〈红外应用〉

白砂糖色值近红外光谱分析的波段选择

梁家杰[1,2]，潘 涛[1,2*]，陈星旦[2,3]，胡愉华[2]，熊德先[4]，林喜令[4]，谢 军[4]

（1.“重大工程灾害与控制”教育部重点实验室（暨南大学），广东 广州 5106322；
2. 暨南大学光电工程系，广东 广州 510632；
3.中国科学院长春光学精密机械与物理研究所应用光学国家重点实验室，吉林 长春 130033；
4.暨南大学信息科学技术学院，广东 广州 510632）

摘要：采用近红外漫反射光谱技术和偏最小二乘法（PLS）建立白砂糖色值的定量分析模型。用多元散射校正方法对光谱进行预处理，再用 Savitzky-Golay 平滑化方法对原谱、一阶导数谱和二阶导数谱进行处理。选取5个波段，每个波段分别采用原光谱、一阶导数谱、二阶导数谱。同时调整 Savitzky-Golay 平滑点数和 PLS 因子数，通过多次 PLS 数值实验比较，按照预测效果确定每个模型的最优平滑点数、因子数，再从中选优。结果表明，采用 780～1100 nm 一阶导数谱的定标效果最好，模型的预测均方根偏差、相对预测均方根偏差分别为 11.2，8.91%。780～1100 nm 可以代替近红外全谱波段（780～2500 nm）得到好的定量分析效果，为设计小型专用近红外分析仪器提供依据。

关键词：近红外光谱分析；偏最小二乘法；波段优选；白砂糖；色值
中图分类号：O657.33 **文献识别码**：A **文章编号**：1001-8891(2009)02-0090-05

Choice of Wave Band in Near Infrared Spectroscopy Analysis of Color Value of White Granulated Sugar

LIANG Jia-jie[1,2]，PAN Tao[1,2*]，CHEN Xing-dan[2,3]，HU Yu-hua[2]，XIONG De-xian[4]，LIN Xi-ling[4]，XIE Jun[4]

(1. *Key Laboratory of Disaster Forecast and Control in Engineering, Ministry of Education of the People's Republic of China (Jinan University), Guangzhou Guangdong* 510632, *China*;
2. *Department of Optoelectronic Engineering，Jinan University, Guangzhou Guangdong* 510632, *China*;
3. *National Key Laboratory of Applied Optics, Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun Jilin* 130033, *China*;
4. *College of information Science and Technology, Jinan University, Guangzhou Guangdong* 510632, *China*)

Abstract：The calibration models for color value of white granulated sugar are constructed by near infrared diffuse reflection spectroscopy technology and partial least squares (PLS) regression. The spectra is pretreated by multiplicative scatter correction method, and makes use of Savitzky-Golay smoothing method to do smooth processing to the original spectra, the first derivative spectra and the second derivative spectra. 5 wave bands are selected, and adopting the original spectra, the first derivative spectra and the second derivative spectra respectively in each band. By adjusting Savitzky-Golay smoothing point number and the PLS factor number simultaneously, compared with many PLS computational experiments, it gets optimal smoothing point number and factor number for each model according to the prediction effect, and then selects the best one from them. The result shows that the prediction effect by using the first derivative spectra in 780～1100 nm is best, and the root mean square error of prediction and the relative root mean square error of prediction for the corresponding model are 11.2, 8.91% respectively. It shows that 780～1100 nm band

收稿日期：2008-07-06
作者简介：梁家杰（1980-），男，硕士研究生，研究方向：近红外、红外光谱技术及其应用。
*通讯作者：潘涛，tpan@jnu.edu.cn
基金资助：国家自然科学基金（10771087）；广东省科技计划粤港关键领域重点突破项目（2007A020905001）；广东省科技计划项目（2007B030501008；2007B020714001）；广州市科技攻关项目（2007Z3-E0281）；教育部留学归国人员科技启动基金（2005-383）

can replace the whole band of near infrared (780～2500 nm) and get good quantification effect, which provide base for designing minitype special near infrared analyzer.

Key words：near infrared spectroscopy analysis；partial least squares regression；optimal wave band；white granulated sugar；color value

引言

在甘蔗制糖过程管理和品质分析中，白砂糖的色值是需要快速检测的重要评价指标。白砂糖色值的常规分析方法是在实验室用化学试剂和化学反应来完成[1]，不能实现快速和在线测定，是长期以来制糖工业中需要解决的问题。

现代近红外光谱技术以其分析速度快、精度适中、成本低、非破坏性、易于实现在线实时分析以及多指标同时测定等优点，已在农业、食品、医药、烟草、石油化工等领域中得到应用[2-7]。本文采用近红外漫反射光谱技术和偏最小二乘法（PLS）方法建立白砂糖色值的快速定量分析模型。用多元散射校正（MSC）方法做光谱预处理，再用 Savitzky-Golay 方法进行原谱、一阶导数谱和二阶导数谱的平滑化处理。根据样品色值与吸光度的相关系数谱和分子振动类型选取若干光谱波段，每个波段都分别采用原光谱、一阶导数谱、二阶导数谱建立定标模型。为了提高模型精度，同时调整平滑点数和因子数，并对每个模型都分别做多次 PLS 数值实验，得到每个模型的最优平滑点数、因子数和预测均方根偏差（RMSEP），再按照 RMSEP 值从中选优，确定定标效果最好的波段，为设计小型专用近红外快速分析仪器提供依据。

1 实验部分

1.1 实验材料、仪器和测量方法

广东某糖厂提供白砂糖样品 97 份及其色值的参考化学值数据。

实验仪器是丹麦 FOSS 公司的 XDS Rapid Content 光栅型近红外光谱分析仪和漫反射附件。光谱采集范围为 400～2500 nm。探测器为 Si（400～1100 nm）和 PbS（1100～2500 nm）。实验室温度为 25℃±1℃，湿度为 46%RH。为了针对小型专用近红外分析仪器的研发，参照美国材料检测协会关于近红外谱区的定义，本文选取近红外谱区（780～2500 nm）为研究范围。

取适量的白砂糖样品均匀置于方形样品池中，用漫反射移动式扫描获得白砂糖的近红外光谱，每个样品重复采集 3 次光谱，采用平均光谱。

1.2 光谱数据处理

近红外光谱是一种间接分析技术，需要借助样品的参考化学值和光谱数据一起建立定标模型。一个正确稳健的近红外分析定标模型要求有准确可靠的参考化学值和光谱数据。但是由于各种原因，在实验室用常规方法获得的参考化学值和用近红外光谱仪获得的光谱数据都存在误差，按照统计规律，必然存在偏离吸光度-浓度线性模型的少量的所谓“异常样品”。确定剔除异常样品的方法很多，本文通过单波长吸光度-浓度线性模型来确定剔除异常样品。首先计算 97 个白砂糖样品在每个波长点上的吸光度和色值参考化学值的相关系数、色值的计算值与化学值的均方根偏差，建立每个波长的吸光度-浓度线性模型。然后按照相关系数高和均方根偏差低挑选出 12 个特征波长。根据这 12 个波长对应模型的色值计算值与化学值的偏差状况来评判样品，按总样品数 10%左右的幅度，剔除 10 个异常样品，得到 87 份样品用于建模。再根据浓度范围和分布均匀的原则，从 87 个样品中确定定标样品（56 个）和验证样品（31 个）的集合。表 1 为用于建模的 87 个白砂糖样品的色值的参考化学值数据统计。

表 1 白砂糖样品的色值的参考化学值数据统计

Table 1 The data statistics for the referenced chemical values of color value of white granulated sugar samples

	色值（IU）			
	最小值	最大值	平均值	标准差
定标集	102	149	125.5	10.9
验证集	102	146	125.7	11.4

多元散射校正（MSC）的作用是校正吸收基线并降低样品散射作用对光谱的影响，降低样品的不均匀性带来的光谱差异性[7,8]。因为本实验的样品为固体，采用漫反射附件测定不是很均匀，因此，首先用 MSC 来做光谱预处理。

导数光谱法可以消除基线漂移或平缓背景干扰的影响，除去光谱信号中高频噪声的干扰。本文采用常用的 Savitzky-Golay 平滑化法[9-10]对原谱、一阶导数谱和二阶导数谱进行预处理。一般而言，平滑点数过大可以使信噪比提高，但同时也会导致信号的失真，而平滑点数过少容易产生新的计算误差从而造成模型精度下降。因此，合理使用平滑点数非常重要，但必须通过多次数值实验的比较，才能确定最优平滑点数。由于运算和处理的工作量很大，在既往的研究中把平

滑点数作为参数来优化定标模型的工作很少。本文将在这方面进行探讨。

采用运用广泛的 PLS 方法[3-5,7]建立定标模型。模型评价指标包括定标相关系数（R_c）、预测相关系数（R_p）、定标均方根偏差（RMSEC）、预测均方根偏差（RMSEP）、相对定标均方根偏差（RRMSEC）和相对预测均方根偏差（RRMSEP），其中：

$$\mathrm{RMSEC}=\sqrt{\frac{\sum_{i=1}^{n}(y_i-y_{ic})^2}{n-1}}\quad;\quad \mathrm{RMSEP}=\sqrt{\frac{\sum_{i=1}^{m}(y_i-y_{ip})^2}{m-1}}\quad;$$

$$\mathrm{RRMSEC}=\frac{\mathrm{RMSEC}}{y_{mc}}\times 100(\%)\quad;\quad \mathrm{RRMSEP}=\frac{\mathrm{RMSEP}}{y_{mp}}\times 100(\%)$$

式中：y_i 为第 i 样品的化学值；y_{ic} 为定标集中第 i 样品的预测值；y_{ip} 为验证集中第 i 样品的预测值；y_{mc} 为定标样品化学值的平均值；y_{mp} 为验证样品化学值的平均值；n 为定标集的样品数；m 为验证集的样品数。从表1看出，白砂糖样品的色值范围很窄，预测值与化学值的相关系数 R_c、R_p 不适合于评价模型的定量分析的精度，因此，本文主要采用 RMSEC、RMSEP、RRMSEC、RRMSEP 来评判模型的预测效果。

2 结果与讨论

图1给出了87个白砂糖样品的在近红外谱区范围（780～2500 nm）内的漫反射光谱。由图可见，不同组分的吸收谱重叠严重，吸收较弱，图中没有显著的某种成分的特征吸收峰。图 2 给出了白砂糖样品的吸光度与色值的参考化学值相关系数谱。由图可见，虽然单波长点上的相关系数普遍不高，但区别很明显，从而可以根据相关系数的高低选择适当谱区来尝试建立定标模型。其中 4 段组合（950～1100 nm＋1150～1400 nm＋1800～1900 nm＋2050～2150 nm）是对应于相关系数谱中上峰位和下峰位波区组合。另一方面，780～1100 nm 是短波近红外区，对应光谱仪探测器为 Si，并对应分子振动的二倍频（含高频）波段。1100～2500 nm 是长波近红外区且对应光谱仪探测器为 PbS，其中 1100～1850 nm、1850～2500 nm 分别对应分子振动的一倍频、合频波段。还有近红外的全谱区（780～2500 nm），共 5 个波段。即本文选取的 5 个波段：780～2500 nm，780～1100 nm，1100～1850 nm，1850～2500 nm，4 段组合（950～1100 nm＋1150～1400 nm＋1800～1900 nm＋2050～2150 nm）。对上述 5 个波段都分别采用原光谱、一阶导数谱、二阶导数谱，共建立 15 个定标模型从中比较优选。

采用 PLS 方法建模时，PLS 因子数是重要的调整参数，如果使用的因子数过少，就不能充分反映样品信息，模型预测精度会降低，如果使用的因子数过多，就会引入一些代表噪声干扰的成分数据，模型的预测能力也会下降。因此，合理确定 PLS 因子数，对于充分利用光谱信息和消除噪声非常重要，但必须通过多次数值实验的反复比较，才能选取确定。本文以预测均方根偏差（RMSEP）作为建模的优化目标，以光谱预处理的 Savitzky-Golay 平滑点数和 PLS 因子数为建模的优化参数，因此，这是一个双参数的目标优化问题。

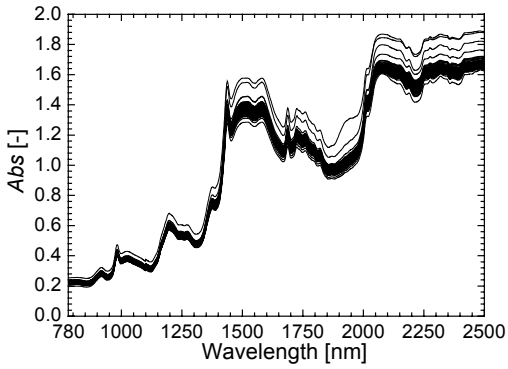

图 1 87 个白砂糖样品的近红外漫反射光谱

Fig.1 The near infrared diffuse reflection spectra of 87 white granulated sugar samples

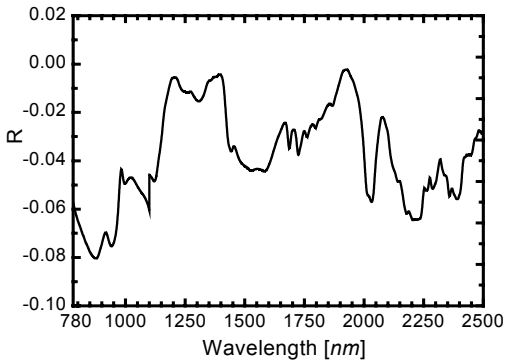

图 2 白砂糖样品吸光度与色值参考化学值的相关系数谱

Fig.2 The correlation coefficient spectrum for the absorbance and the referenced chemical values of color value for white granulated sugar samples

每个定标模型都设定 3、4、5、6、7 共 5 个因子数，并设定 3～51（取奇数）共 25 种平滑点数，从而有很多参数组合。对每一个模型，每一个参数组合都做一次 PLS 数值实验，记录每次数值实验的参数组合（平滑点数、因子数等）和结果（每个验证样品的色值预测值、RMSEP 等），比较选优。

以 780～1100 nm 一阶导数谱为例，表 2 给出了每一个的参数组合（平滑点数、因子数）对应的数值实验结果（RMSEP）。其中，方块内的每一个 RMSEP 值

的所在行最左端的数和所在列最上端的数分别是所采用的平滑点数和因子数。表内用小方框围住的RMSEP值，表示它是同一列中的最小值，通过它标明了同一因子数对应的最优平滑点数；表内用黑体和有下横线的RMSEP值，表示它是同一行中的最小值，通过它标明了同一平滑点数对应的最优因子数，如果同一行或同一列当中有两个相同的最小值，同时标出。从表2可以看出，最优参数组合的因子数和平滑点数分别为11和7，最优的RMSEP值是11.2。

表2 采用780～1100 nm波段一阶导数谱的偏最小二乘法的预测均方根偏差（RMSEP）

Table 2 The root mean square error prediction for the partial least squares method by using the first derivative spectra for whole region (RMSEP)

		因子数				
		9	10	11	12	13
平滑点数	3	11.7	11.9	11.3	11.8	12.3
	5	11.7	11.9	11.4	11.8	12.2
	7	11.8	11.8	11.2	11.7	12.5
	9	12.1	12.2	11.6	11.6	13.0
	11	12.3	12.5	12.0	11.8	13.7
	13	12.5	12.8	12.4	12.3	14.6
	15	12.7	13.1	12.4	12.8	14.8
	17	12.8	13.1	12.6	12.8	13.8
	19	12.9	13.2	12.7	12.8	12.9

表3给出了通过全部的数值实验比较得到的15个模型对应的最优的平滑点数、因子数和RMSEP值。从表看出，所有15个模型的最优参数组合（平滑点数，因子数）都不相同。5个波段的最好的模型是：采用全谱段（780～2500 nm）一阶导数谱、最优因子数、平滑点数和RMSEP值为7、17、11.9；采用780～1100 nm一阶导数谱，最优因子数、平滑点数和RMSEP值为11，7，11.2；采用1100～1850 nm一阶导数谱，最优因子数、平滑点数和RMSEP值为8、11、11.7；采用1850～2500 nm二阶导数谱，最优因子数、平滑点数和RMSEP值为2、3、12.4；采用4段组合（950～1100 nm＋1150～1400 nm＋1800～1900 nm＋2050～2150 nm）二阶导数谱，最优因子数、平滑点数和RMSEP值为4、3、11.6。

其中效果最好的是采用780～1100 nm一阶导数谱，RMSEP、RRMSEP分别为11.2、8.91%。采用全谱区（780～2500 nm）一阶导数谱也有好的预测效果，RMSEP、RRMSEP分别为11.9、9.47%。这两个模型得到的色值预测值和化学值的比较分别在图3、4中给出。由此可见，780～1100 nm可以代替近红外全谱区（780～2500 nm）得到更好的定量分析效果。

表3 15个定标模型的最优的Savitzky-Golay平滑点数、PLS因子数和RMSEP的比较

Table 3 Comparison of optimal Savitzky-Golay smoothing point number, PLS factor number and RMSEP in for 15 calibration models

		最优因子数	最优平滑点数	最优RMSEP
780～2500	Spectrum	15	45	13.5
	1st	7	17	11.9
	2nd	8	17	12.1
780～1100	Spectrum	20	39	12.2
	1st	11	7	11.2
	2nd	7	5	12.8
1100～1850	Spectrum	20	3	22.9
	1st	8	11	11.7
	2nd	10	51	11.9
1850～2500	Spectrum	15	39	17.8
	1st	5	7	13.6
	2nd	2	3	12.4
4段组合	Spectrum	16	9	20.9
	1st	8	3	18.5
	2nd	4	3	11.6

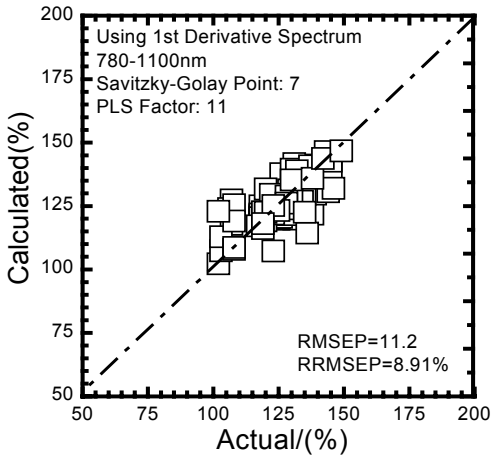

图3 采用780～1100 nm波段一阶导数谱的最优定标模型的白砂糖色值的预测值和参考化学值的比较

Fig.3 Comparison of the calculated values with the referenced chemical values of color value of white granulated sugar samples in optimal calibration model by using the first derivative spectra in region (780～1100 nm)

3 结论

采用近红外漫反射光谱技术和PLS方法建立白砂糖色值的快速定量分析模型。用MSC方法做光谱预处

理，再用 Savitzky-Golay 方法对原谱、一阶导数谱和二阶导数谱进行平滑化处理。根据样品吸光度与色值的相关系数谱和分子振动类型选取下列 5 个波段：780～2500 nm、780～1100 nm、1100～1850 nm、1850～2500 nm、4 段组合（950～1100 nm＋1150～1400 nm＋1800～1900 nm＋2050～2150 nm）。每个波段分别采用原光谱、一阶导数谱、二阶导数谱，共建立 15 个定标模型。1）同时调整 Savitzky-Golay 平滑点数和 PLS 因子数，通过多次 PLS 数值实验比较，按照预测效果（RMSEP 值最小）确定每个模型的最优平滑点数、因子数，再从 15 个模型中选优，结果表明，采用 780～1100 nm 一阶导数谱的定标效果最好；2）采用 780～1100 nm 一阶导数谱定标建模，相应的 RMSEP、RRMSEP 分别为 11.2、8.91%，这样的定量分析精度已经可以用于甘蔗制糖过程的快速监测。3）780～1100 nm 可以代替近红外的全谱区（780～2500 nm）得到更好的定量分析效果，为设计小型专用近红外分析仪器提供依据。4）要获得高精度的定标模型必须考虑波段的不同选取、不同光谱预处理方式的选取、不同平滑点数和不同因子数的选取等多种情形和多种参数变化来比较选优。虽然运算和处理的工作量很大，但这是改善提高近红外光谱分析预测能力的有效措施。

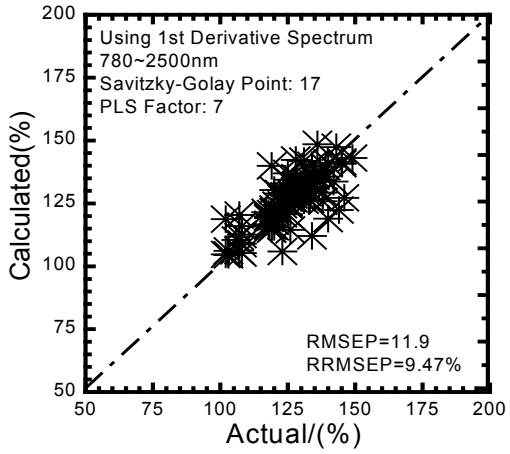

图 4 采用 780～2500 nm 一阶导数谱的最优定标模型的白砂糖色值的预测值和参考化学值的比较

Fig.4 Comparison of the calculated values with the with the referenced chemical values of color value of white granulated sugar samples in optimization calibration model by using the first derivative spectra in region (780～2500 nm)

参考文献：

[1] GB317-1998 白砂糖.

[2] WILLIAMS P, NORRIS K. *Near-infrared Technology in the Agricultural and Food Industries (Second Edition)*[M]. Minnesota (USA): the American Association of Cereal Chemists, Inc. St. Paul, 2001.

[3] 陆婉珍. **现代近红外光分析技术（第二版）**[M]. 北京: 中国石化出版社, 2007.

[4] 严衍禄. **近红外光谱分析基础与应用**[M]. 北京: 中国轻工业出版社, 2005.

[5] 丁海泉, 卢启鹏, 陈星旦, 等. 土壤有机质近红外光谱分析组合波长的优选[J]. **光学精密工程**, 2007, **15**(12): 1946-1951.

[6] 陈华才, 吕进, 陈星旦, 等. 基于径向基函数网络的茶多酚总儿茶素近红外光谱检测模型的研究[J]. **光学精密工程**, 2006, **14**(1): 58-62.

[7] 赵强, 张工力, 陈星旦. 多元散射校正对近红外光谱分析定标模型的影响[J]. **光学精密工程**, 2005, **13**(1): 53-58.

[8] 张银, 周孟然. 近红外光谱分析技术的数据处理方法[J]. **红外技术**, 2007, **29**(6): 345-348.

[9] 陈洁梅, 潘涛, 陈星旦. 二阶导数光谱预处理在用 FTIR/ATR 方法定量测定葡萄糖-6-磷酸和果糖-6-磷酸中的应用[J]. **光学精密工程**, 2006, **14**(1): 1-7.

[10] SAVITZKY A, GOLAY M J E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures[J], *Analytical Chemistry*, 1964, **36**: 1627-1637.

（上接第 86 页）

参考文献：

[1] 刘恩科, 朱秉升, 罗晋生, 等. **半导体物理学**[M]. 北京: 国防工业出版社, 1994: 179.

[2] 管志斌, 王立, 江风益, 等. 两步镀膜 Ti/Al/Ti/Au 的 n 型 GaN 欧姆接触研究[J]. **功能材料与器件学报**, 2003, **9**(3).

[3] 黄江平, 杨春丽, 黎力, 等. 128×128 混合式热释电非致冷焦平面探测器列阵铟膜及铟柱制备工艺研究[J]. **红外技术**, 2003, **25**(6): 54-58.