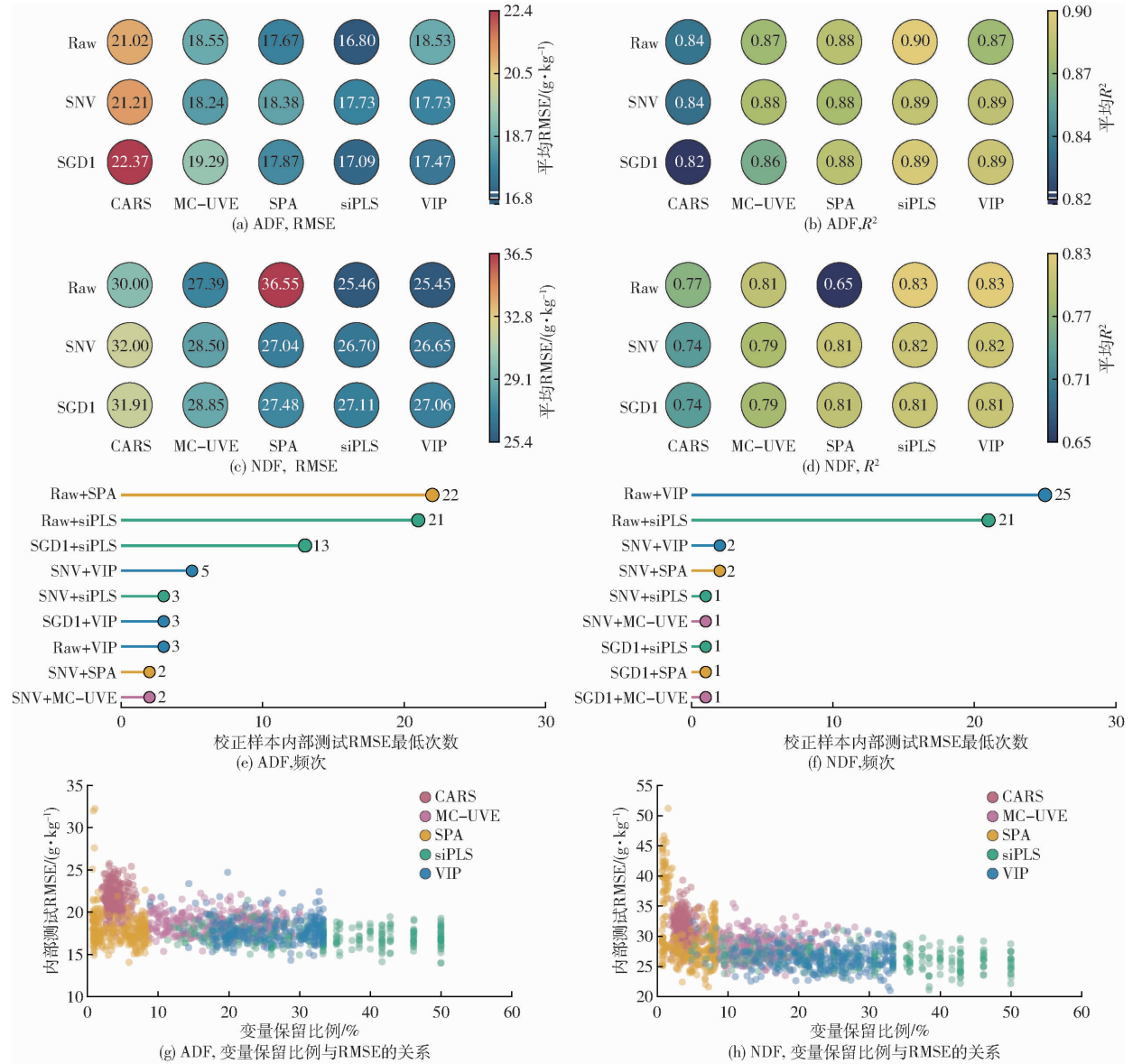


图 2 100 次 Monte Carlo 随机划分下多建模流程的预测表现与流程不稳定性

Fig. 2 Predictive performance and workflow instability of multiple modeling workflows across 100 Monte Carlo splits

表 3 共识波段、全谱与单一波长选择方法的独立外部验证结果

方法	ADF					NDF				
	变量数	主成分数	RMSE/ ($\text{g} \cdot \text{kg}^{-1}$)	R^2	RPD	变量数	主成分数	RMSE/ ($\text{g} \cdot \text{kg}^{-1}$)	R^2	RPD
共识波段	449	17	13.54	0.892	3.05	434	13	17.05	0.890	3.03
全谱 PLS	3 001	13	17.59	0.817	2.35	3 001	14	20.01	0.848	2.58
CARS	89	19	27.40	0.556	1.51	154	18	37.42	0.470	1.38
MC-UVE	521	18	18.78	0.791	2.20	512	18	28.64	0.689	1.80
SPA	26	17	15.87	0.851	2.60	34	23	41.23	0.357	1.25
siPLS	1 500	14	14.89	0.869	2.78	1 501	12	26.20	0.740	1.97
VIP	1 000	14	18.73	0.793	2.21	651	13	17.93	0.878	2.88
GA	195	14	16.08	0.847	2.57	46	6	53.51	-0.084	0.97
RF 特征重要性	900	15	17.18	0.825	2.41	900	13	20.45	0.842	2.53

不同单一波长选择方法在 ADF 和 NDF 上的表现并不一致。ADF 含量预测中, siPLS 获得单一方

法中的最低外部验证 RMSE, 为 14.89 g/kg, 其 R^2 和 RPD 分别为 0.869 和 2.78; 共识波段模型的 RMSE

为 13.54 g/kg。NDF 含量预测中,VIP 的外部验证表现最接近共识波段模型,其 RMSE、 R^2 和 RPD 分别为 17.93 g/kg、0.878 和 2.88;共识波段模型的 RMSE 较其降低 4.86%,同时变量数由 651 个减少至 434 个。RF 特征重要性对照在 ADF 和 NDF 上的 RMSE 分别为 17.18 g/kg 和 20.45 g/kg,均未达到共识波段模型水平。其余方法在两个目标间的性能排序变化较大,说明基于单一算法得到的变量集合具有一定目标依赖性。

上述结果表明,跨预处理及跨算法重复支持的共识波段在 ADF 含量和 NDF 含量两个目标上均获得了较为一致的外部验证表现。

2.3 共识波段识别、稳定性及低成本设备选型意义

如图 3a 和图 3b 所示,不同预处理方式和变量选择算法对波长变量的选择结果存在明显差异,说明单一建模流程获得的特征波段容易受到流程设定影响。为降低这种偶然性,本研究以平均选择频率和支持流程比例共同构建共识强度,经 31 点移动平

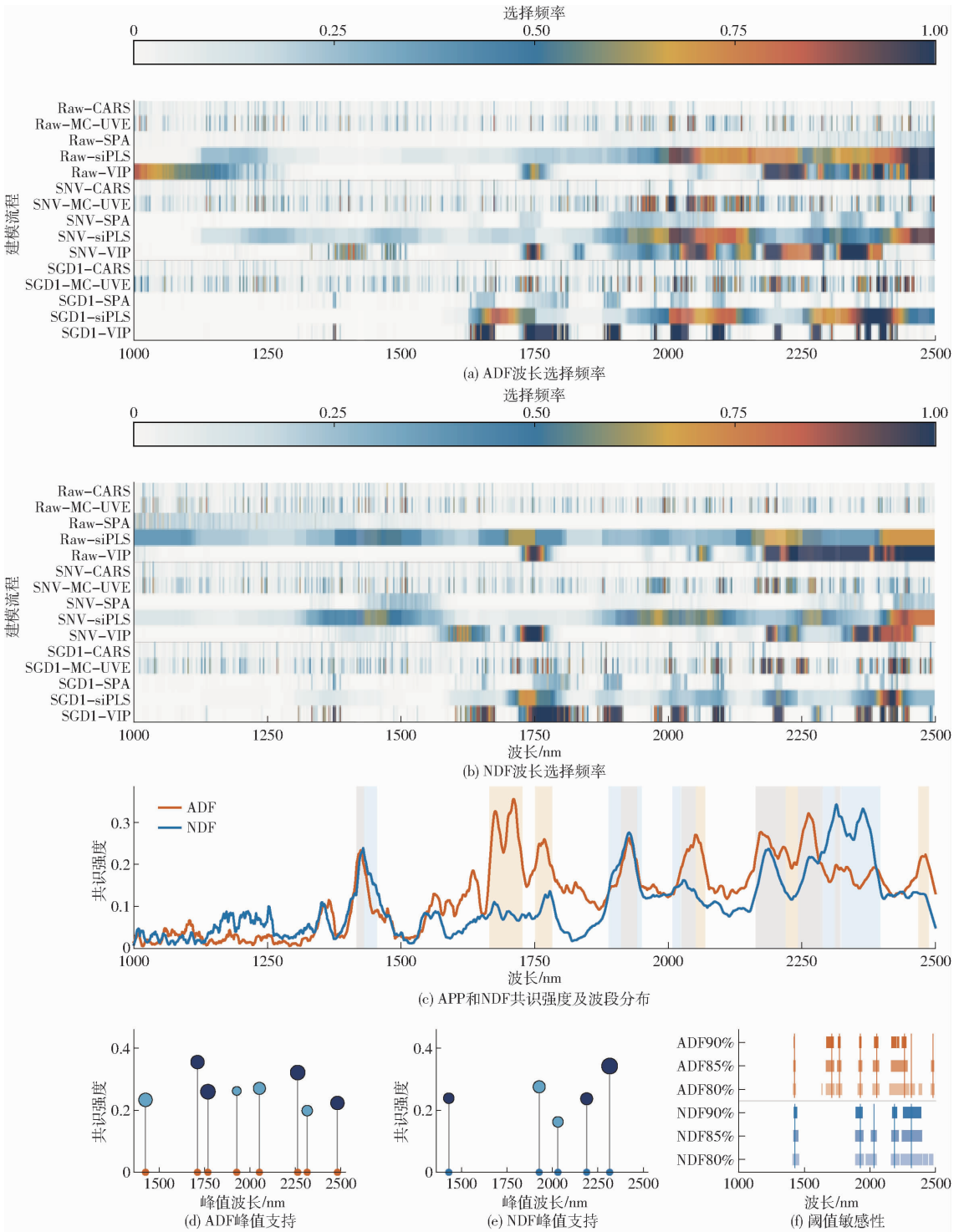


图 3 ADF 和 NDF 特征波段的共识指纹图谱及阈值稳健性分析

Fig. 3 Consensus fingerprint maps and threshold robustness analysis of ADF and NDF characteristic bands

滑后,将共识强度高于目标变量 85% 分位阈值且支持流程数不少于 5 个的波长点定义为强共识点,并按光谱轴连续性合并为共识波段。

ADF 共识波段包含 449 个变量,占原始扫描变量总数的 14.96%,主要分布于 1 416.4 ~ 1 431.4 nm、1 665.6 ~ 1 727.1 nm、1 751.3 ~ 1 782.5 nm、1 912.0 ~ 1 941.7 nm、2 024.3 ~ 2 068.7 nm、2 160.8 ~ 2 288.3 nm、2 312.7 ~ 2 322.3 nm、2 467.9 ~ 2 487.6 nm。如图 3c 和图 3d 所示,1 710.6 nm 和 2 262.4 nm 附近为 ADF 最稳定的峰值响应点,分别由 11 ~ 12 个建模流程支持,并覆盖 3 种预处理方式和 5 种变量选择算法。这说明 ADF 的稳定光谱信息主要分布于 C—H 一倍频、C—H/C—O 合频及长波骨架结构响应区,与 ADF 含量主要表征纤维素和木质素等难降解组分的化学定义相一致。NDF 共识波段包含 434 个变量,占原始扫描变量总数的 14.46%,主要分布于 1 416.8 ~ 1 455.2 nm、1 888.9 ~ 1 950.1 nm、2 008.0 ~ 2 050.9 nm、2 163.6 ~ 2 219.3 nm、2 243.2 ~ 2 396.9 nm。如图 3c 和图 3e 所示,2 243.2 ~ 2 396.9 nm 为 NDF 最突出的连续共识区段,峰值波长 2 314.8 nm 由 13 个建模流程支持。该区间位于长波 C—H/C—O 复合合频区,与纤维素、半纤维素及总细胞壁多糖骨架响应密切相关。相较 ADF,NDF 共识波段在长波合频区的连续覆盖范围更宽,说明 NDF 含量检测更

依赖总细胞壁基质、多糖骨架和半纤维素参与所形成的复合光谱信息。

ADF 与 NDF 共识波段可进一步分为共同波段和目标特异波段。如图 3c 所示,共同波段包含 225 个变量,占原始扫描变量总数的 7.50%,分布于 1 416.8 ~ 1 431.4 nm、1 912.0 ~ 1 941.7 nm、2 024.3 ~ 2 050.9 nm、2 163.6 ~ 2 219.3 nm、2 243.2 ~ 2 288.3 nm、2 312.7 ~ 2 322.3 nm。这些区间在 ADF 和 NDF 中均达到强共识标准,说明其承载了两类洗涤纤维指标共享的细胞壁基础信息。共同波段主要对应多糖羟基 O—H 一倍频、O—H 合频以及 C—H/C—O 复合合频,反映纤维素、半纤维素和木质素共同构成的细胞壁基质信息(图 4,其中图 4b 中 C 表示共同波段,A 表示 ADF 特异波段,N 表示 NDF 特异波段,各类波段分别按波长从小到大编号)。ADF 特异波段主要位于 1 665.6 ~ 1 727.1 nm、1 751.3 ~ 1 782.5 nm、2 051.7 ~ 2 068.7 nm、2 220.2 ~ 2 242.2 nm、2 467.9 ~ 2 487.6 nm,集中于 C—H 一倍频和长波合频区,主要补充纤维素和木质素富集区响应。NDF 特异波段主要位于 1 431.8 ~ 1 455.2 nm、1 888.9 ~ 1 911.3 nm、1 942.5 ~ 1 950.1 nm、2 008.0 ~ 2 023.5 nm、2 289.4 ~ 2 311.6 nm、2 323.4 ~ 2 396.9 nm,集中于 O—H 倍频/合频和长波 C—H/C—O 合频区,主要补充半纤维素、结合水状态和总细胞壁基质共同形成的复合响应信息。

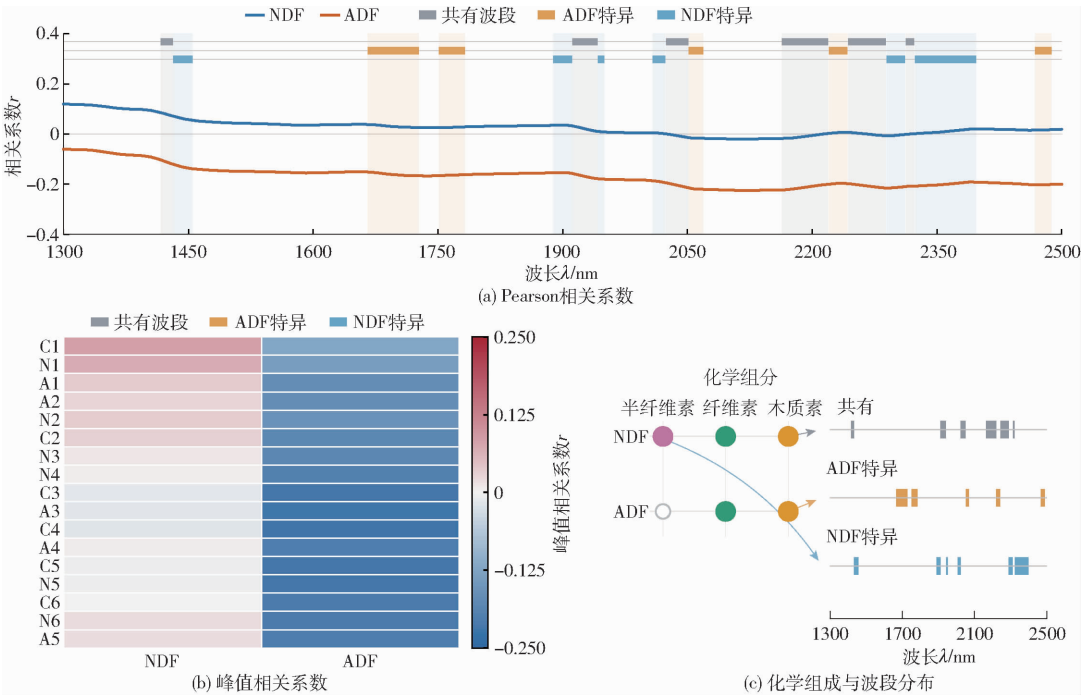


图 4 ADF 和 NDF 共识光谱指纹的木质纤维素化学解释

Fig. 4 Lignocellulosic chemical interpretation of ADF and NDF consensus spectral fingerprints

如图 5a 和图 5b(图中括号内数字表示支持算法数,字母 C 表示目标共识区段)所示,ADF 各区段

在 Raw、SNV 和 SGD1 组中的支持频率范围分别为 0.19 ~ 0.48、0.20 ~ 0.50、0.13 ~ 0.44,每个预处理

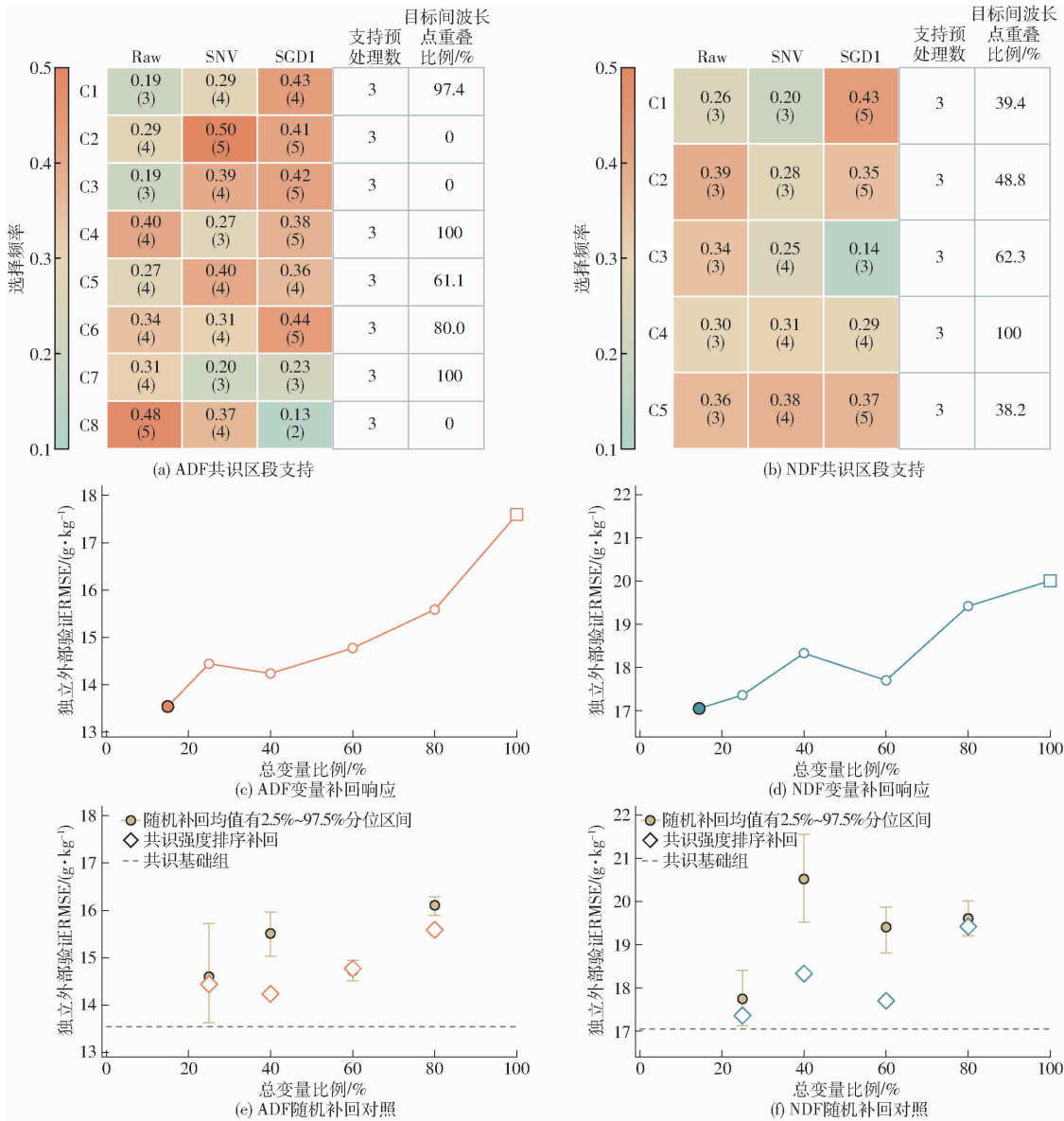


图 5 不同预处理组对共识区段的支持及共识变量补回的外部验证响应

Fig. 5 Preprocessing-group support for consensus wavebands and external-validation responses to consensus-variable addback

组内有 2 ~ 5 种算法达到支持标准;NDF 相应范围分别为 0. 26 ~ 0. 39、0. 20 ~ 0. 38、0. 14 ~ 0. 43,每个预处理组内有 3 ~ 5 种算法达到支持标准。尽管不同预处理组对同一区段的支持强度存在差异,但未出现仅由单一预处理组支撑的共识区段,说明主要波段并非某一种光谱表达方式下的偶然结果。

ADF 的 8 个共识区段中有 5 个与 NDF 共识区段重叠,波长点重叠比例为 61.11% ~ 100%;NDF 的 5 个共识区段均与 ADF 共识区段存在重叠,波长点重叠比例为 38.19% ~ 100%。这一结果与上述共同波段承载细胞壁基础信息、目标特异波段补充组分边界差异的解释一致。各区段在 3 类预处理下均获得支持,同时仍保留目标间重叠程度的差异,表明跨流程共识既提取了较稳定的共享光谱信息,也保留了 ADF 和 NDF 的目标特异响应。

不同波段集合的建模结果进一步说明,共识波段具有较好的预测信息承载能力。由表 2 可知,ADF 目标共识波段的内部测试 RMSE 为 (17.26 ± 1.19) g/kg,低于共同波段的 (19.30 ± 1.27) g/kg 和 ADF 特异波段的 (19.91 ± 1.37) g/kg,说明 ADF 预测需要共享细胞壁基础信息和 ADF 目标特异信息共同支撑。NDF 共同波段单独建模时的 RMSE 为 (25.39 ± 1.74) g/kg,与 NDF 目标共识波段模型的 (25.45 ± 1.69) g/kg 接近;NDF 特异波段单独建模 RMSE 为 (26.88 ± 1.74) g/kg,说明 NDF 预测的主体信息主要来自共享细胞壁基础波段,NDF 特异波段更多承担增量补充作用。结合表 2 中独立外部验证结果可知,固定目标共识波段模型在未参与建模和波段筛选的外部样品上仍保持较好的预测能力,表明这些共识区间并非内部划分下的偶然选择,而

具有一定样品迁移基础。

共识阈值稳健性分析表明,共识强度阈值在 80%、85% 和 90% 分位间变化时,ADF 在 1 424. 5、1 710. 6、1 768. 0、1 926. 8、2 051. 7、2 262. 4 nm 附近的核心峰位保持稳定,NDF 在 1 429. 0、1 926. 8、2 188. 2、2 314. 8 nm 附近的核心峰位也未发生明显迁移(图 3f)。阈值变化主要影响共识波段边界和弱响应区段数量,而不改变主要峰位分布。结果表明,多流程共识方法识别得到的核心波段具有较好的阈值稳健性,可作为后续低成本近红外检测通道筛选的候选谱区。

从低成本少波段近红外设备选型角度看,ADF/NDF 含量快速检测的波段筛选不应只依据某一模型或某一次划分得到的最低误差波点,而应优先选择在多建模流程下反复出现的连续共识区间。对于需要同时兼顾 ADF 和 NDF 含量检测的设备,可优先考虑共同波段中的强共识峰位,如 1 425、1 927、2 188、2 262、2 315 nm 附近波段;若设备主要用于 ADF 含量检测,可在共同通道基础上增加 1 711、1 768、2 052、2 240、2 478 nm 附近波段;若设备主要用于 NDF 含量检测,则可重点补充 1 450 nm 及 2 008~2 024 nm、2 290~2 397 nm 区间。上述波长不应被直接理解为仪器选型的中心波长,而应作为硬件设计的候选谱区;具体中心波长、半峰宽和通道数量仍需结合探测器响应范围及实际样机定标结果进一步优化。

2.4 共识变量补回的性能响应

ADF 共识基础组包含 449 个变量,占全谱变量的 14. 96%,其独立外部验证 RMSE、 R^2 和 RPD 分别为 13. 54 g/kg、0. 892 和 3. 05。变量总数依次增加至 750、1 200、1 801、2 401、3 001 个时,RMSE 分别为 14. 44、14. 24、14. 77、15. 59、17. 59 g/kg,所有确定性补回阶段均未优于共识基础组。补回至全谱后,RMSE 较共识基础组增加 29. 9%, R^2 下降至 0. 817,RPD 下降至 2. 35,表明逐步恢复低共识强度变量未形成持续的性能增益。

NDF 表现出相似的总体规律。包含 434 个变量的共识基础组占全谱变量的 14. 46%,其独立外部验证 RMSE、 R^2 和 RPD 分别为 17. 05 g/kg、0. 890 和 3. 03。变量总数增加至 750、1 200、1 801、2 401、3 001 个后,RMSE 依次为 17. 36、18. 33、17. 70、19. 42、20. 01 g/kg。补回至全谱后,RMSE 增加 17. 3%, R^2 下降至 0. 848,RPD 下降至 2. 58。ADF 和

NDF 均未表现出随补回比例增加而持续改善的性能响应,表明被筛除变量未提供可稳定转化为外部预测收益的系统性协同信息。

等规模随机补回结果表明,变量组成而非变量数量本身影响补回后的模型性能。ADF 含量检测在 25%、40%、60% 和 80% 阶段随机补回结果优于相应强度排序结果的比例分别为 49%、0、72% 和 0;NDF 相应比例分别为 11%、0、0 和 22%,两目标各阶段随机补回 RMSE 的均值及 2. 5% 分位数均高于共识基础组。综上所述,共识筛选在大幅压缩变量的同时保留了主要的联合预测信息。

3 结论

(1)本研究面向秸秆 NDF 和 ADF 含量少波段近红外快速检测需求,构建了基于多建模流程共识的特征波段识别方法。通过组合 Raw、SNV 和 SGD1 共 3 种光谱表达方式与 CARS、MC-UVE、SPA、siPLS 和 VIP 共 5 种变量选择算法,在 100 次 Monte Carlo 样本划分下重复建立 PLS 模型,统计不同流程中波长变量的重复选择特征。结果表明,单一预处理或单一变量选择算法得到的特征波段具有一定流程依赖性,多流程重复出现的共识波段更适合表征 NDF 和 ADF 含量检测中的稳定光谱信息。

(2)ADF 和 NDF 目标共识波段分别包含 449 和 434 个变量,占原始扫描变量的 14. 96% 和 14. 46%。二者共同波段包含 225 个变量,占原始扫描变量的 7. 50%。阈值稳健性分析显示,ADF 在 1 424. 5、1 710. 6、1 768. 0、1 926. 8、2 051. 7、2 262. 4 nm 附近,NDF 在 1 429. 0、1 926. 8、2 188. 2、2 314. 8 nm 附近均形成稳定核心峰位。基于共识波段建立的模型在变量数量大幅压缩后仍保持较好预测能力,ADF 模型在独立外部验证集上的 RMSE、 R^2 和 RPD 分别为 13. 54 g/kg、0. 892 和 3. 05,NDF 对应指标分别为 17. 05 g/kg、0. 890 和 3. 03,表明所识别波段具有较好的预测信息承载能力和外部样品迁移能力。在同一独立外部验证集上,共识模型的 RMSE 均低于全谱 Raw-PLS 及所比较的单一波长选择方法,说明多流程重复支持有助于降低单一变量选择流程的结果依赖性。共同波段主要反映多糖羟基、结合水和 C—H/C—O 合频等细胞壁基础信息,ADF 特异波段主要补充纤维素和木质素富集区响应,NDF 特异波段主要补充半纤维素、结合水状态和总细胞壁基质差异信息。

参 考 文 献

[1] McRoberts K C, Cherney D J R. Low-infrastructure filter bag technique for neutral detergent fiber analysis of forages[J].

- Animal Feed Science and Technology, 2014, 187: 77 – 85.
- [2] Saha U, Vann R A, Reberg-Horton S C, et al. Near-infrared spectroscopic models for analysis of winter pea (*Pisum sativum* L.) quality constituents[J]. Journal of the Science of Food and Agriculture, 2018, 98(11): 4253 – 4267.
- [3] Arroyo J M, González J. Effects of microbial colonization in the rumen on concentration and degradability estimates of fibre fractions[J]. Livestock Science, 2013, 153(1 – 3): 101 – 107.
- [4] Ferreira G, Teets C L, Huffard J B, et al. Effects of planting population, genotype, and nitrogen fertilization on dry matter yield, nutrient composition, in vitro ruminal neutral detergent fiber disappearance, and nitrogen and phosphorus removal of corn for silage[J]. Animal Feed Science and Technology, 2020, 268: 114615.
- [5] Płazek A, Dubert F, Kopec P, et al. In vitro-propagated *Miscanthus × giganteus* plants can be a source of diversity in terms of their chemical composition[J]. Biomass and Bioenergy, 2015, 75: 142 – 149.
- [6] Zeng L, Chen C. Using remote sensing to estimate forage biomass and nutrient contents at different growth stages[J]. Biomass and Bioenergy, 2018, 115: 74 – 81.
- [7] Silva R S T D, Fernandes A M, Gomes R D S, et al. On the specificity of different methods for neutral detergent fiber and related problems[J]. Animal Feed Science and Technology, 2018, 240: 128 – 144.
- [8] Andueza D, Picard F, Martin-Rosset W, et al. Near-infrared spectroscopy calibrations performed on oven-dried green forages for the prediction of chemical composition and nutritive value of preserved forage for ruminants[J]. Applied Spectroscopy, 2016, 70(8): 1321 – 1327.
- [9] Liu J, Jin S, Bao C, et al. Rapid determination of lignocellulose in corn stover based on near-infrared reflectance spectroscopy and chemometrics methods[J]. Bioresource Technology, 2021, 321: 124449.
- [10] 冀荣华, 李常昊, 郑立华, 等. 基于稀疏自注意力和可见-近红外光谱的土壤氮含量预测[J]. 农业机械学报, 2024, 55(10): 392 – 398, 409.
- Ji Ronghua, Li Changhao, Zheng Lihua, et al. Prediction of soil nitrogen content based on sparse self-attention and visible – near-infrared spectroscopy[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(10): 392 – 398, 409. (in Chinese)
- [11] 李硕, 王雅晋, 覃卫林, 等. 基于近地传感光谱协同的土壤重金属含量空间分布预测方法[J]. 农业机械学报, 2025, 56(3): 451 – 457.
- Li Shuo, Wang Yajin, Qin Weilin, et al. Prediction method for soil heavy metal content based on covariates over proximally sensed spectra[J]. Transactions of the Chinese Society for Agricultural Machinery, 2025, 56(3): 451 – 457. (in Chinese)
- [12] Pasquini C. Near infrared spectroscopy: a mature analytical technique with new perspectives: a review[J]. Analytica Chimica Acta, 2018, 1026: 8 – 36.
- [13] Czarnecki M A, Morisawa Y, Futami Y, et al. Advances in molecular structure and interaction studies using near-infrared spectroscopy[J]. Chemical Reviews, 2015, 115(18): 9707 – 9744.
- [14] Jeong S Y, Lee E J, Ban S E, et al. Near infrared spectroscopy model for analyzing chemical composition of biomass subjected to Fenton oxidation and hydrothermal treatment[J]. Renewable Energy, 2021, 172: 1341 – 1350.
- [15] Shukla S R, Shashikala S, Sujatha M. Non-destructive estimation of fibre morphological parameters and chemical constituents of *Tectona grandis* L. f. wood by near infrared spectroscopy[J]. Journal of Near Infrared Spectroscopy, 2021, 29(3): 168 – 178.
- [16] Sluiter A, Wolfrum E. Near infrared calibration models for pretreated corn stover slurry solids, isolated and in situ[J]. Journal of Near Infrared Spectroscopy, 2013, 21(4): 249 – 257.
- [17] Sem V. Interpretability of selected variables and performance comparison of variable selection methods in a polyethylene and polypropylene NIR classification task[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2021, 258: 119850.
- [18] Yun Y H, Li H D, Deng B C, et al. An overview of variable selection methods in multivariate analysis of near-infrared spectra [J]. TrAC Trends in Analytical Chemistry, 2019, 113: 102 – 115.
- [19] Bampi M, Scheer A de P, de Castilhos F. Application of near infrared spectroscopy to predict the average droplet size and water content in biodiesel emulsions[J]. Fuel, 2013, 113: 546 – 552.
- [20] Li L, Peng Y, Li Y, et al. A new scattering correction method of different spectroscopic analysis for assessing complex mixtures[J]. Analytica Chimica Acta, 2019, 1087: 20 – 28.
- [21] Scheibelhofer O, Wahl P R, Larchevêque B, et al. Spatially resolved spectral powder analysis: experiments and modeling[J]. Applied Spectroscopy, 2018, 72(4): 521 – 534.
- [22] Huang X, Luo Y P, Xia L. An efficient wavelength selection method based on the maximal information coefficient for multivariate spectral calibration[J]. Chemometrics and Intelligent Laboratory Systems, 2019, 194: 103872.
- [23] Wang L L, Lin Y W, Wang X F, et al. A selective review and comparison for interval variable selection in spectroscopic modeling[J]. Chemometrics and Intelligent Laboratory Systems, 2018, 172: 229 – 240.
- [24] Yang W, Xiong Y, Wang H, et al. Interval interaction moving window partial least squares for wavelength interval selection in near infrared spectroscopy[J]. Chemometrics and Intelligent Laboratory Systems, 2023, 241: 104976.

- [25] Raffrenato E, Lombard R, Erasmus L J, et al. Prediction of indigestible NDF in South African and Australian forages from cell wall characteristics[J]. *Animal Feed Science and Technology*, 2018, 246: 104 – 113.
- [26] Wang X, Han L, Dai J, et al. Comparative study of laboratory and industrialized near-infrared spectroscopy models for rapid prediction of proximate compositions in crop straw[J]. *Microchemical Journal*, 2025, 216: 114646.
- [27] Chen M J, Yin H L, Liu Y, et al. Non-destructive prediction of the hotness of fresh pepper with a single scan using portable near infrared spectroscopy and a variable selection strategy[J]. *Analytical Methods*, 2022, 14(2): 114 – 124.
- [28] Jin P, Fu Y, Niu R, et al. Non-destructive detection of the freshness of air-modified mutton based on near-infrared spectroscopy[J]. *Foods*, 2023, 12(14): 2756.
- [29] Li Y, Via B K, Han F, et al. Comparison of various chemometric methods on visible and near-infrared spectral analysis for wood density prediction among different tree species and geographical origins[J]. *Frontiers in Plant Science*, 2023, 14: 1121287.
- [30] Shi H, Yu P. Comparison of grating-based near-infrared (NIR) and Fourier transform mid-infrared (ATR – FT/MIR) spectroscopy based on spectral preprocessing and wavelength selection for the determination of crude protein and moisture content in wheat[J]. *Food Control*, 2017, 82: 57 – 65.
- [31] Eastridge M L, Starkey R A, Gott P N, et al. Dairy cows fed equivalent concentrations of forage neutral detergent fiber from corn silage, alfalfa hay, wheat straw, and corn stover had similar milk yield and total tract digestibility[J]. *Animal Feed Science and Technology*, 2017, 225: 81 – 86.
- [32] Van Parijs F, Van Waes C, Vandecasteele B, et al. The optimal lignin quantification method to breed for an improved cell wall digestibility in perennial ryegrass[J]. *Grass and Forage Science*, 2018, 73(1): 101 – 111.
- [33] ISO 16472:2006 Animal feeding stuffs—determination of amylase-treated neutral detergent fibre content (aNDF)[S].
- [34] ISO 13906:2008 animal feeding stuffs—determination of acid detergent fibre (ADF) and acid detergent lignin (ADL) contents [S].
- [35] Barnes R J, Dhanoa M S, Lister S J. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra[J]. *Applied Spectroscopy*, 1989, 43(5): 772 – 777.
- [36] Rinnan Å, Van den Berg F, Engelsen S B. Review of the most common pre-processing techniques for near-infrared spectra [J]. *TrAC Trends in Analytical Chemistry*, 2009, 28(10): 1201 – 1222.
- [37] Delwiche S R, Reeves J B III. A graphical method to evaluate spectral preprocessing in multivariate regression calibrations; example with Savitzky-Golay filters and partial least squares regression[J]. *Applied Spectroscopy*, 2010, 64(1): 73 – 82.
- [38] Savitzky A, Golay M J E. Smoothing and differentiation of data by simplified least squares procedures [J]. *Analytical Chemistry*, 1964, 36(8): 1627 – 1639.
- [39] Leardi R, Boggia R, Terrile M. Genetic algorithms as a strategy for feature selection[J]. *Journal of Chemometrics*, 1992, 6(5): 267 – 281.
- [40] Breiman L. Random forests[J]. *Machine Learning*, 2001, 45(1): 5 – 32.