杨天卓,何晋,吴恋恋,等. 基于高光谱成像的烟丝掺配比例检测研究[J]. 轻工学报,2025,40(3):115-126.
YANG T Z,HE J,WU L L,et al. Detection of tobacco blend ratio based on hyperspectral imaging[J]. Journal of Light Industry,2025,40(3):115-126. DOI:10.12187/2025.03.013

# 基于高光谱成像的烟丝掺配比例检测研究

杨天卓[1],何晋[2],吴恋恋[3],戴永生[4],易斌[4],李华杰[5],张二强[6],堵劲松[1]

1 中国烟草总公司郑州烟草研究院,河南 郑州 450001;
2. 上海烟草集团有限责任公司 技术中心,上海 200082;
3. 甘肃烟草工业有限责任公司 技术中心,甘肃 兰州 730050;
4. 云南中烟工业有限责任公司 楚雄卷烟厂,云南 楚雄 675099;
5. 福建中烟工业有限责任公司 技术中心,福建 厦门 361021;
6. 陕西中烟工业有限责任公司 技术中心,陕西 宝鸡 721013

**摘要**:针对目前卷烟生产线缺乏快速检测烟丝掺配比例方法的问题,利用高光谱成像技术和机器学习方法,采集不同掺配比例的混合烟丝光谱数据,探讨单一及组合预处理方法对模型构建效果的影响。采用偏最小二乘回归(PLSR)和支持向量机回归(SVR)建立回归模型,并利用最小角回归(LARS)、连续投影算法(SPA)、竞争性自适应重采样(CARS)及遗传算法(GA)进行特征波段选择,建立简化模型。结果表明:不同预处理方法的单一或组合使用均会影响回归模型的精度,其中小波变换与SG滤波联用(Wave+SG)方法相比原始数据将平均绝对百分比误差(*MAPE*)降低了1.2%;基于Wave+SG-GA-PLSR建立的回归模型表现最佳,其在两组分掺配中的训练集和测试集*MAPE*分别为1.415%和1.531%;基于两组分掺配建立的方法同样适用于多组分掺配,在三组分和四组分的预测集中,*MAPE*均低于8.361 5%。高光谱成像技术结合机器学习方法能够较为准确地预测混合烟丝中各组分的掺配比例,可为烟丝掺配均匀性的在线监测及卷烟生产过程的质量控制提供参考。

**关键词**:高光谱成像技术;掺配比例;波段选择;机器学习;回归分析
**中图分类号**:TS45 **文献标识码**:A **文章编号**:2096-1553(2025)03-0115-12

## 0 引言

在卷烟产品生产过程中,卷烟配方通常由一种或多种类型的烟丝组成。配方组分的均匀性直接影响产品的物理特性和感官品质的稳定性,对提高原料利用率、提升理化特性和感官品质等方面具有重要作用[1-3]。因此,快速监测卷烟生产线上不同烟丝物料的掺配比例,对于保障各组分实际比例与目标值一致,保证烟丝掺配均匀性,实现工艺过程均质化,以及确保产品质量稳定性等方面具有重要

**收稿日期**:2024-10-12;**修回日期**:2024-11-11;**出版日期**:2025-06-15
**基金项目**:中国烟草总公司重点研发项目(中烟办[2022]111号);云南大观实验室课题项目(YNDG202301QT10);云南中烟科技项目(2492030301010085-3)
**作者简介**:杨天卓(1999—),男,湖南省长沙市人,中国烟草总公司郑州烟草研究院硕士研究生,主要研究方向为烟草工艺。E-mail:526369602@qq.com
**通信作者**:张二强(1989—),男,河南省郑州市人,陕西中烟工业有限责任公司工程师,主要研究方向为烟草工艺。E-mail:erqiang_zhang@163.com

意义[4]。

目前,关于混合烟丝均匀性评价及各组分掺配比例的研究已有较多报道。YC/T 426—2012[5]中以化学成分特性值(即糖碱比与钾含量的乘积)评估烟草组分混合的均匀性;温若愚等[6]研究了混丝模式对烟丝掺配效果的影响;刘栋等[7]通过示踪物表征掺配均匀性;寇霄腾等[8]基于 RGB 图像处理技术实现了梗丝掺配比例的定量测定;李斌等[9]通过拟合模型预测了烟丝中梗丝含量。然而,上述研究方法主要依赖抽样和离线分析,效率较低,难以实现实时质量监控。因此,开发一种实时、无损且适应复杂掺配比例的检测方法至关重要。

高光谱成像技术结合了成像和光谱信息,能够无损获取目标物的光谱与空间特征,具有高精度和高分辨率[10],已广泛应用于食品、农业、医学等多个领域[11-13]。张卫正等[14]提出了一种基于高光谱成像的茎节识别与定位方法;郭文孟等[15]利用高光谱成像技术评估了鲜烟叶成熟度、烘烤过程和烤烟等级;朱亚昆等[16]利用高光谱成像技术结合 PPF 投影算法建立了卷烟牌号识别模型;陶发展等[17]利用高光谱成像技术结合机器学习方法实现了对烟丝中掺杂梗签的快速识别。梅吉帆等[18]利用高光谱成像技术,通过采集单组份烟丝的像素点光谱数据,完成了叶丝与梗丝、薄片丝的判别。这些研究都说明了高光谱成像技术应用于分类识别具有可行性。

但上述方法主要适用于研究类别间光谱特征差异显著、物理分布相对分离的物质体系。而对于混合烟丝掺配,尤其是多组分掺配的情况,会由于不同种类烟丝的光谱存在较多重叠峰,且混合体系中单个像素点的光谱信息较复杂,导致对光谱分辨率的要求大大增加,难以实现精确的组分识别和定量检测。

因此,本文拟基于高光谱成像技术,采集不同掺配比例混合烟丝的光谱数据,并采用不同预处理方法进行处理,利用预处理后数据建立偏最小二乘回归(Partial Least Squares Regression, PLSR)和支持向量回归(Support Vector Regression, SVR)两种回归模型,根据性能指标确定最佳预处理方法和模型;通过最小角回归(Least Angle Regression, LARS)[19]、连续投影算法(Successive Projections Algorithm, SPA)[20]、竞争性自适应重采样(Competitive Adaptive Reweighting Sampling, CARS)[21]及遗传算法(Genetic Algorithm, GA)[22]提取特征波段以建立简化模型,并将该建模方法应用于两组分到四组分掺配中,以期建立一种实时、无损、准确检测混合烟丝掺配比例的方法,为烟丝掺配均匀性的在线监测及卷烟生产过程中的质量控制提供参考。

## 1　材料与方法

### 1.1　实验材料

实验样品,包含云南中烟、福建中烟、陕西中烟、上海烟草集团不同牌号的烟丝(见表 1),用于烟丝掺配比例检测实验与建模。表中梯度数表示为各掺配组分设置的不同比例组合数量,梯度设计基于目前大部分产品的实际配方比例,以确保实验与现实生产相吻合。图像数表示所有梯度下采集用于建模的高光谱图像总数。

### 1.2　主要仪器与设备

LD-sw6401715300UC-CL-SG 型高光谱成像仪、800W 卤素灯,西安立鼎光电科技有限公司;烟丝环形输送装置,郑州嘉德机电科技有限公司。

高光谱图像采集系统与烟丝环形输送装置实物图与示意图如图 1 所示。该系统采集波长范围为 960~1700 nm。

表 1　不同牌号烟丝信息

Table 1　Information on different brands of tobacco shreds

| 来源 | 牌号 | 掺配组分及比例 | 梯度数/个 | 图像数/个 |
|---|---|---|---|---|
| 云南 | YX | 梗丝(4%~24%)+叶丝(含薄片) | 12 | 643 |
| 福建 | QPL | 梗丝(3%~26%)+膨胀丝(3%~22%)+叶丝(含薄片) | 13 | 521 |
| 陕西 | HY | 梗丝(4%~21%)+膨胀丝(4%~11%)+薄片丝(2%~9%)+叶丝 | 9 | 160 |
| 上海 | ZH-1 | 梗丝(3%~21%)+膨胀丝(3%~17%)+叶丝(含薄片) | 9 | 394 |
| 上海 | ZH-2 | 梗丝(7%~14%)+膨胀丝(6%~13%)+薄片丝(6%~13%)+叶丝 | 8 | 318 |

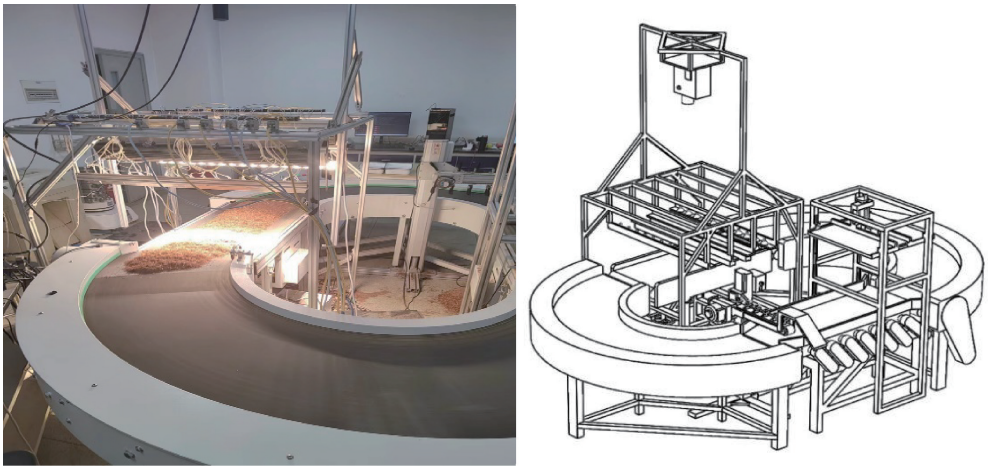


图 1　高光谱图像采集系统与烟丝环形输送装置实物图与示意图

Fig. 1　Physical and schematic diagrams of hyperspectral imaging acquisition system and tobacco circular conveying device

实验参数设置为:曝光时间 3.118 s,帧频 320 f/s,烟丝环形输送装置的黑色传送带速度 0.13 m/s,需与相机设定的参数相匹配,传送皮带宽度 55 cm,烟丝平铺厚度约 2~3 cm。

## 1.3　实验方法

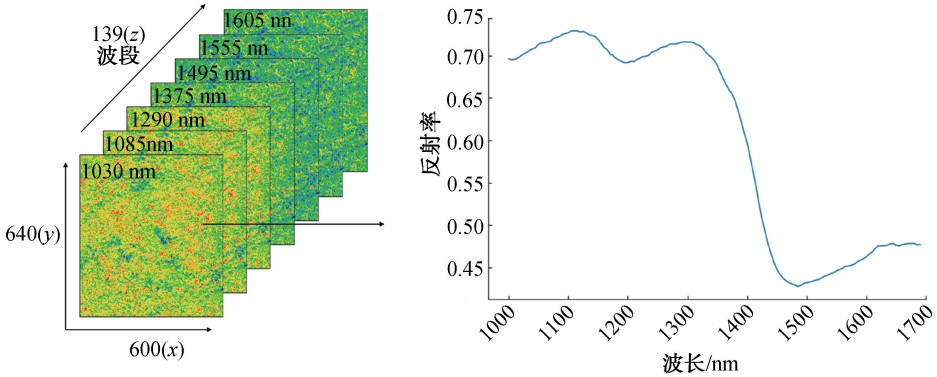
**1.3.1　高光谱图像采集方法**　按照实验设定的不同掺配比例,精确称取不同种类的烟丝,在黑色传送皮带上混合均匀,平整放置,启动传送皮带和相机进行高光谱数据采集,皮带输送过程中,对混合烟丝进行人工翻动混合。由于高光谱成像仪在波长范围两端的响应强度较低,导致光谱数据在 1000 nm 前和 1690 nm 后的数据信噪比较低,文献[17]使用短波近红外光谱范围对不同烟丝进行分类判别,说明该范围对烟丝等物料进行分析具有可行性,因此选择 1000~1690 nm 的数据进行后续分析,共计 139 个波段。

**1.3.2　校正方法**　为提高数据质量,在采集前进行黑白板校正来降低光线及暗电流的干扰。数据校正公式[23]如下所示:

$$R = \frac{I - B}{W - B}$$

式中,$I$ 是原始样本数据,$B$ 是反射率为 0 的黑板反射光谱,$W$ 是反射率为 99%的白板反射光谱,$R$ 是校正后数据。

**1.3.3　感兴趣区域提取**　不同波段高光谱数据如图 2 所示,以其平均光谱作为光谱数据集。为了消除底板对烟丝光谱数据的影响,采用阈值分割方法提取烟丝部分光谱数据。




a)不同波段高光谱图像　　b)对应像素光谱曲线

图 2　不同波段高光谱数据

Fig. 2　Hyperspectral data across different bands

首先将相机采集到的高光谱图像顺序拼接在一起,形成完整的图像数据集,然后用指定窗口大小(28.8 cm×27 cm)来滑动采集光谱数据,滑动窗口区域就是感兴趣区域(Region of Interest, ROI)。由于背景区域反射率低于 0.35,同时为防止出现过曝现象,阈值设置为(0.35, 0.95),通过该阈值排除那些反射率过低或过高的像素点。每个滑动窗口采集经阈值处理后的各像素点的平均光谱,即为样本的原始高光谱反射数据。本研究对每个特定的掺配比例梯度,通过滑动窗口技术采集混合烟丝样品的平均光谱数据,形成该梯度的训练数据集,每个数据点对应的标签为其相应的真实掺配比例(即质量配比),滑动窗口的步长设置为 0.1 倍窗口长度,对应 64 像素,以确保数据的连续性和充足的数据量。

采用同样的方法,对各个掺配比例梯度的混合烟丝样品按滑动窗口方式进行平均光谱数据采集,并打上对应的标签。据此获得不同掺配比例梯度下平均光谱数据的集合,其中每个集合均代表混合烟丝体系在特定掺配比例下的光谱特征。将这些平均光谱数据集合作为训练数据,构建一个多元回归模型,该模型能够预测混合烟丝中不同组分的掺配比例。

光谱数据采集流程如图 3 所示,其中光谱集合中每一条光谱就是一个滑动窗口采集的平均光谱,即某一梯度的原始光谱数据。

**1.3.4　光谱预处理**　混合物料分布密度、表面平整性等样品自身物理状态及仪器响应、散射光、背景等外部因素干扰会使混合烟丝的近红外光谱产生一定的噪声信号进而影响后续回归模型的预测准确率[24]。不同预处理方法在应对不同任务及样品时会有不同效果,本文采用光谱预处理中常用的一阶导数(First Derivative, D1)、二阶导数(Second Derivative, D2)、多元散射校正(Multiplicative Scatter Correction, MSC)、标准正态变换(Standard Normal Variate, SNV)、尺度缩放(Min-Max Scaling, MMS)、小波变换(Wavelet Transform, Wave)、SG 滤波(Savitzky-Golay Filtering, SG)及其组合的方法来对采集的平均光谱数据进行预处理,通过对比不同预处理方法的模型性能,选择最优的数据预处理方法。

**1.3.5　特征波段选择**　高光谱数据的波段众多,这虽然提供了丰富的信息,却也带来了数据的冗余及计算的复杂性。为了提升处理速度,确保高光谱在生产线上的高效应用,并在尽量不牺牲预测精度的同时提高效率和模型的稳定性[25],本研究采用 LARS、SPA、CARS、GA 4 种特征波段选取方法,通过

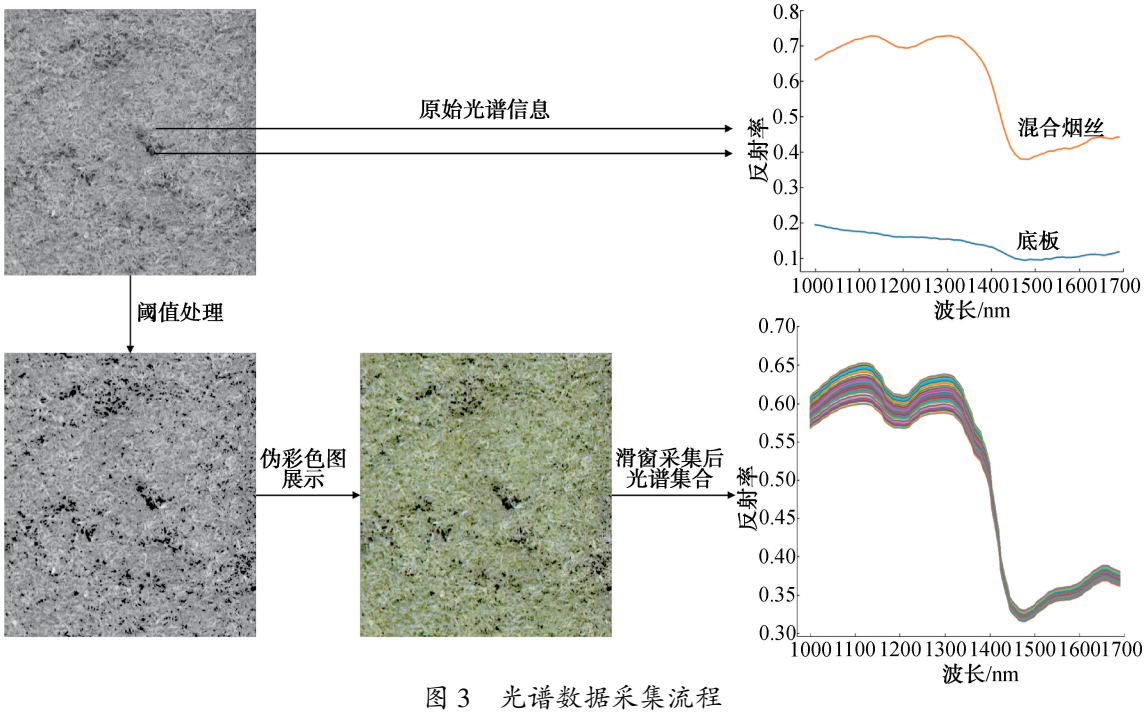



图 3　光谱数据采集流程

Fig. 3　Spectral data acquisition process

比较这些基于所选特征波段建立的模型的性能,确认最佳的特征波段。

**1.3.6 模型建立与评估** PLSR 用于发现变量之间的关联关系,是一种常用的回归模型[26]。与传统的最小二乘回归不同,PLSR 适用于解释变量数量多于观测值的情况,或当解释变量高度相关时。PLSR 通过提取潜在变量或成分来捕捉解释变量与响应变量之间的关系。这些潜在变量是从原始数据中提取出来的,旨在最大化响应变量和解释变量之间的协方差。建模过程中潜在变量数通过五折交叉验证方式确定,如图 4 所示。随着模型潜在变量数的增加,训练和验证的均方根误差(*RMSE*)均降低,代表模型预测性能提升。选择潜在变量数为 20 时,既保证了模型的精确性,也提升了运算的效率。

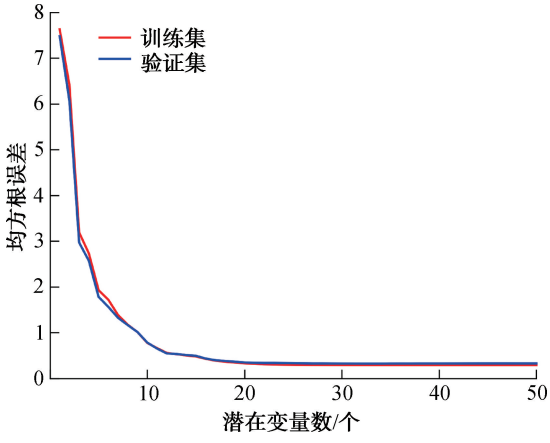



图 4 PLSR 模型交叉验证确定变量数

Fig. 4 Cross-validation of PLSR model to determine the number of variables

SVR 是建立在支持向量机(Support Vector Machine, SVM)理论基础上的回归分析方法[27]。它通过寻找一个最佳超平面来最小化实际输出与模型预测值之间的偏差,同时也控制模型复杂度以防止过拟合。SVR 在处理非线性问题时表现出色,由于其在超平面构建中引入了核函数,能够有效处理那些在原始空间中不易分析的复杂数据模式。此外,SVR 通过引入松弛变量允许一定程度的误差,使得模型能够更好地适应数据的噪声和异常值。这些特性使得 SVR 成为一种适用于广泛场景的稳健回归工具,能够在保持模型复杂度和计算效率平衡的同时,为精确的数据预测提供支持。

模型采用决定系数($R^2$),*RMSE* 和平均绝对百分比误差(Mean Absolute Percentage Error, *MAPE*)进行评价,其中 $R^2$ 越接近 1,*RMSE* 和 *MAPE* 越小,则代表回归模型预测精度越高。相应公式如下所示:

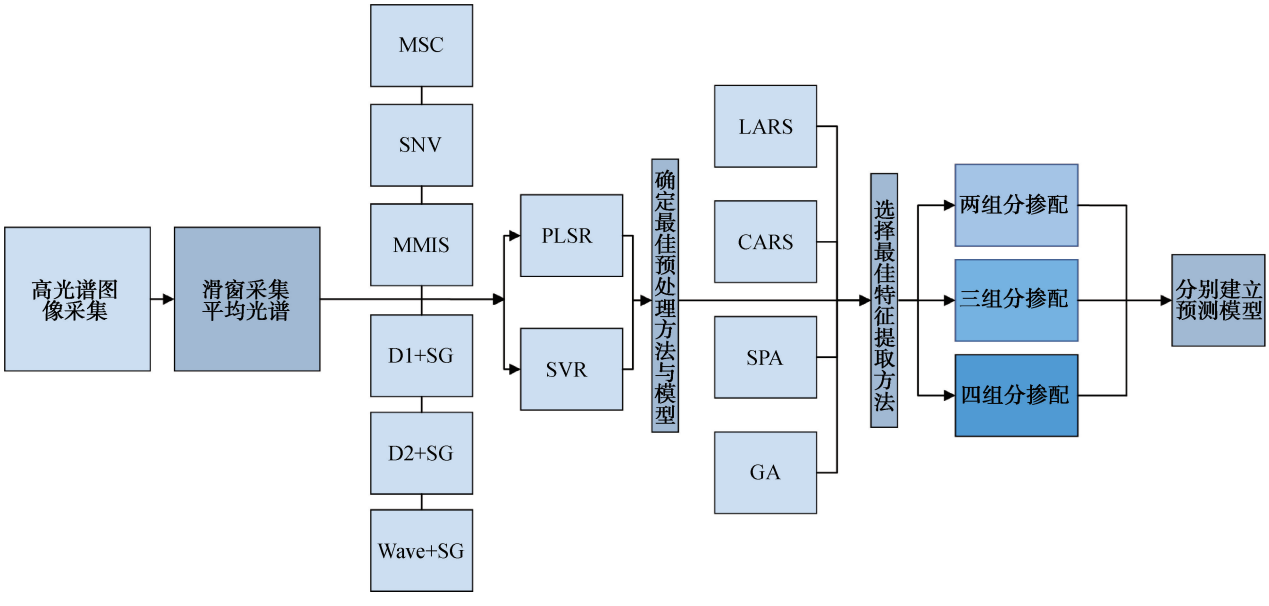



图 5 烟丝掺配比例检测技术路线图

Fig. 5 Technical route diagram for tobacco blend ratio detection

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}{n}}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \widehat{Y}_i}{Y_i}\right| \times 100\%$$

式中，$n$ 为平均光谱样本数；$Y_i$ 为当前梯度下该组分实际掺配比例；$\widehat{Y}_i$ 为组分掺配比例的预测值；$\bar{Y}$ 为实际掺配比例的平均值。其中 $R^2$ 用于评估数据集的拟合程度，为便于计算各组分掺配比例×100。

烟丝掺配比例检测技术路线图如图 5 所示。

### 1.4　数据处理分析

本研究中所有高光谱数据的采集、校正、光谱数据提取、数据预处理、特征波段选择、回归模型建立和评估及图片绘制均通过 Python3.9 完成。

## 2　结果与分析

### 2.1　烟丝掺配比例检测模型研究

以两组分烟丝掺配为例进行研究。

**2.1.1　光谱数据预处理结果**　混合烟丝光谱数据预处理后结果如图 6 所示。由图 6a）可知，经 MSC 处理后的光谱较为平滑，校正了样品散射效应，减少了多重散射带来的干扰；由图 6b）可知，经 SNV 处理后表现出均一化趋势，降低了因样品厚度或表面不均匀引起的全局偏差，使不同样品基线更加一致；由图 6c）可知，经 MMS 处理后，光谱数据在 0~1 范围内得到进一步标准化，提高了不同样品之间的对比性和一致性；由图 6d）和 e）可知，一阶和二阶导数强化了光谱中的细节特征，出现更多尖峰，但也增加了一定噪声，而 SG 滤波能够平滑处理，在保留关键特征的同时减少噪声；由图 6f）可知，小波变换后的光谱显示出不同层次的细节，而 SG 滤波进一步平滑了光谱，保留了全局形态并适度增强了细节。

**2.1.2　基于预处理后光谱的回归建模**　将采集到的不同梯度下混合烟丝的光谱数据，以每个滑动窗口采集到的一条平均光谱作为一个样本点，按照肯纳德斯算法（Kennard-Stone，K-S）将建模集按 7∶3 的比例划分为 862 个训练集和 370 个测试集。

使用基于原始光谱及预处理得到的相关光谱建立 PLSR 和 SVR 回归模型，以模型的训练集和测试集的 *RMSE* 和 *MAPE* 来评价模型的效能。原始光谱及预处理后光谱的 PLSR 和 SVR 模型结果见表 2。由表 2 可知，基于原始光谱建立的 PLSR 和 SVR

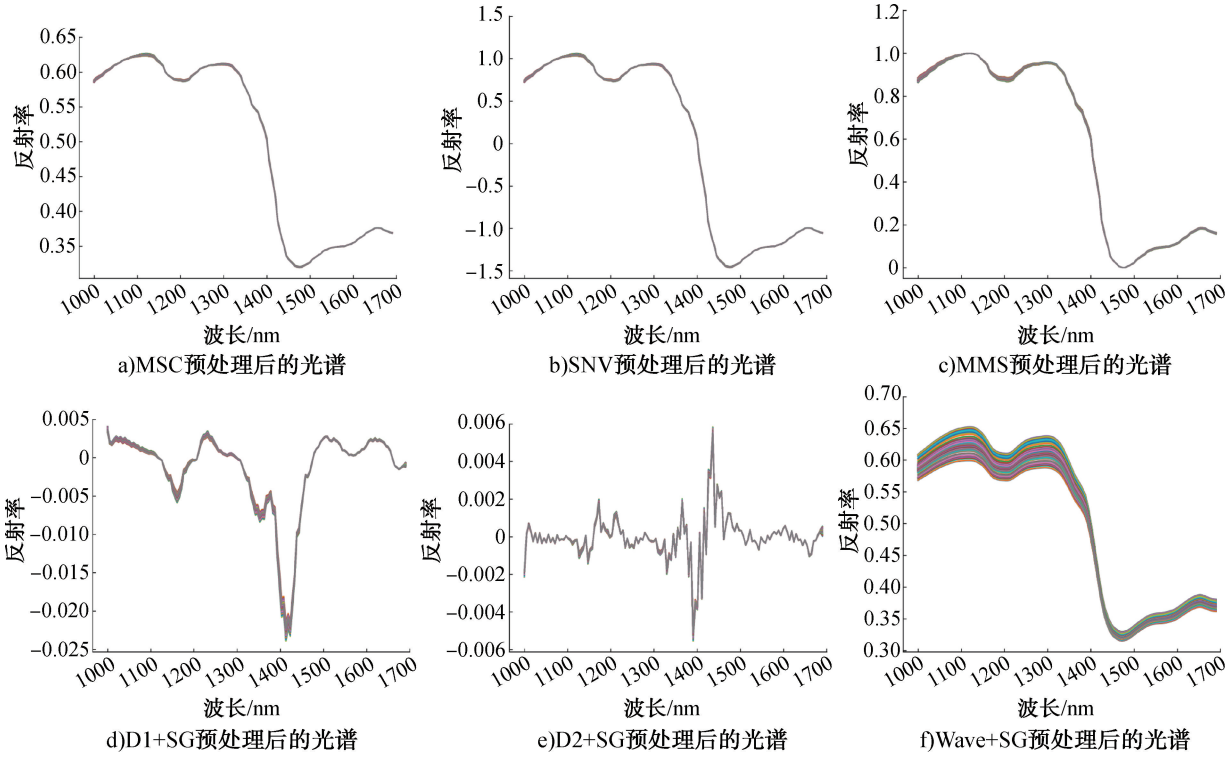



图 6　混合烟丝光谱数据预处理结果

Fig. 6　Spectral data preprocessing results for blended tobacco

表 2 原始光谱及预处理后光谱的 PLSR 和 SVR 模型结果

Table 2 Results of PLSR and SVR models for both original and preprocessed hyperspectral data

| 模型 | 预处理方法 | 训练集 | | | 测试集 | | |
|---|---|---|---|---|---|---|---|
| | | $R^2$ | *RMSE* | *MAPE*/% | $R^2$ | *RMSE* | *MAPE*/% |
| PLSR | Original | 0.990 | 0.589 | 2.067 | 0.990 | 0.631 | 2.493 |
| | MSC | 0.989 | 0.615 | 2.169 | 0.939 | 1.573 | 7.339 |
| | SNV | 0.989 | 0.623 | 2.215 | 0.981 | 0.641 | 2.401 |
| | MMS | 0.990 | 0.631 | 2.419 | 0.988 | 0.655 | 2.439 |
| | D1+SG | 0.995 | 0.423 | 1.549 | 0.994 | 0.452 | 1.567 |
| | D2+SG | 0.996 | 0.377 | 1.437 | 0.996 | 0.394 | 1.611 |
| | Wave+SG | 0.997 | 0.307 | 1.083 | 0.997 | 0.324 | 1.247 |
| SVR | Original | 0.958 | 1.298 | 5.417 | 0.945 | 1.398 | 5.631 |
| | MSC | 0.948 | 1.449 | 5.819 | 0.931 | 1.568 | 6.259 |
| | SNV | 0.993 | 0.500 | 1.777 | 0.993 | 0.522 | 2.048 |
| | MMS | 0.981 | 0.813 | 3.049 | 0.981 | 0.876 | 3.749 |
| | D1+SG | 0.714 | 3.198 | 14.361 | 0.686 | 3.568 | 17.931 |
| | D2+SG | 0.705 | 3.251 | 15.060 | 0.677 | 3.617 | 18.740 |
| | Wave+SG | 0.956 | 1.322 | 5.315 | 0.943 | 1.419 | 5.454 |

模型在测试集上的 *MAPE* 分别为 2.493% 和 5.631%,证明了利用高光谱成像技术检测不同烟丝掺配比例的可行性和有效性。不同预处理方法对模型的精度也有不同影响,在 PLSR 模型中,MSC 方法会导致建立的模型性能下降较多,可能是因为对光谱的多元散射效应有过度校正,而经 D1+SG、D2+SG、Wave+SG 这 3 种方法预处理后,建立的模型性能得到提升,其中 Wave+SG 的提升效果最明显;在 SVR 模型中,MSC、D1+SG、D2+SG 这 3 种方法会使模型性能明显下降。在模型的对比上,PLSR 回归模型优于 SVR 回归模型,其主要体现在模型更加简便高效,训练集和测试集的 $R^2$、*RMSE*、*MAPE* 更优。

通过比对以上不同预处理方法和回归模型,基于小波变换与 SG 滤波联(Wave+SG)方法结合 PLSR 回归模型的性能最优,其测试集的 *RMSE* 和 *MAPE* 分别达到 0.324 和 1.247%。基于此,下文将用经 Wave+SG 预处理后的光谱进行特征波段选择,并建立 PLSR 回归模型。

**2.1.3 特征波段确定结果** 选定波段的 LARS 回归系数如图 7 所示,其中在 1000~1690 nm 下,使用 LARS 进行特征波段选择,特征点数目选择为 40,LARS 系数设置为 0.000 1。由图 7 可知,每个条形的高度对应了 LARS 算法所选择特征波段对应的回归系数的大小,系数越大,则相应特征波段在预测时越重要。

CARS 波段的选择过程如图 8 所示。由图 8a)可知,所选波段数量随采样次数增加而降低;由图 8b)可知,在前 30 次迭代中 10 折交叉验证均方根误差(Root Mean Square Error of Cross-Validation, *RMSECV*)维持在一个较低的稳定水平,而在迭代末期出现明显上升,表明模型在经过一系列特征剔除后出现了性能退化。由图 8c)可知,不同波长回归系数路径的波长被保留时间越长,则认为对预测目标变量的影响力最大。随着迭代进行,不同波长的回归系数趋于零或稳定,表明被模型排除或确认为关键特征。上述过程消除了与混合烟丝各组分含量无关的信息,选择了最具相关信息的 45 个特征波段,结果如图 9 所示。

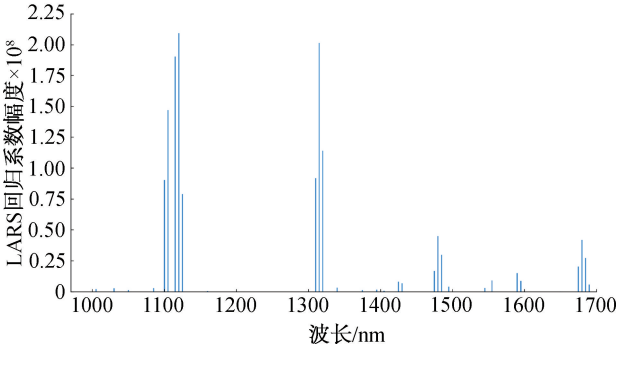



图 7 选定波段的 LARS 回归系数

Fig. 7 LARS regression coefficients for selected bands

SPA 算法设模型波长最小变量为 5，最大变量为 40，以 *RMSE* 最小时确定最终特征波长数量，SPA 特征波段提取结果如图 10 所示。由图 10a）可知，在保留 33 个变量时的 *RMSE* 达到一个较低的稳定值；数据集的平均光谱和所选的特征波段见图 10b）。

在使用 GA 算法时，其相关参数设置为：初始种群大小为样品点数目，即光谱条数，变异概率 10%，交叉概率 50%，迭代次数 20 次。GA 特征波段提取结果如图 11 所示。由图 11a）可知，随着迭代的进行，最大适应度逐渐提高，表明算法逐渐找到更优的特征组合以最大化目标函数；而最小适应度的波动较大，显示了种群中某些个体适应度的不稳定性；平均适应度则相对稳定地提升，反映了整体种群适应度的平均改善趋势。由图 11b）可知，GA 算法选择的特征波段主要集中在 1010～1140 nm、1190～1370 nm、1530～1690 nm 这 3 个区间，共筛选出 43 个特征波段，占总特征波段的 30.9%。

**2.1.4　基于特征波段的回归建模**　基于特征波段的 PLSR 模型结果见表 3。由表 3 可知，经过 LARS、CARS、SPA 和 GA 筛选特征波段后，模型在测试集上的 *RMSE* 与 *MAPE* 均存在小幅度上升，代表模型性能下降，这可能是因为在特征波段提取过程中删

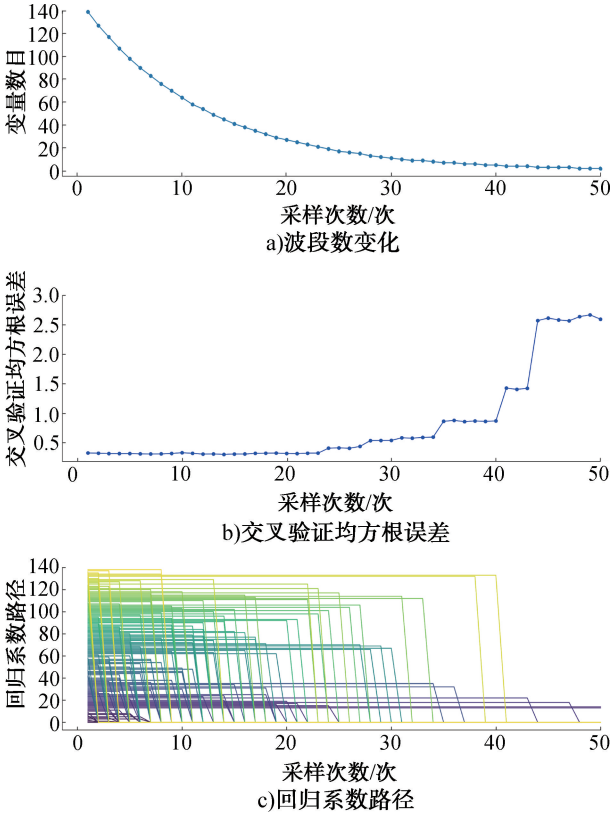



图 8　CARS 波段选择过程

Fig. 8　CARS bands selection process

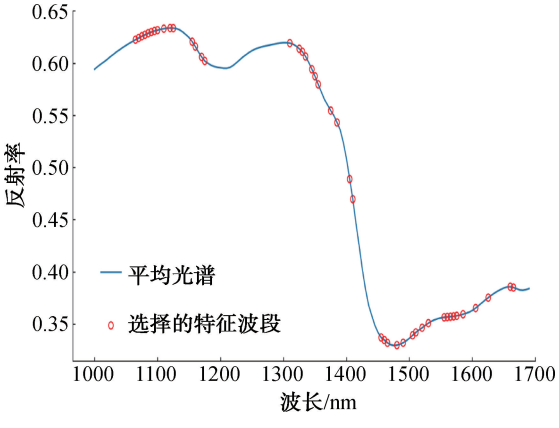



图 9　CARS 提取的特征波段

Fig. 9　Feature bands extracted by CARS

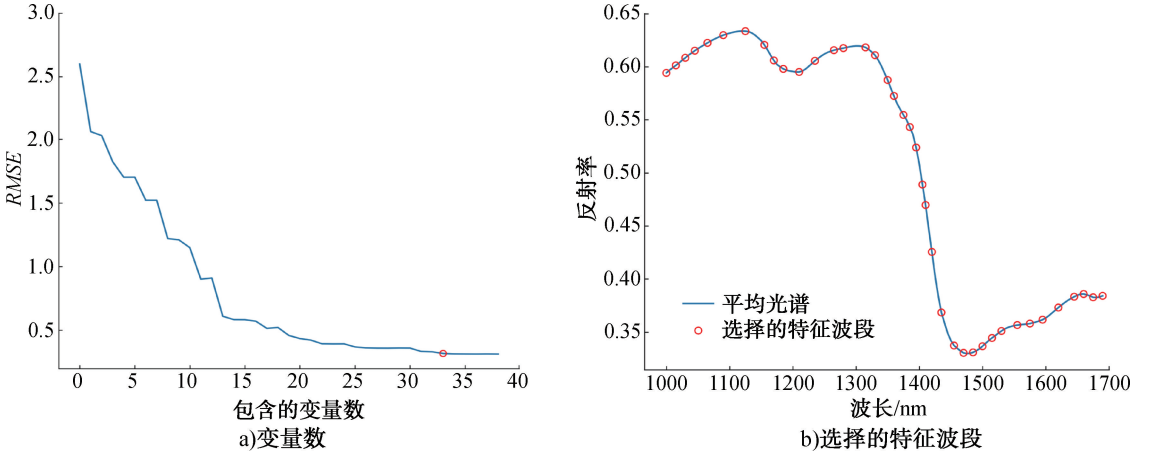



图 10　SPA 特征波段提取结果

Fig. 10　Feature bands extracted by SPA

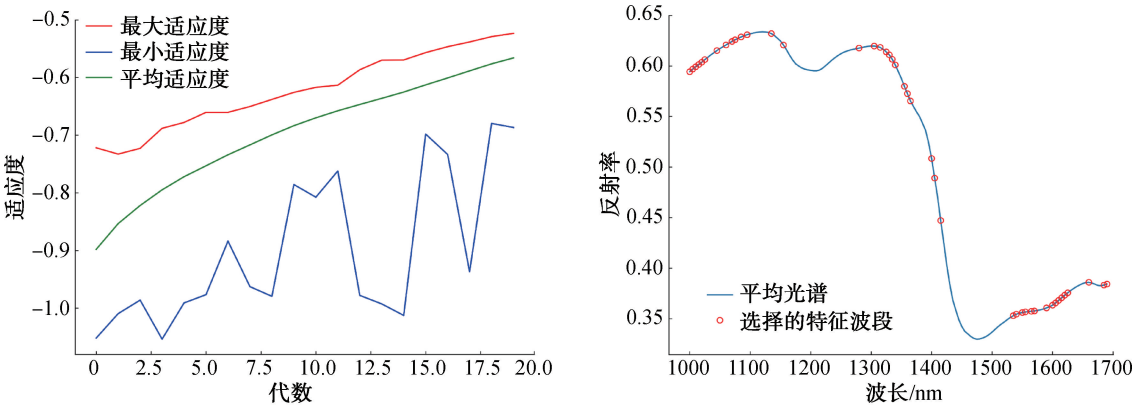



图 11 GA 特征波段提取结果

Fig. 11 Feature bands extracted by GA

除了一定量的相关波段,导致信息含量有一定程度的下降。在这 4 种方法中,经 GA 算法特征提取后的模型性能指标最佳,说明经 GA 筛选出的特征波段最具代表性,含有最多的信息并已删除冗余的波段数据,且其造成的性能略微下降在可接受范围内。因此基于 Wave+SG 预处理后的光谱数据再经 GA 特征波段选择建立的 PLSR 模型为检测两组分掺配比例的最佳方法。

## 2.2 两组分掺配结果分析

为说明模型的泛化能力,将上述所建立的 Wave-SG-GA-PLSR 方法应用于新的预测数据集(为重复实验得到,未参与模型训练),预测集的梯度范围需控制在训练集的梯度范围内,将模型预测值的平均值与真实值进行比对,结果见表 4。由表 4 可知,3 个梯度的 *MAPE* 分别达到一个较低的水平,平均 *MAPE* 为 5.083 2%,说明预测精度较高。其中梗丝的 *MAPE* 在 3 个梯度中分别为 11.870 2%、12.762 6%、2.467 3%。故基于 Wave-SG-GA 建立的 PLSR 回归模型在未知数据上具有较好的泛化能力,在梗丝含量为 4%~24%时,能够较为准确地预测混合烟丝中各组分的掺配比例。

## 2.3 三组分和四组分掺配结果分析

将 Wave-SG-GA-PLSR 方法应用于三组分掺配和四组分掺配,此处由于原始样品和掺配组分变化,不同牌号的烟丝样品需要重新训练对应的模型,不同组分的模型并不通用,但建模方法通用。三组分掺配和四组分掺配预测集结果见表 5 和表 6。由表 5 和表 6 可知,平均 *MAPE* 在 9%以下。

表 3 基于特征波段的 PLSR 模型结果

Table 3 Results of PLSR models based on feature bands

| 方法 | 训练集 | | | 测试集 | | |
|---|---|---|---|---|---|---|
| | $R^2$ | *RMSE* | *MAPE*/% | $R^2$ | *RMSE* | *MAPE*/% |
| LARS | 0.995 | 0.440 | 1.581 | 0.995 | 0.447 | 1.697 |
| CARS | 0.995 | 0.424 | 1.423 | 0.995 | 0.427 | 1.563 |
| SPA | 0.995 | 0.418 | 1.417 | 0.995 | 0.426 | 1.585 |
| GA | 0.996 | 0.394 | 1.415 | 0.996 | 0.392 | 1.531 |

表 4 两组分掺配预测集结果

Table 4 Prediction results for two-component blending

| 牌号 | 真实掺配比例/% | | 平均预测值/% | | *MAPE*/% |
|---|---|---|---|---|---|
| | 梗丝 | 纯丝 | 梗丝 | 纯丝 | |
| YX | 9.090 0 | 90.910 0 | 10.169 0 | 89.831 0 | 6.528 6 |
| | 11.110 0 | 88.890 0 | 12.527 9 | 87.472 1 | 7.178 9 |
| | 20.000 0 | 80.000 0 | 20.464 8 | 79.535 2 | 1.542 1 |

表 5　三组分掺配预测集结果

Table 5　Prediction results for the three-component blending

| 牌号 | 真实掺配比例/% | | | 平均预测值/% | | | *MAPE*/% |
|---|---|---|---|---|---|---|---|
| | 梗丝 | 膨胀丝 | 叶丝 | 梗丝 | 膨胀丝 | 叶丝 | |
| QPL | 6.960 0 | 6.090 0 | 86.960 0 | 7.964 0 | 6.372 8 | 85.662 9 | 7.814 7 |
| | 14.810 0 | 11.110 0 | 74.070 0 | 14.199 2 | 10.841 9 | 74.962 0 | 2.918 9 |
| | 23.730 0 | 19.770 0 | 56.500 0 | 23.011 3 | 19.317 3 | 57.668 2 | 2.941 0 |
| ZH | 9.170 0 | 7.500 0 | 83.330 0 | 8.955 0 | 7.575 7 | 83.480 1 | 2.161 6 |
| | 14.810 0 | 11.110 0 | 74.070 0 | 14.303 9 | 10.783 6 | 74.913 3 | 2.553 3 |
| | 20.250 0 | 16.460 0 | 63.290 0 | 20.176 2 | 16.465 9 | 63.358 9 | 0.710 4 |

表 6　四组分掺配预测集结果

Table 6　Prediction results for four-component blending

| 牌号 | 真实掺配比例/% | | | | 平均预测值/% | | | | *MAPE*/% |
|---|---|---|---|---|---|---|---|---|---|
| | 梗丝 | 膨胀丝 | 薄片丝 | 叶丝 | 梗丝 | 膨胀丝 | 薄片丝 | 叶丝 | |
| HY | 20.408 0 | 6.122 0 | 5.442 0 | 68.027 0 | 20.162 9 | 7.162 0 | 6.150 5 | 66.524 7 | 8.361 5 |
| ZH | 10.000 0 | 10.000 0 | 10.000 0 | 70.000 0 | 10.420 2 | 9.008 2 | 9.004 7 | 71.579 4 | 6.583 4 |

在三组分掺配中，QPL 样品梗丝的 *MAPE* 分别为 14.481 7%、4.185 3%、3.504 3%，膨胀丝的 *MAPE* 分别为 7.318 0%、3.270 7%、2.862 8%。说明在梗丝含量为 4%～26%和膨胀丝含量为 3%～22%时，能够较为准确地预测新样品的混合烟丝掺配比例，而 ZH 样品能取得更高的精度。

在四组分掺配中，HY 样品梗丝的 *MAPE* 为 1.230 5%，膨胀丝的 *MAPE* 为 16.988 7%，薄片丝的 *MAPE* 为 13.018 4%，说明在梗丝含量为 4%～21%、膨胀丝含量为 4%～11%及薄片丝含量为 2%～9%时，该方法也能取得较好的精度。其中 ZH 样品的平均 *MAPE* 较 HY 样品略低。

整体上，三组分掺配的平均 *MAPE* 为 3.183 3%，四组分掺配的平均 *MAPE* 为 7.472 5%，因此该方法对于多组分掺配的情况也能有较高的预测精度。

## 3　结论

本文进行了 4 个牌号不同烟丝的两组分、三组分、四组分的掺配实验，通过单一预处理及组合预处理方法处理原始光谱数据，根据建模结果选择了最佳预处理方法和模型，并比对了 4 种特征波段提取方法，建立了基于 Wave-SG-GA-PLSR 掺配组分预测模型，结果显示：原始光谱数据经 MSC、SNV、MMS、D1+SG、D2+SG、Wave+SG 这 6 种方法的预处理后构建了 PLSR 和 SVR 模型，经模型结果对比分析可得，最佳预处理方法是 Wave+SG，最佳回归模型是 PLSR；对比 LARS、CARS、SPA 及 GA 这 4 种特征波段提取方法，由基于特征波段建立的 PLSR 模型结果可知，GA 提取的 43 个特征波段最具代表性，其训练集和测试集的 *MAPE* 分别为 1.415%和 1.531%；建立了基于 Wave-SG-GA-PLSR 的两组分、三组分、四组分烟丝掺配预测模型，其预测集的平均 *MAPE* 分别为 5.083 2%、3.183 3%、7.472 5%，说明该方法建立的预测模型具有较好的泛化能力，对复杂掺配情况的预测误差较小。

因此，本研究提供了一种对两组分到多组分烟丝掺配比例较为通用的预测方法，通过优化光谱预处理、特征波段选择及建模策略，实现了对烟丝掺配比例的精确预测。该方法有助于提高烟丝配方的完整性和掺配均匀性，为卷烟生产线上烟丝质量监测和均匀性控制提供了参考。未来的研究将致力于探索更多组分的掺配建模，通过引入更先进的特征提取方法和升级硬件条件，进一步提高模型的检测精度和效率。

**参考文献：**

[1]　丁美宙，刘欢，刘强，等. 梗丝形态对细支卷烟加工及综合质量的影响[J]. 食品与机械，2017，33(9)：197-202.

[2]　陈帅伟，胡苏林，崔宁，等. “三丝”掺配比例对卷烟理化指标的影响研究[J]. 西南农业学报，2015，28(6)：

2742-2745.
[3] 徐成龙,宋朝鹏,戴亚,等. 梗丝与膨胀丝及薄片丝对卷烟主流烟气中 3 种有害成分释放量的影响[J]. 湖南农业大学学报(自然科学版),2013,39(1):23-25.
[4] 胡立中,张胜军,余小平,等. 均匀设计-PLS-NIR 法预测卷烟配方烟丝中梗丝及薄片丝含量[J]. 中国烟草学报,2010,16(2):26-30.
[5] 国家烟草专卖局. 烟草混合均匀度的测定:YC/T 426—2012[S].
[6] 温若愚,席年生,张大波,等. 不同混丝模式对烟丝掺配效果的影响[J]. 烟草科技,2008,41(9):13-16.
[7] 刘栋,陈越立,李华杰,等. 取样量、取样次数对烟丝混合均匀性检测的影响[J]. 郑州轻工业学院学报(自然科学版),2013,28(1):45-49.
[8] 寇霄腾,张勇,张卉,等. 基于 RGB 图像特征的卷烟梗丝掺配比例检测[J]. 食品与机械,2021,37(9):78-82.
[9] 李斌,蔡佳校,何邦华,等. 一种基于热分析技术测定烟丝中梗丝含量的方法:CN107271312A[P]. 2017-10-20.
[10] LU Y Z,SAEYS W,KIM M,et al. Hyperspectral imaging technology for quality and safety evaluation of horticultural products:A review and celebration of the past 20-year progress[J]. Postharvest Biology and Technology,2020,170:111318.
[11] ZHANG C,GUO C T,LIU F,et al. Hyperspectral imaging analysis for ripeness evaluation of strawberry with support vector machine[J]. Journal of Food Engineering,2016,179:11-18.
[12] GOMEZ C,GHOLIZADEH A,BORŮVKA L,et al. Using legacy data for correction of soil surface clay content predicted from VNIR/SWIR hyperspectral airborne images[J]. Geoderma,2016,276:84-92.
[13] NEITTAANMÄKI-PERTTU N,GRÖNROOS M,TANI T,et al. Detecting field cancerization using a hyperspectral imaging system[J]. Lasers in Surgery and Medicine,2013,45(7):410-417.
[14] 张卫正,张伟伟,张焕龙,等. 基于高光谱成像技术的甘蔗茎节识别与定位方法研究[J]. 轻工学报,2017,32(5):95-102.
[15] 郭文孟,薛宇毅,罗靖,等. 基于高光谱成像的烟叶泛青特征分析与表征[J]. 烟草科技,2023,56(7):84-91.
[16] 朱亚昆,梅吉帆,郭文孟,等. 基于 PPF 投影算法和高光谱技术的卷烟牌号识别模型[J]. 轻工学报,2024,39(4):118-126.
[17] 陶发展,杨栋,洪伟龄,等. 基于高光谱成像的烟丝中梗签分类识别研究[J]. 河南科技大学学报(自然科学版),2024,45(3):32-42,5.
[18] 梅吉帆,李智慧,李嘉康,等. 基于高光谱成像技术的配方烟丝组分判别[J]. 分析测试学报,2021,40(8):1151-1157.
[19] LI B,WANG Y Q,LI L S,et al. Research on apple origins classification optimization based on least-angle regression in instance selection[J]. Agriculture,2023,13(10):1868.
[20] WANG J G,TIAN T,WANG H J,et al. Improving the estimation accuracy of rapeseed leaf photosynthetic characteristics under salinity stress using continuous wavelet transform and successive projections algorithm[J]. Frontiers in Plant Science,2023,14:1284172.
[21] ZHU L,GU W Z,SONG T Q,et al. Coal seam *in situ* inorganic analysis based on least angle regression and competitive adaptive reweighted sampling algorithm by XRF-visNIR fusion[J]. Scientific Reports,2022,12(1):22365.
[22] 褚小立,袁洪福,王艳斌,等. 遗传算法用于偏最小二乘方法建模中的变量筛选[J]. 分析化学,2001,29(4):437-442.
[23] 令峰,邢富康,韦克苏,等. 基于高光谱成像技术和深度学习的烤后烟叶品种分类判别[J]. 中国烟草科学,2024,45(4):83-92.
[24] MISHRA P,NORDON A,TSCHANNERL J,et al. Near-infrared hyperspectral imaging for non-destructive classification of commercial tea products[J]. Journal of Food Engineering,2018,238:70-77.
[25] 吴龙国,王松磊,何建国. 基于高光谱技术的土壤水分无损检测[J]. 光谱学与光谱分析,2018,38(8):2563-2570.
[26] FU P,MEACHAM-HENSOLD K,GUAN K Y,et al. Estimating photosynthetic traits from reflectance spectra:A synthesis of spectral indices,numerical inversion,and partial least square regression[J]. Plant,Cell & Environment,2020,43(5):1241-1258.
[27] ZHANG P P,SHEN B B,JI H W,et al. Nondestructive prediction of mechanical parameters to apple using hyperspectral imaging by support vector machine[J]. Food Analytical Methods,2022,15(5):1397-1406.

# Detection of tobacco blend ratio based on hyperspectral imaging

YANG Tianzhuo[1],HE Jin[2],WU Lianlian[3],DAI Yongsheng[4],YI Bin[4],LI Huajie[5],
ZHANG Erqiang[6],DU Jinsong[1]

1. *Zhengzhou Tobacco Research Institute of CNTC*,*Zhengzhou* 450001,*China*;
2. *Technology Center*,*Shanghai Cigarette Group Co.*,*Ltd.*,*Shanghai* 200082,*China*;

3. *Technology Center*, *Gansu Tobacco Industrial Co.*, *Ltd.*, *Lanzhou* 730050, *China*;
4. *Chuxiong Cigarette Factory*, *Yunnan Cigarette Industrial Co.*, *Ltd.*, *Chuxiong* 675099, *China*;
5. *Technology Center*, *Fujian Cigarette Industrial Co.*, *Ltd.*, *Xiamen* 361021, *China*;
6. *Technology Center*, *Shaanxi Cigarette Industrial Co.*, *Ltd.*, *Baoji* 721013, *China*

**Abstract**: This study focuses on detecting tobacco blend ratios using hyperspectral imaging. Due to the lack of rapid methods for detecting tobacco blend ratios on cigarette production lines, spectral data were collected from mixed tobacco with different blend ratios using hyperspectral imaging technology and machine learning methods. The effects of single and combined preprocessing techniques on model performance were explored. Regression models were established using partial least squares regression (PLSR) and support vector machine regression (SVR). Feature wavelength selection was performed with least angle regression (LARS), successive projections algorithm (SPA), competitive adaptive reweighted sampling (CARS), and genetic algorithm (GA) to build simplified models. The results showed that preprocessing methods, either individually or combined, affected model accuracy. The combined wavelet transform and SG filtering (Wave+SG) method reduced the mean absolute percentage error (*MAPE*) by 1.2% compared to raw spectral data. The Wave+SG-GA-PLSR model performed best, with MAPE values of 1.415% and 1.531% for the training and test sets in two-component blends, respectively. This method was also applicable to multi-component blends, with *MAPE* values below 8.361 5% for both three-component and four-component blends. Hyperspectral imaging combined with machine learning can accurately predict the proportions of components in mixed tobacco, providing a reference for online monitoring of blend uniformity and quality control in cigarette production.

**Key words**: hyperspectral imaging technology; blend ratio; band selection; machine learning; regression analysis

[责任编辑:王晓波　刘春奎]