近红外技术的广西速生桉抽出物含量测定与模型优化

朱 华[1,2]，吴 斑[2,3]，房桂干[2,3]，梁 龙[2,3]，朱北平[2,3]，佘光辉[1,2*]

1. 南京林业大学林学院，江苏 南京 210037
2. 南京林业大学林业资源高效加工利用协同创新中心，江苏 南京 210037
3. 中国林业科学研究院林产化学工业研究所，江苏 南京 210042

摘 要 为解决速生桉抽出物测定方法繁琐耗时，木浆生产能耗居高不下等问题，以引种的3种广西速生尾巨桉原料(DH32-29，DH32-26，DH33-27)为研究对象，采集了144个样本的近红外光谱，按国标方法测定全部样品的苯醇抽出物和1%NaOH抽出物含量。在Matlab 8.0中采用信号平滑，一阶、二阶导数，矢量归一化，多元散射校正等方法预处理原始光谱，用偏最小二乘法、支持向量机法和人工神经网络法以及常用于宏观经济分析的LASSO法分别结合上述预处理方法建立模型，筛选出最优建模方法。运用遗传算法对波段进行选择，提高了模型的精确度从而优化了模型。确定了建立苯醇抽出物含量模型时，可联用平滑、MSC和一阶导数预处理光谱数据，以1 345.0～1 821.4和2 127.8～2 241.3 nm区间波段参与建模，建模方法为偏最小二乘法，最佳主成分数为9时，模型有最好的精确度。其RMSEP值可达0.25%，绝对偏差范围为－0.39%～0.38%。筛选出的波段包含了如1 410和1 447 nm附近酚羟基伸缩振动的一级倍频，2 133 nm处苯环上碳氢键的伸缩振动与碳碳双键伸缩振动的合频等苯醇抽出物的特征波段。建立1%NaOH抽出物分析模型时，可联用平滑、矢量归一化和一阶导数预处理，选择1 138.2～2 363.0 nm波段数据，建模方法为LASSO，选取的调整参数值 μ 为12.61，此时模型精确度最高。RMSEP值为0.37%，绝对偏差范围为－0.56%～0.53%。筛选出的波段包含了1 158和1 170 nm附近乙酰脂基团 CH_3 中C—H的伸缩振动二级倍频吸收，1 666，1 681和1 790 nm附近 CH_3 中C—H伸缩振动的一级倍频吸收等特征吸收。模型的预测能力从组分结构角度得到了解释。模型的RPD值分别为4.67和5.77，模型性能均可满足实际需求，有望应用于制浆造纸生产线上的速生桉抽出物含量分析。研究结果表明，通过预处理方法选择和建模方法选择，结合遗传算法的应用，可以建立并优化广西速生桉木抽出物含量的近红外测定模型；同时，LASSO算法作为一种新兴算法，在近红外光谱分析中表现出了较好的处理共复线性数据的能力，可以建立准确性较好的分析模型。

关键词 近红外技术；LASSO算法；预处理；遗传算法

中图分类号：O433 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2020)03-0793-06

引 言

近年来宏观政策对进口废纸配额的限制，影响了国内纸浆生产市场，制浆造纸原料短缺的形势越发严峻。与此同时，草浆因污染重，运输成本高等原因产量增速缓慢，本应得到大力发展的木浆，因天然林禁伐，生态林保护等原因，年进口比例超过了50%的警戒线。因此，加大国内速生林发展投入，合理选择培育树种，科学规划提高利用质量已成为增加国产优质木浆产量，填补制浆造纸原料短缺的唯一出路[1]。桉树作为南方人工速生林主要培育树种，有着轮伐期短，年生长量高的优势，在两广、海南、云南等地区广泛种植。桉树采伐后以木片原料的形式来到制浆生产线，而各批次木片因生长来源情况，运输保存时间的不同，化学成分存在着较大的差异，这些差异对桉木浆生产的药品用量，能耗以及成浆质量有很大影响。其中，综纤维素/纤维素决定了木浆的得率，而一些含量较少的抽出物也起着重要的作用，如苯醇抽出物可用于表征制浆过程中的树脂沉积物障碍，且

收稿日期：2019-02-27，**修订日期**：2019-05-15

基金项目：国家重点研发计划项目(2017YFD0601005)，国家自然科学基金重大项目(31890774)资助

作者简介：朱 华，女，1975年生，南京林业大学林学院博士研究生 e-mail：jecy_zhu@163.com

*通讯联系人 e-mail：ghshe@njfu.edu.cn

与纸浆着色返黄相关；1%NaOH 抽出物则与原料腐朽变质程度有关。因此，研究木片原料抽出物含量的快速测定技术，摆脱传统抽出物含量测定的繁琐耗时，根据各批次木片抽出物含量情况实时调整制浆用药水平和电耗，有利于木浆生产提质降耗目标的实现。

近红外光谱作为高效，无损的分析方法，在农业生产[2]，生物化学[3]等方面都得到了广泛的应用。制浆造纸行业中分析检测项目多，测试步骤繁琐，前人尝试运用近红外技术实现原料鉴别分类[4]，卡伯值预测[5]的智能化控制，已取得一定成果。而在制浆原料抽出物分析方面，因取样难度大，抽出物含量较低且成分复杂，光谱信息干扰严重，基于近红外光谱的传统回归算法对其拟合建模效果往往不佳。本研究以广泛种植的速生桉——尾巨桉杂交品系(DH32-29，DH32-26，DH33-27)为研究对象采集近红外光谱，通过多种建模算法对比分析，发现传统的偏最小二乘法适用于建立苯醇抽出物分析模型，而在分析 1%NaOH 抽出物时，采用 LASSO 算法压缩建模回归系数，可以更有效的解决光谱数据共复线性问题，得到更加精确的分析模型。在此基础上结合遗传算法筛选特征波长变量，提高了模型的预测性能和稳健性，实现了广西速生桉中苯醇抽出物、1%NaOH 抽出物含量的快速分析测定。

1 实验部分

1.1 速生桉样品

选择广西速生尾巨桉杂交品系三个主要品种(DH32-29，DH32-26，DH33-27)，由金华林业原料林基地提供，引种自广西东门林场，树龄均为 6 年。每个品种在树干不同部位截取 48 个圆盘为样本，去皮，经切片机切割后置于空气中充分平衡水分，用粉碎机将木片磨成木粉，经振动筛筛分出粒径 40～60 目之间的木粉，同样平衡水分后作为样品，共计 144 个。每个品种以含量梯度法按 3∶1 的比例筛选 12 个样品作为验证集；其余 108 个样品作为训练集，用于速生桉苯醇抽出物、1%NaOH 抽出物模型的建立。

1.2 近红外光谱采集

所使用近红外光谱仪器应能适应制浆造纸生产线复杂环境，成本低廉且结构简单，易于组装搭建，因此选用上海复享光学股份有限公司的全息光栅分光(阵列检测器)近红外光谱仪 NIR2500，采集速生桉样品的近红外漫反射光谱。设定仪器参数如下：光谱范围 900～2 500 nm，信噪比为 6 800∶1，波长点数为 256 个。采集光谱时，取 3～4 g 木粉样品置于底部为石英玻璃的上置式样品杯中，并用 250 g 砝码压住木粉样品以保证样品紧实度的一致性，光源经光纤探头从样品杯底部照射样品。每个样品重复取样 3 次，取其平均光谱作为该样品代表性光谱。

1.3 含量分析

用传统实验室方法(GB/T 2677.6—1994，GB/T 2677.5—1993)测定全部 144 个速生桉样品的苯醇抽出物和 1%NaOH 抽出物含量，每个样品测 3 组平行数据，以平均值作为样品测定值。

1.4 模型建立

选择了常见的光谱预处理方法分别与偏最小二乘法(PLS)，LASSO 法，支持向量机法(SVR)和人工神经网络法(ANN)结合建模，得出两种抽出物含量分析的最优建模方法，并通过遗传算法(genetic algorithm，GA)进行波段选择，进一步优化模型。

其中 LASSO 法在近红外光谱分析中应用较少，该算法主要原理是通过增加约束条件，令回归系数的绝对值之和小于可调整的常数，并最小化残差平方和，将次要项、干扰项的回归系数压缩为零，从而使得所建模型更为准确。

设有 P 个自变量 x_1，x_2，…，x_p 和因变量 y，满足

$$y=\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p+\varepsilon \tag{1}$$

式中 α 是常数项，β_1，β_2，…，β_p 是回归系数，ε 是随机误差项。

设(x_{i1}，x_{i2}，…，x_{ip}；y_i)，$i=1,2,\cdots,n$，是自变量 x_1，x_2，…，x_p 的观测值，对其中心标准化，即：$\sum_{i=1}^{n} y_i=0$，$\sum_{i=1}^{n} x_{ij}=0$，$\sum_{i=1}^{n} x_{ij}^2=1$，$j=1,2,\cdots,p$，记 $\beta=(\beta_1,\beta_2,\cdots,\beta_p)^{\mathrm{T}}$。

对回归系数的绝对值和进行惩罚

$$(\hat{\alpha},\hat{\beta})=\arg\min\sum_{i=1}^{n}(y_i-\alpha-\sum_{j=1}^{p}\beta_j x_{ij})^2,$$

$$\text{subject to}\sum_{j=1}^{p}|\beta_j|\leqslant\lambda \tag{2}$$

式(2)中 $\lambda\geqslant 0$ 是约束常数。数据经过了中心标准化，因此对任意约束常数 $\lambda\geqslant 0$，α 都存在解 $\hat{\alpha}=0$。可将式(2)转化为

$$\hat{\beta}(\text{Lasso})=\arg\min\Big\{\sum_{i=1}^{n}(y_i-\sum_{j=1}^{p}\beta_j x_{ij})^2+\mu\sum_{j=1}^{p}|\beta_j|\Big\} \tag{3}$$

其中 μ 是惩罚系数。随着 μ 的增加，最优解的 $\sum_{j=1}^{p}|\beta_i|$ 项减小，过程中，无用信息变量的权重逐步降低至 0，有用信息变量在建模过程中的重要性凸显，从而提高模型的稳健性和预测性能。式(3)中，每个值对应着唯一 LASSO 解，因此 LASSO 建模的关键在于最优调整参数 μ 的确定[6]。

1.5 模型评价

以 R^2_{cv} 和 RMSECV 辅助建立模型。以 R^2_{val}、RMSEP、相对分析误差(RPD)、绝对偏差范围(AD)评价模型性能，以偏移值(Bias)表征模型系统误差。其中 R^2 仅用于辅助评价，RMSECV、RMSEP 是均方根误差(RMSE)在建模和预测阶段的两种形式，其值越接近 0，模型的分析预测精确度越高。RMSEP 是评价模型的主要指标，而 RPD 值定义为验证集标准偏差与 RMSEP 的比值。当 RPD 值小于 2 时，认为所建模型没有价值；其值在 2～3 之间时，可用于定性评估；当 RPD 值高于 3 时，模型准确性有所提高，适用于质量控制。绝对偏差是近红外预测值和实验测定值之差，决定了预测误差的上下限。Bias 值反映了分析过程中可能存在的系统误差。针对制浆造纸原料成分分析，通常认为当 Bias 的绝对值小于 0.01%时，系统偏移对分析结果的影响可以忽略，模型不存

在系统误差。

$$\text{Bais} = \frac{\sum_{i=1}^{n}(y_{i,\text{measured}} - y_{i,\text{predicted}})}{n}$$

2　结果与讨论

2.1　速生桉抽出物测定值分布

三种速生桉共计 144 个样本的抽出物含量如表 1 所示。其中 DH32-26 样本的苯醇抽出物含量较高，分布在 1.07%～5.01%之间，DH32-29 和 DH33-27 的苯醇抽出物含量分布没有显著差异，均分布在 0.85%～3.96%之间。DH32-26 和 DH33-27 的 1%NaOH 抽出物含量较低，分布在 11.65%～16.92%之间，DH32-29 的 1%NaOH 抽出物含量分布较宽，约 11.61%～20.16%。速生桉样本的两种抽出物含量分布较宽，据此可建立有代表性的分析模型。

表 1　样品含量分布情况

Table 1　Content distribution of samples

材种	苯醇抽出物/%				1%NaOH 抽出物/%			
	含量范围	平均值	中位数	标准差	含量范围	平均值	中位数	标准差
DH32-26	1.07%～5.01%	3.17	3.06	1.14	11.65%～16.88%	14.26	14.55	1.46
DH32-29	0.87%～3.95%	2.37	2.28	0.87	11.61%～20.16%	16.31	16.57	2.72
DH33-27	0.85%～3.96%	2.32	2.39	0.90	11.70%～16.92%	14.40	14.58	1.41

2.2　光谱预处理及算法选择

控制实验室温度(20±0.5)℃，采集全部速生桉样品的近红外光谱，如图 1 所示，横坐标表示波长，纵坐标表示漫反射吸光度。速生桉样品成分结构复杂，除了含量较少的苯醇抽出物(包括树脂、蜡、脂肪等)，1%NaOH 抽出物(无机盐、糖、植物碱等)之外，还含有作为主成分存在的纤维素、半纤维素等多糖类及芳香族高聚物。为降低无关信息影响，需对速生桉原始光谱进行预处理。

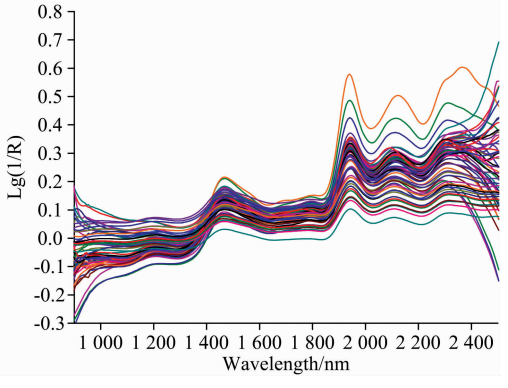

图 1　速生桉样品近红外光谱

Fig. 1　The near infrared spectra of Fast-growing Eucalyptus

常见预处理方法针对性较强，如信号平滑可用于降低噪声，导数则用于消除基线和背景干扰，矢量归一化和多元散射校正用于消除光程变化和非特异性散射的影响[7]。在 Matlab8.0 中将预处理方法组合联用处理原始光谱数据，以期获得更好的效果，并结合 PLS，LASSO，SVR 和 ANN 法对训练集样本进行五折交叉验证建模分析，确定分别适用于分析苯醇抽出物和 1%NaOH 抽出物的预处理和建模方法，16 个模型的性能见表 2。可见当采用平滑、MSC、一阶导数的方法预处理原始光谱[图 2(a)]，运用偏最小二乘法建立苯醇抽出物含量分析模型时，R^2_{cv} 值最大，为 0.968 9；RMSECV 值最小，为 0.22%，模型最优。建模过程中选择的最佳主成分数为 9。当采用平滑、矢量归一化、一阶导数处理原始光谱数据[图 2(b)]，用 LASSO 法建立 1%NaOH 抽出物含量分析模型时，R^2_{cv} 值最大，为 0.978 4；RMSECV 值最小，为 0.37%，模型建立过程中选取的调整参数值为 12.61。

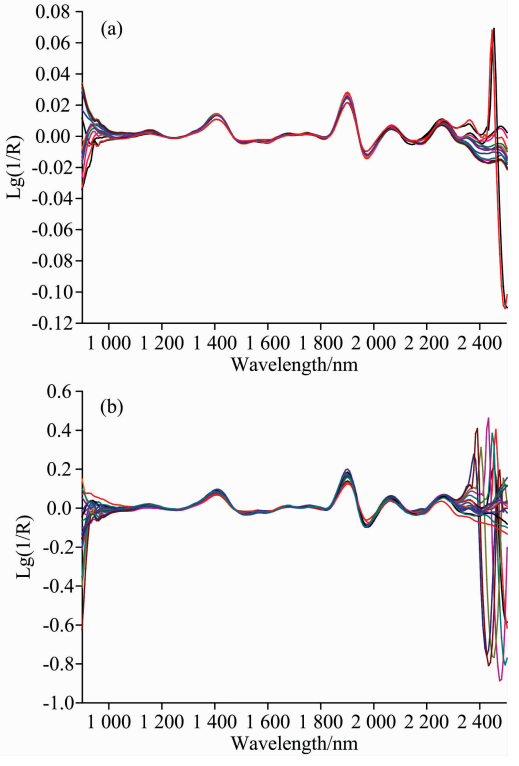

图 2　速生桉原始光谱的预处理

(a)：苯醇抽出物；(b)：1%NaOH 抽出物

Fig. 2　Pretreatment of original spectra of Fast-growing Eucalyptus

(a)：Benzene-alcohol extractives；(b)：1% NaOH extractives

表 2 不同抽出物不同建模方法的模型评价

Table 2 Evaluation of models for different extractives using different modeling methods

算法	预处理方法	苯醇抽出物		1%NaOH 抽出物	
		R^2_{cv}	RMSECV /%	R^2_{cv}	RMSECV /%
偏最小二乘法 PLS	平滑+归一化+一阶导数	0.962 1	0.25	0.965 0	0.42
	平滑+归一化+二阶导数	0.965 8	0.23	0.964 1	0.42
	平滑+MSC+一阶导数	0.968 9	0.22	0.966 6	0.40
	平滑+MSC+二阶导数	0.965 0	0.23	0.9679	0.39
LASSO 法 LASSO	平滑+归一化+一阶导数	0.950 7	0.31	0.973 4	0.37
	平滑+归一化+二阶导数	0.953 3	0.31	0.967 6	0.39
	平滑+MSC+一阶导数	0.957 0	0.29	0.969 4	0.39
	平滑+MSC+二阶导数	0.949 5	0.32	0.967 3	0.40
支持向量机法 SVR	平滑+归一化+一阶导数	0.944 3	0.34	0.960 1	0.47
	平滑+归一化+二阶导数	0.941 7	0.35	0.957 4	0.49
	平滑+MSC+一阶导数	0.942 1	0.35	0.963 0	0.43
	平滑+MSC+二阶导数	0.948 2	0.32	0.966 2	0.40
人工神经网络法 ANN	平滑+归一化+一阶导数	0.946 5	0.33	0.959 4	0.47
	平滑+归一化+二阶导数	0.949 1	0.32	0.963 9	0.43
	平滑+MSC+一阶导数	0.949 0	0.32	0.964 6	0.42
	平滑+MSC+二阶导数	0.950 2	0.31	0.963 2	0.43

2.3 波段选择与模型的建立

仪器的近红外光谱区域为 900～2 500 nm，该区间包含一些无关或干扰信息，可能影响模型的精确度和稳健性。因此采用遗传算法(genetic algorithm，GA)结合相应的建模方法分别筛选出光谱中与苯醇抽出物和 1%NaOH 抽出物有较强相关性的波长变量，以期建立更加稳健的分析模型。对 256 个波长点分别进行二进制编码，以“1”代表该波长点被选中，以“0”代表该波长被剔除，设定筛选波长数目范围为 5～200，种群规模 100，进化代数 500，交叉概率 0.8，变异概率 0.1。

对于苯醇抽出物，将每代筛选出的波长变量子集进行偏最小二乘建模，以 RMSECV 作为适应度评价指标。最终筛选出的最优波段为 1 345.0～1 821.4 和 2 127.8～2 241.3 nm。苯醇抽出物中含有树脂、蜡、脂肪、固醇等，同时还有少量单宁和色素，该波段存在苯醇抽出物的特征吸收：如 1 360 nm 附近 C—H 伸缩振动二级倍频和 C—H 变形振动的合频，其中 C—H 来自抽出物中有机物所含的 CH_3；1 410和 1 447 nm 附近酚羟基伸缩振动的一级倍频，酚羟基来源于单宁和色素；1 668 nm 处苯环上 C—H 伸缩振动的一级倍频，苯环上的 C—H 同样来源于单宁和色素。1 695 和1 721 nm 处存在 CH_3 中 C—H 的伸缩振动一级倍频；1 820 nm 附近存在 O—H 伸缩振动和 C—O 伸缩振动二级倍频的合频，这类羟基和碳氧基源于抽出物中所含的固醇；而 2 133 nm 附近苯环上存在 C—H 的伸缩振动和 C═C 伸缩振动的合频，这两种官能团可能来自于色素[8-9]。

对于 1%NaOH 抽出物，将每代筛选出的波长变量子集进行 LASSO 法建模，以 RMSECV 作为适应度评价指标。最终筛选出的最优波段为 1 138.2～2 363.0 nm。1%NaOH 抽出物中除了包含部分无机盐类、环多醇以及多糖类如胶、植物粘液、淀粉等，还含有一部分溶出的木质素、聚戊糖、树脂酸及糖醛酸，成分复杂。筛选出的波段包含了 1 158 和 1 170 nm 附近聚戊糖乙酰脂基团 CH_3 中 C—H 的伸缩振动二级倍频吸收；1 666 nm(聚戊糖)、1 681 nm(聚戊糖)、1 790 nm(木质素)附近 CH_3 中 C—H 伸缩振动的一级倍频吸收；2 133 nm 附近苯环 C—H 伸缩振动和 C═C 伸缩振动合频的吸收，两种官能团均源于木质素；2 329 nm 处 C—H 伸缩振动和 C—H 变形振动的合频吸收，C—H 来自于聚戊糖；2 360 nm 附近则存在 O—H 变形振动和 C—H 伸缩振动合频的吸收，同样源于纤维素[10-11]。用所得模型分析验证集样品光谱，见表 3，可见通过遗传算法筛选出的波段所建立的抽出物分析模型比全波段建立的抽出物模型性能更好，

表 3 不同波段所建抽出物模型评价

Table 3 Evaluation of models for extractives developed from different bands

类型	波段/nm	R^2_{val}	RMSEP /%	RPD	AD /%
苯醇抽出物	1 345.0～1 821.4 2 127.8～2 241.3	0.954 0	0.25	4.67	−0.39～0.38
	900～2 500	0.950 3	0.28	4.49	−0.43～0.41
1%NaOH 抽出物	1 138.2～2 363.0	0.969 9	0.37	5.77	−0.56～0.53
	900～2 500	0.966 1	0.40	5.43	−0.55～0.57

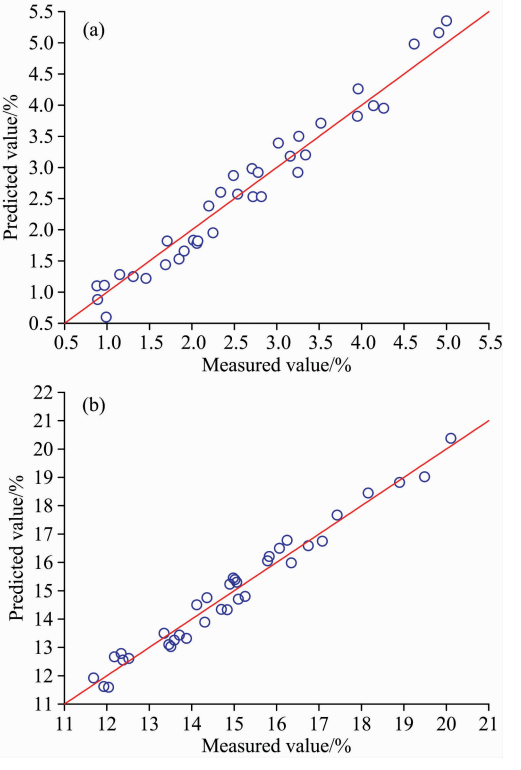

图 3 测定值-预测值散点分布

(a)：苯醇抽出物；(b)：1%NaOH 抽出物

Fig. 3 The distribution of scattered points of measured value and predicted value

(a)：Benzene-alcohol extractives；(b)：1% NaOH extractives

有着更低的 RMSEP 值和更小的绝对偏差范围。一定程度上可以减少分析时间，同时提升了模型精确度。两种抽出物模型 RPD 值分别为 4.67 和 5.77，均高于 3，模型可用于质量控制。根据独立验证情况作测定值-预测值散点图(图 3)，所得点在斜率为 45°的直线两侧分布较为均匀，Bias 值分别为 −0.003 89%和−0.005 28%，均小于 0.01%，可以认为模型不存在系统误差。

3　结　论

利用近红外技术结合多种预处理方法和化学计量学算法，建立了广西速生桉木的抽出物分析模型并通过遗传算法对模型进行了优化。确定了建立苯醇抽出物分析模型的方法为：平滑、MSC、一阶导数预处理，1 345.0～1 821.4 和 2 127.8～2 241.3 nm 近红外波段参与建模，建模方法为偏最小二乘法。建立 1%NaOH 抽出物分析模型的方法为：平滑、矢量归一化、一阶导数预处理，1 138.2～2 363.0 nm 区间波段参与建模，LASSO 法建立模型。针对广西速生桉抽出物进行近红外分析，模型性能较好，能够满足原料抽出物快速分析测定的要求；同时创新性的在近红外光谱建模分析中引入 LASSO 算法，为有效解决光谱数据的共复线性问题提供了更多的选择。

References

[1] GENG Li-min, SHEN Wen-xing(耿利敏，沈文星). Forestry Economics(林业经济), 2018, 40(12): 14.
[2] Folch-Fortuny A, Prats-Montalbán J M, Cubero S, et al. Chemometrics & Intelligent Laboratory Systems, 2016, 156: 241.
[3] Morozova M, Elizarova T, Pleteneva T. European Scientific Journal, 2013, 9(27): 8.
[4] Nascimbem L B, Rubini B R, Poppi R J. Journal of Wood Chemistry and Technology, 2013, 33(4): 247.
[5] Xu F, Yu J, Tesso T, et al. Applied Energy, 2013, 104(2): 801.
[6] Tibshirani R. Journal of the Royal Statistical Society, 2011, 73(3): 267.
[7] Tsuchikawa S, Torii M, Tsutsumi S. Journal of Near Infrared Spectroscopy, 2017, 6(1): 47.
[8] Shin Y, Na B, Lim W, et al. Education Economics, 2014, 1(2): 137.
[9] Axrup L, Markides K, Nilsson T. Journal of Chemometrics, 2015, 14(5-6): 561.
[10] Pfautsch S, Macfarlane C, Ebdon N, et al. Trees: Structure and Function, 2012, 26(3): 963.
[11] Hein P R G, Campos A C M, Mendes R F, et al. European Journal of Wood & Wood Products, 2011, 69(3): 431.

Analysis of Extractives Content of Guangxi Fast-Growing Eucalyptus and Models Optimization Based on Near-Infrared Technique

ZHU Hua[1,2], WU Ting[2,3], FANG Gui-gan[2,3], LIANG Long[2,3], ZHU Bei-ping[2,3], SHE Guang-hui[1,2*]

1. College of Forestry, Nanjing Forestry University, Nanjing 210037, China
2. Collaborative Innovation Center for High Efficient Processing and Utilization of Forestry Resources, Nanjing Forestry University, Nanjing 210037, China
3. Institute of Chemical Industry of Forest Products, Chinese Academy of Forestry, Nanjing 210042, China

Abstract In pulping and papermaking industry, extractives of wood chips influence the impregnation efficiency, pulp energy consumption and pulp yield. But traditional analysis methods for the content of extractives are not applicable for industrial online monitoring because of being time consuming and costly. Therefore, the present study used near infrared (NIR) spectroscopy to predict rapidly extractives content of three species of fast-growing *Eucalyptus urophylla* × *E. grandi* chips (DH32-29, DH32-26, DH33-27) grown in China's Guangxi Province. NIR spectra of 144 fast-growing Eucalyptus were collected using a holographic grating spectrometer equipped with a halogen illumination and array detector. The benzene-alcohol extractives and 1% NaOH extractives content of 144 samples were gravimetrically determined according to the Chinese national standard test method respectively. The near-infrared spectrum were pretreated using smoothing, first derivative, second derivative, vector normalization and multivariate scattering correction in Matlab 8.0, and the models were developed for various pretreatment methods by loading PLS, LASSO, SVR and ANN algorithm. The optimal modeling methods were selected. Genetic algorithm was used to select the bands, which improved the accuracy of the models and optimized the models. In conclusion, in order to develop analysis model of benzene-alcohol extractives, smoothing, MSC and first derivative methods should be used to preprocess the original spectrum, the bands of 1 345.0～1 821.4 and 2 127.8～2 241.3 nm were selected, meanwhile, the partial least squares algorithm was used with the optimal factor 9. The model had the best accuracy for the RMSEP value as low as 0.25%, and the ab-

solute deviation range was −0.39%～0.38%. The optimal bands between 1 345.0～1 821.4 and 2 127.8～2 241.3 nm have been associated with O—H stretching (1st overtone) of phenolic compound (1 410 and 1 447 nm), as well as C—H stretching and C═C stretching group frequencies of benzene ring (2 133 nm) and other characteristic absorption. In order to establish the content analysis model of 1% NaOH, smoothing, vector normalization, first derivative should be used to pretreat the original data, the bands between 1 138.2～2 363.0 nm were picked and LASSO was adopted. The model had the best accuracy when the μ value was 12.61, the independent verification show the RMSEP value was 0.37%, and the absolute deviation range was −0.56%～0.53%. The optimal bands between 1 138.2～2 363.0 nm have been associated with C—H stretching (2nd overtone) of —C═OCH$_3$ (1 158 and 1 170 nm), as well as C—H stretching (1st overtone) of —CH$_3$ (1 666, 1 681 and 1 790 nm) and other characteristic absorption. The characteristic absorption of benzene-alcohol extractives and 1% NaOH extractives on the optimal bands was analyzed from the point of view of molecular structure, and the performance of models was explained theoretically. The models can meet the actual demand and can be applied to the analysis of the content of Eucalyptus extractives in pulping and papermaking industry. The results showed that performance of near-infrared models can be developed and optimized by the selection of pretreatment and modeling methods combined with the genetic algorithm for the prediction of Eucalyptus extractives. At the same time, as an emerging algorithm, LASSO algorithm has a good ability to process co-complex linear data in near-infrared spectroscopy, and can establish models with good analysis performance.

Keywords Near-infrared technique; LASSO algorithm; Pretreatment; Genetic algorithm

(Received Feb. 27, 2019; accepted May 15, 2019)

* Corresponding author