烟草水溶性糖近红外定量模型中光谱范围选择方法的研究

夏骏，陆扬，苏燕，潘力，林垦，朱书秀，陆明华
浙江中烟工业有限责任公司，浙江 杭州 310024

摘　要：运用四种不同的光谱范围选择方法来建立烟草中水溶性糖的近红外定量模型，发现模型的交互验证系数、交互验证均方差和预测均方差有明显的差异。通过对烟草中水溶性糖的分子结构分析，结合傅里叶变换近红外漫反射光谱的特性，初步确定烟草水溶性糖近红外定量模型的建模光谱范围，以交互验证系数和交互验证均方差为评价指标进一步优化光谱范围，可以得到烟草水溶性糖在近红外定量模型中的最佳光谱范围为 3850~5010 cm^{-1}、5720~7010 cm^{-1} 和 7760~7980 cm^{-1}，总糖和还原糖定量模型的交互验证系数、交互验证均方差和预测均方差分别为 0.989、0.787、0.565 和 0.982、0.801、0.693。

关键词：近红外；烟草；定量模型；光谱范围；水溶性糖；交互验证
引用本文：夏骏，陆扬，苏燕 , 等 . 烟草水溶性糖近红外定量模型中光谱范围选择方法的研究 [J]. 中国烟草学报，2015,21（2）

烟草中的水溶性糖对烟叶品质具有重要影响 [1]，是决定烟气醇和度的主要因素 [2]。在烟草工业中，水溶性糖含量的分析测定是日常检测项目之一，对卷烟成品质量的控制具有重要的意义。由于传统的化学检测方法比较复杂，费时长，消耗大 , 因此近些年来，简单快速的近红外光谱分析技术在烟草行业中的应用颇为广泛 [3-5]。但是近红外光谱分析技术是利用化学计量学方法从复杂的光谱数据中提取有用的样品含量信息，所以其分析结果容易受到多种因素的影响。李军会等研究了样品装样、测试条件等因素对近红外检测结果的影响，并实现了通过建立全局模型应用自校正方法来降低分析误差 [6]。段焰青等考察了粒度对烟末总糖、总氮和烟碱含量近红外预测值的影响，结果表明当烟末粒度在 40 目以上才能保证预测数据的准确性和精密性 [7]。练文柳等采用基线校正、去卷积、一阶微分、二阶微分、主成分回归和偏最小二乘法对烟叶样品的近红外光谱数据进行处理，建立了相应的校正模型，并作比较，发现运用二阶微分和偏最小二乘法建立的模型效果最好 [8]。马翔等选择不同的谱区范围对烟草定量模型进行优化，结果显示不同谱区范围对定量模型的影响明显 [9]。但对建立近红外定量模型过程中如何确定最优光谱范围，目前还未见详尽的报道。本文通过研究水溶性糖的分子结构，推断出水溶性糖中含氢基团在近红外谱区的合频与倍频吸收范围，并以近红外模型的交互验证系数和交互验证均方差为评价指标，确定了烟草水溶性糖在建立近红外定量模型时的最优光谱范围。

1　实验部分

1.1　仪器设备

Antaris 傅里叶变换近红外光谱仪（Thermo，美国），配有积分球漫反射采样系统和样品杯旋转采样附件；Futura 全自动连续流动分析仪（Alliance，法国）；Cyclotec 1093 旋风式样品磨（Foss Tecator，丹麦）；SC101 型鼓风电热恒温干燥箱（浙江嘉兴新胜电器厂）。

1.2　样品与处理

收集 700 个 2011~2013 年烤烟烟叶样本，产区包括云南、四川、山东、江西、湖南、湖北、河南、贵州、广西、福建、安徽共 11 个，等级包括 B1F、B2F、B3F、C1F、C2F、C3F、X1F、X2F、X3F 共 9 个，样本涵盖主要产区和等级。将所有样本按照烟草行业标准《YC/T 31-1996 烟草及烟草制品 试样的制备和水分测定 烘箱法》制备成粉末样本 (粒径≥ 40 目，水分含量 5%~7%)，其中 600 个作为建模样本集，100 个作为预测样本集。

1.3　近红外光谱采集

将烟叶粉末样本装入样品杯，并用压样器压实，

基金项目：浙江中烟工业有限责任公司科技项目“近红外光谱法在烤烟烟叶化学成分测定中的应用研究”（ZJZY2013C004）
作者简介：夏骏（1978—），硕士，工程师，主要从事烟草化学分析，Tel：0571-81188571，Email：xiajun@zjtobacco.com
通讯作者：陆明华 （1965—），本科，工程师，主要从事烟草化学分析，Tel：0571-81188293，Email：lumh@zjtobacco.com
收稿日期：2014-06-05

然后采集光谱，光谱采集范围：3800~10000 cm^{-1}，光谱分辨率：8 cm^{-1}，扫描次数：64 次，样品杯方式：旋转。

1.4 水溶性总糖和还原糖含量的测定

按照烟草行业标准《YC/T 159-2002 烟草及烟草制品 水溶性糖的测定 连续流动法》对所有样本的水溶性总糖和还原糖含量进行测定。结果如表 1 所示。

表 1 连续流动法测得的烟草中水溶性糖含量

Tab. 1 Water-soluble sugar content in tobacco by continuous flow method

测定项目	最小值 /%	最大值 /%	平均值 /%	标准差 /%
水溶性总糖	11.05	45.81	31.71	5.29
水溶性还原糖	10.36	36.01	25.77	4.16

1.5 数据处理

使用 TQ Analyst 软件进行水溶性总糖和还原糖近红外定量模型的建立和优化，建模方法选择偏最小二乘法，光程使用多元散射进行校正，光谱数据采用二阶微分并用诺里斯导数滤波法进行平滑处理。光谱范围：（a）全光谱范围：3800-10000 cm^{-1}；（b）TQ Analyst 软件自动优化选择光谱范围：4331.34-4416.19 cm^{-1}、5688.98-5847.11 cm^{-1} 和 5958.96-6012.96 cm^{-1}；（c）从文献引用的光谱范围：4246.70-7502.10 cm^{-1}[9]；（d）通过分子化学结构分析确定的光谱范围。

2 结果与讨论

2.1 近红外定量模型的评价

评价近红外定量模型的优劣一般采用内部交互验证和外部验证，所谓内部交互验证是指从建模样本集中剔除一个或多个样本，用剩余的样本建立模型来预测剔除的样本，重复该过程直至所有样本都被剔除并预测过，该验证主要考察模型的稳定性和内部预测能力，避免模型出现过拟合现象，评价指标为交互验证系数（Q^2）和交互验证均方差（*RMSECV*）；而外部验证则是用已建立的模型来预测多个未知样本，主要考察模型的实际预测能力，是评价模型好坏的主要方法，评价指标为预测均方差（RMSEP）。

交互验证系数、交互验证均方差和预测均方差的计算公式如下：

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

$$RMSECV = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2)$$

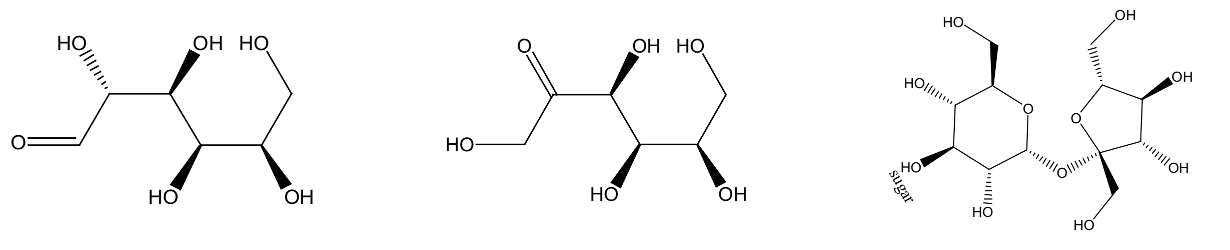
$$RMSECV = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

其中 n 为模型的样本数量，y_i 为模型中第 i 个样本的参考值，$\hat{y}_i$ 为模型中第 i 个样本的交互验证预测值，$\bar{y}$ 为模型中所有样本参考值的平均值，N 为预测样本集的样本数量，y_{Ni} 为预测样本集中第 i 个样本的参考值，$\hat{y}_{Ni}$ 为预测样本集中第 i 个样本的预测值。

从以上公式可以看出：交互验证系数越大，交互验证均方差越小，说明模型的稳定性越好，其内部预测能力越强；预测均方差越小，模型的实际预测效果越好。

2.2 水溶性糖在近红外光谱吸收范围的确定

烟草中的水溶性糖主要包括单糖(葡萄糖、果糖)、双糖(蔗糖)，其分子结构如下图所示：



葡萄糖 Glucose　　果糖 Fructose　　蔗糖 Sucrose

图 1 烟草中水溶性糖的分子结构图

Fig.1 Molecular structure of water-soluble sugar in tobacco

从图 1 可以看出，烟草中水溶性糖中所包含的含氢基团主要就是 C—H 和 O—H 基团，因此烟草中水溶性糖在近红外谱区的吸收主要就是 C—H 和 O—H 基团基频振动的合频与倍频吸收。C—H 基团伸缩振动的基频吸收带在中红外区的 3000 cm^{-1} 附近，弯曲振动在 1450 cm^{-1} 左右；O—H 基团伸缩振动的基频吸收带约在中红外区的 3650 cm^{-1}，弯曲振动约在 1300 cm^{-1}。考虑到实际的分子振动并不完全符合简谐振动，而是属于非线性谐振，所以倍频的实际频率比基频乘以倍频数的计算值略小，由此推测出 C—H 基团和 O—H 基团的倍频与合频吸收带的近似位置，如表 2 所示。

表 2 烟草中水溶性糖中 C—H 基团和 O—H 基团倍频与合频吸收带的近似位置

Tab. 2 Approximate absorption band of water-soluble sugar in tobacco

基团	C—H	O—H
一级倍频 /cm^{-1}	5900	7000
二级倍频 /cm^{-1}	8800	10700
合频 /cm^{-1}	7450; 4450	4950

同时从近红外漫反射光谱的背景扫描图中发现在 5100~5560 cm^{-1} 和 7010~7440 cm^{-1} 处存在较大的背景干扰。

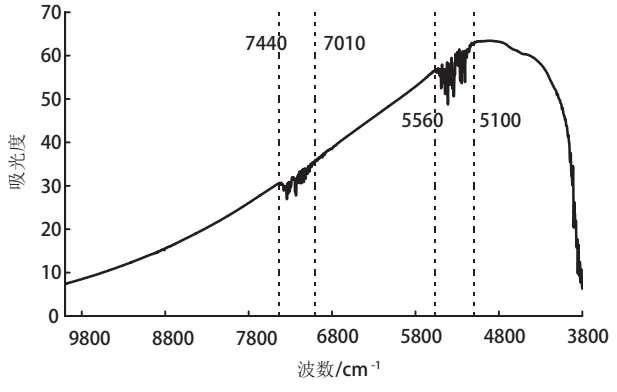


图 2 近红外漫反射光谱的背景扫描图

Fig. 2 Background spectrum of FT-NIR diffuse reflectance spectroscopy

由此初步确定了 C—H 基团和 O—H 基团在近红外谱区的吸收范围：3950~5100 cm^{-1}；5560~7010 cm^{-1}；7440~7950 cm^{-1}；8300~9300 cm^{-1}。

C—H 基团和 O—H 基团是有机分子组成的基本结构，在水溶性糖以外的其他成分中大量存在。为使最终得到的光谱范围和烟草中水溶性糖的含量有较大的相关性，所以需要结合连续流动法测得的烟草中水溶性糖含量数据来建立相应的近红外定量模型，并以近红外模型的交互验证系数和交互验证均方差为评价指标对每一个光谱区间的上下限进一步优化。发现当光谱区间选择 3850~5010 cm^{-1}、5720~7010 cm^{-1} 和 7760~7980 cm^{-1} 时，该两项指标达到最优。

2.3　不同的光谱范围选择方法建模结果

用三种不同的光谱范围选择方法建立水溶性总糖和还原糖的近红外定量模型，并对含 100 个样本的预测样本集进行预测。按照 1.5 中提到的四种选择方法选择光谱范围，其预测结果见表 3 和表 4。

表 3 不同光谱范围选择方法对水溶性总糖近红外定量模型的影响

Tab. 3 Effect of different spectral range selection methods on total sugar model

光谱范围选择方法	交互验证系数（Q^2）	交互验证均方差（RMSECV）	预测均方差（RMSEP）
a	0.962	1.04	0.606
b	0.916	1.55	1.18
c	0.976	0.819	0.595
d	0.978	0.787	0.565

表 4 不同光谱范围选择方法对水溶性还原糖近红外定量模型的影响

Tab. 4 Effect of different spectral range selection methods on reduced sugar model

光谱范围选择方法	交互验证系数（Q^2）	交互验证均方差（RMSECV）	预测均方差（RMSEP）
a	0.955	0.906	0.770
b	0.857	1.57	1.43
c	0.962	0.811	0.709
d	0.964	0.801	0.693

从表 3 和表 4 的数据可以看出，通过不同的光谱范围选择方法来建立的近红外定量模型，其交互验证系数、交互验证均方差和预测均方差存在较大的差异，通过分子化学结构分析确定光谱范围后建立模型的内部验证指标明显优于其他 3 种方法，其外部验证的实际预测效果也最好。最终得到水溶性总糖、还原糖的化学测量值和模型预测值的关系如图 3 和图 4 所示。

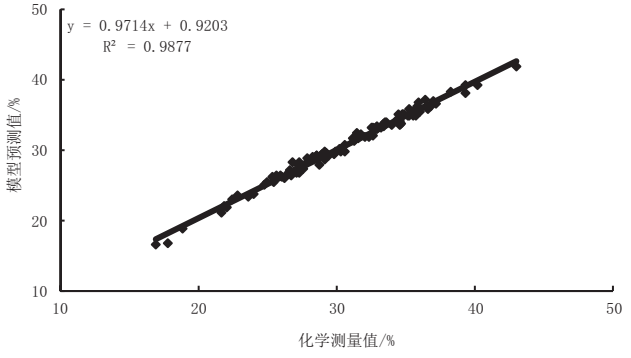


图 3 水溶性总糖的化学测量值和模型预测值的关系

Fig. 3 Relationship between total sugar standard value and calculated value

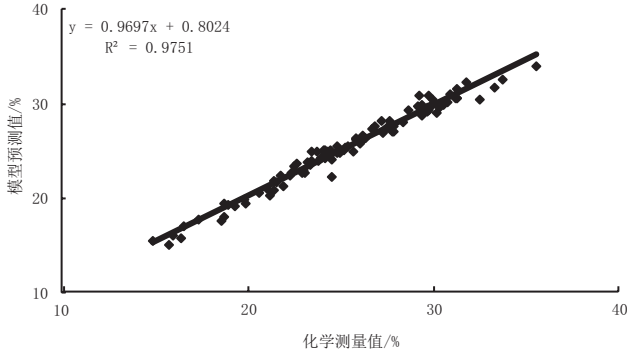


图 4 水溶性还原糖的化学测量值和模型预测值的关系

Fig. 4 Relationship between reduced sugar standard value and calculated value

3 结论

本文通过对烟草中水溶性糖的分子结构分析，结合傅里叶变换近红外漫反射光谱的特性，确定了水溶性糖中含氢基团在近红外谱区倍频与合频吸收带的近似位置，并结合连续流动法测得的烟草中水溶性糖含量数据进一步优化，得到最佳的光谱范围，在此基础上建立了水溶性总糖和还原糖的近红外定量模型，取得了较好的结果，模型的交互验证系数、交互验证均方差和预测均方差都优于其他 3 种光谱范围选择方法。该方法可以有效的应用于烟草水溶性总糖和还原糖含量的实际分析。

参考文献

[1] 闫克玉 . 烟草化学 [M]. 郑州：郑州大学出版社，2002：51-56.

[2] 卷烟工艺（第二版）编写组 . 卷烟工艺（第二版）[M]. 北京：北京出版社，2000：88-93.

[3] 邓亮，冷红琼，段沅杏，等 . FT-NIR 光谱测定烟草中烟碱、总氮、总糖含量的模型研究 [J]. 云南农业大学学报，2013，28(6)：814-818.

[4] 邓发达，朱立军，戴亚，等 . 近红外技术测定成品卷烟中总糖和还原糖及绿原酸的含量 [J]. 安徽农业科学，2010，38(12)：6181-6182，6188.

[5] 张建平，谢雯燕，束如欣，等 . 烟草化学成分的近红外快速定量分析研究 [J]. 烟草科技，1999，136(3)：37-38.

[6] 李军会，秦西云，张文娟，等 . 样品装样、测试条件等因素对近红外检测结果的影响与分析误差源比较研究 [J]. 光谱学与光谱分析，2007，27(9)：1751-1753.

[7] 段焰青，周红，王明锋，等 . 粒度对烟末总糖、总氮和烟碱含量 NIR 预测值的影响 [J]. 烟草科技，2005，(7)：22-23.

[8] 练文柳，吴名剑，孙贤军，等 . 不同预处理方法对烟草近红外光谱预测模型的影响 [J]. 烟草科技，2005, (2)：19-23.

[9] 马翔，王毅，温亚东，等 . FT—NIR 光谱仪测定烟草化学成分不同谱区范围对数学模型影响的研究 [J]. 光谱学与光谱分析，2004，24(4)：444-446.

Spectral range selection method in NIR quantitative model of tobacco water-soluble sugar

XIA Jun, LU Yang, SU Yan, PAN Li, LIN Ken, ZHU Shuxiu, LU Minghua
China Tobacco Zhejiang Industrial Co., Ltd, Hangzhou 310024, China

Abstract：Four different spectral range selection methods were applied in NIR quantitative model of tobacco water-soluble sugar. It was found that correlation coefficients of cross validation（R^2）, root mean square errors of cross validation (RMSECV) and root mean square errors of prediction (RMSEP) were significantly different. Spectral ranges were preliminarily determined by analyzing molecular structures of tobacco water-soluble sugar and characteristic of FT-NIR diffuse reflectance spectroscopy. Two evaluating indicators, R^2 and RMSECV, were used to further optimize spectral range. It was found that the optimal spectral ranges of tobacco water-soluble sugar was 3850~5010 cm^{-1}, 5720~7010 cm^{-1}and7760~7980 cm^{-1}. The final quantitative models of total sugar and reduced sugar, R^2, RMSECV and RMSEP were 0.989, 0.787, 0.565 and 0.982, 0.801, 0.693, respectively.

Keywords：near-infrared; tobacco; quantitative model; spectral range; water-soluble sugar; cross validation

Citation: XIA Jun, LU Yang, SU Yan, et al. Spectral range selection method in NIR quantitative model of tobacco water-soluble sugar [J]. Acta Tabacaria Sinica, 2015,21 (2)