实验误差对近红外模型准确性的影响

姚 胜, 武国峰, 周舒珂, 姜亦飞, 金小娟, 赵 强, 蒲俊文*

北京林业大学材料学院, 北京 100083

摘 要 以相思树聚戊糖含量为例, 通过用不同精确度的数据建立的近红外模型预测性能, 讨论了不同精确度的数据对近红外模型准确性的影响。结果表明, 建模原始数据的精确度在一定程度上影响着近红外模型的预测性能, 精确度越高, 建立的模型越好。但对于精确度较小的的样品, 所建立的模型预测性能也能较好的预测未知样品。

关键词 近红外模型; 实验误差; 聚戊糖
中图分类号: O657.3 文献标识码: A **DOI**: 10.3964/j.issn.1000-0593(2011)05-1216-04

引 言

近红外定量分析技术是一种间接的分析技术, 近红外光谱模型是利用化学计量学建立起来的近红外光谱数据与其样品待测性质或成分含量之间的多元回归方程[1]。故任何影响到样本光谱和数据的因属, 都有可能影响到最后模型的质量。这些因属主要是来之样品、仪器以及操作者[2]。目前, 不同学者针对影响近红外结果准确性的条件作了大量的研究, 主要包括样品粒径、装样条件[3]、分辨率[4]、温度、预处理方法[5]以及数学建模方法[6]。但对于参考数据准确性的研究较少, 目前, 主要有两种观点, 一种是近红外模型是根据参考数据建立起来的, 故"近红外光谱分析法的测定结果不如参考方法的准确"。但也有人认为近红外光谱分析模型是通过多元校正方法建立在大量数据统计的结果上, 近红外预测值不一定比参考数据差, 褚小立等通过人为增加基础数据误差的方法做了验证, 认为对于精度相对较差测试方法提供的基础数据, 通过大量样本的光谱分析和化学计量学统计处理, 近红外方法有可能得到更精确的预测结果。但这些误差是人为加上的, 不能体现出实验中的实际误差。本文以相思木聚戊糖含量为例, 用来自实验室误差的样本作为校正集建立了近红外模型, 研究了实验误差对近红外模型分析结果的影响。

1 材料与方法

1.1 试验材料

研究用相思树采自广西和福建的三个不同林场, 树种为厚荚相思、马占相思、卷荚相思、杂交相思、直干大叶以及黑木等, 树龄为五到八年。广西共取样品78个, 福建33个。将样品带回实验室用植物粉碎机粉碎成40~60目的木粉备用。

1.2 仪器与光谱数据采集

仪器: 近红外光谱仪采用德国布鲁克光谱仪器公司生产的傅里叶变换近红外光谱仪, 带有内置镀金漫反射积分球。

以相思树聚戊糖含量为例, 通过用不同精确度的数据建立的近红外模型预测性能, 讨论了不同精确度的数据对近红外模型准确性的影响。结果表明, 建模原始数据的精确度在一定程度上影响着近红外模型的预测性能, 精确度越高, 建立的模型越好。但对于精确度较小的的样品, 所建立的模型预测性能也能较好的预测未知样品。

光谱采集: 将木粉放入置于旋转台上的50 mm石英杯中, 在4 000~12 000 cm^{-1}谱区内, 扫描64次平均成为一个光谱数据, 仪器分辨率为8 cm^{-1}。

1.3 聚戊糖含量的测定

精确称取试样0.5 g(精确至0.000 1 g), 与10 g氯化钠及数枚沸石一同置于500 mL圆底烧瓶中, 再加入100 mL 12%盐酸溶液, 置于电热套上加热, 控制每10 min蒸馏出30 mL馏出液。每当馏出30 mL液体, 即补入12%盐酸溶液30

收稿日期: 2010-08-18, 修订日期: 2010-11-12
基金项目: 国家"十一五"科技支撑项目(2006BAD32B03)资助
作者简介: 姚 胜, 1984年生, 北京林业大学材料科学与技术学院博士生 e-mail: yao_sh@163.com
* 通讯联系人 e-mail: pujunwen@126.com

mL。等糠醛全部蒸馏完毕(约蒸馏出 300 mL 溶液)，将馏出液移入 500 mL 容量瓶中，用 12% 盐酸定容到刻度。最后用分光光度计测量溶液中的糠醛含量，并换算成木材聚戊糖含量。每个样品做两次平行实验，取平均值作为最后的结果。对于两次测量结果误差较大的样品，进行了第三次或者第四次试验，确保数据的准确性。

1.4　模型的建立方法

合适的光谱预处理方法和建模波段的选择是建立定量模型的基础。为了获得高质量的校正模型，本研究使用 BRUKER 公司 OPUS/QUANT6.5 光谱定量分析软件建立模型。首先将相思木样品光谱图和对应的聚戊糖含量调入 OPUS 软件，利用软件中所有预处理方法和自动优化功能对校正集样品进行内部交叉验证。以交叉验证均方根误差 RMSECV (root mean square error of cross validation) 对主成分作图，确定模型最佳主成分数。然后使用具有最小的 RMSECV 参数建立模型。

2　结果与讨论

2.1　基础数据含量及绝对误差分布

为了讨论实验误差对近红外模型准确性的影响，按每个样品两次测量绝对误差大小将 111 个样品分成三个样品集：A，B 和 C。其中样本集 C 为 11 个样品，是 111 个数据当中误差最小的前 11 个数据，用来修正其他模型。样品集 A 和样品集 B 作为校正集来建立的模型。样品集 B 为 50 个样品，是所有数据当中绝对误差较大的样品，也就是比较差的数据。对于较差的样品集 B 的样本重新测定，重新测定的结果为样品集 B′。数据样品集 A 的样品数也是 50 个，绝对误差介于样品集 B 和验证集 C 之间。样品集 A、样品集 B 和样本集 C 的统计数据见表 1。国标木材原料中聚戊糖含量的测定中规定，两次测量间绝对误差小于 0.4% 即合格。从表 1 可以看出，检验集 C 中的样品最大绝对误差为 0.1%，平均绝对误差仅有 0.06%，精确度非常高，可以认为是样品聚戊糖含量的真值。样品集 A 的平均绝对误差为 0.34%，最大绝对误差为 0.59%，基本上满足国标的要求。样品集 B 中最小的绝对误差也有 0.59%，最大绝对误差达到了 1.98%，平均绝对误差为 1.15%，比国标要求的误差大了 3 倍多，即所有值都不符合国标的要求。样品集 B′ 的误差大大减小，与样品集 A 差别不大。

2.2　模型的建立

分别利用样品集 A、样品集 B 及样品集 B′ 的谱图和对应数据建立模型，为了保证模型的可靠性，选择了具有较小

Table 1　Chemical analyzing results of calibration sets A, B, B′ and prediction set C

		N	Maximum/%	Minimum/%	Mean/%	SD/%
Hemicelloluse content	C	11	26.74	15.86	21.43	2.97
	A	50	26.08	14.78	21.37	2.02
	B	50	24.63	16.71	21.22	1.85
	B′	50	24.92	17.46	21.41	1.81
Absolute error	C	11	0.1	0.02	0.06	0.03
	A	50	0.59	0.11	0.34	0.15
	B	50	1.98	0.59	1.15	0.41
	B′	50	0.67	0.01	0.27	0.18

Table 2　Results from cross validation

Models	Pre-treatment methods	Wavenumber/cm^{-1}	RPD	RMSECV	R^2
A1	1stDer+ SLS	7 502.1~5 446.3, 4 601.6~4 246.7	3.97	0.502	0.937
A2	1stDer+ MSC	7 502.1~6 098.1, 5 450.1~4 246.7	3.90	0.512	0.934
A3	1stDer+ SLS	7 502.1~6 098.1, 5 450.1~4 597.7	3.87	0.515	0.933
A4	1stDer+ SLS	7 502.1~5 450.1, 4 601.6~4 424.1	3.87	0.517	0.933
A5	SLS	7 502.1~6 098.1, 4 601.6~4 246.7	3.81	0.524	0.931
A6	1stDer+ VecNor	7 502.1~4 246.7	3.81	0.524	0.931
B1	VecNor	7 502.1~5 446.3, 4 601.6~4 246.7	2.86	0.642	0.877
B2	None	7 502.1~6 098.1, 4 601.6~4 246.7	2.85	0.643	0.877
B3	ConOff	7 502.1~5 446.3, 4 601.6~4 246.7	2.80	0.656	0.872
B4	ConOff	7 502.1~6 098.1, 4 601.6~4 246.7	2.74	0.669	0.867
B5	1stDer	7 502.1~5 446.3, 4 601.6~4 246.7	2.73	0.672	0.866
B6	ConOff	6 102~5 446.3, 4 601.6~4 246.7	2.71	0.677	0.864
B′1	1stDer + MSC	7 502.1~6 098.1, 5 450.1~4 246.7	3.56	0.487	0.921
B′2	None	7 502.1~6 098.1, 4 601.6~4 246.7	3.63	0.499	0.924
B′3	SLS	7 502.1~5 446.3, 4 601.6~4 246.7	3.51	0.502	0.918
B′4	1stDer+ MSC	7 502.1~4 597.7	3.26	0.549	0.906
B′5	MSC	7 502.1~6 098.1, 4 601.6~4 246.7	3.25	0.551	0.905
B′6	VecNor	7 502.1~5 446.3	3.20	0.560	0.902

1stDer: First derivative, VecNor: vector normalization, SLS: Straight line subtraction, ConOff: constant offset

RMSECV 值的前六个参数建立了模型，对应的近红外光谱模型分别为 A1-A6，B1-B6 和 B′1-B′6。建立模型的预处理方法，波长的选择和模型的 RPD，RMSECV 和 R^2 见表 2。从建模条件来看，用样品集 A 建立的模型预处理方法主要以一阶导数在加减去一条直线或其他预处理方法为主，样品集 B 建立模型时更多用到消除常数偏移量，样品集 B′建模预处理方法使用一阶导数和多元散射校正较多。不同模型的最佳波长段大致相同，均在 7 502.1~4 246.7 cm^{-1}之间。从模型结果可以看出，用样品集 A 建立的近红外模型 A1-A6 差异不大，用不同的预处理方法和选择不同的谱区建立的模型均很好，RMSECV 最低可达到 0.502，R^2 最大为 0.937。用样品集 B 建立的近红外模型 B1-B6 结果稍差，RMSECV 相对较高，最小也有 0.642，R^2 最大也仅有 0.877。用样品集 B′建立的近红外模型 B′1-B′6 结果 RMSECV 相对较低，最小只有 0.487，但 R^2 最大为 0.924，与模型 A1-A6 相差不大。说明试验误差越小，所建立的模型越好，但实验误差很大时，模型质量虽然下降，但 RMSECV 和 R^2 还在合理的范围。没有因为建模数据误差的增大而变的很差。

2.3 模型的验证

分别用样品集 A 和样品集 B 当做验证集，来检验模型 B1-B6 和 B′1-B′6 以及模型 A1-A6。即把样品集 B 和 B′的样品当做验证集来检验模型 A1-A6，并反过来用样品集 A 的样品当做验证集来检验模型 B1-B6 和 B′1-B′6，检验结果见表 3 和表 4。可以看出，用样本集 B 的样品来检验样本集 A 建立的近红外模型时，最优模型 A6 的 RMSEP 为 0.699，R^2 为 0.886，结果较差；用样本集 B′的样品来检验样本集 A 建立的近红外模型时，最优模型 A6 的 RMSEP 为 0.538，R^2 为 0.919，化学测定值和近红外模型预测值更一致。这也说明经修正后的样本集 B′的基础数据比样本集 B 的基础数据更为准确。当用样本集 A 的样本来检验样本集 B 和 B′建立的模型时，最优模型 B6 的 RMSEP 为 0.678，R^2 为 0.915；模型 B′3 的 RMSEP 为 0.659，R^2 为 0.905，稍好于模型 B6 的预测结果。对比模型 B1-B6 和 B′1-B′6，可以看出，虽然模型 B′1-B′6 较好，但预测未知样品性能和模型 B1-B6 相差不大。说明提高数据的准确度不能大幅度的提高模型的预测性能。

Table 3 Results of calibration models A1-A6 for prediction validation set B and B′

Models	Validation set	RMSEP	Bias	RPD	R^2	Slope	Intercept
A1	B	0.777	-0.241	2.48	0.842	0.902	2.323
A2		0.711	-0.28	2.80	0.877	0.934	1.679
A3		1.12	-0.409	1.76	0.712	0.867	3.234
A4		0.878	-0.285	2.20	0.799	0.863	3.191
A5		0.997	-0.397	2.00	0.772	0.899	2.542
A6		0.699	-0.299	2.90	0.886	0.952	1.312
A1	B′	0.679	-0.049 5	2.64	0.863	0.975	1.469
A2		0.544	-0.088 3	3.33	0.914	0.913	0.617
A3		0.969	-0.218	1.90	0.755	0.894	2.090
A4		0.781	-0.093 8	2.31	0.820	0.940	2.372
A5		0.856	-0.206	2.16	0.807	0.992	1.496
A6		0.538	-0.107	3.40	0.919	0.975	0.285

Table 4 Results of calibration models B1-B6 for prediction validation set A

Models	RMSEP	Bias	RPD	R^2	Slope	Intercept
B1	0.713	0.327	3.15	0.900	0.929	1.189
B2	0.825	0.375	2.71	0.864	0.868	2.446
B3	0.764	0.334	2.90	0.882	0.912	1.555
B4	0.850	0.399	2.66	0.859	0.881	2.147
B5	0.783	0.409	2.99	0.889	0.919	1.328
B6	0.678	0.330	3.37	0.915	0.966	0.406
B′1	0.735	0.141	2.77	0.869	0.869	2.665
B′2	0.728	0.196	2.85	0.877	0.864	2.721
B′3	0.659	0.226	3.22	0.905	0.939	1.082
B′4	0.756	0.289	2.86	0.879	0.916	1.512
B′5	0.771	0.242	2.73	0.865	0.873	2.472
B′6	0.665	0.260	3.26	0.908	0.952	0.774

从上文看以看出，提高样品集 B 中聚戊糖含量的准确度以后，模型预测性能没有显著的提高，为了提高模型的准确度，用另外 10 个未参加建模的数据修正了样品集 B 和 B′建立的模型，结果如表 5。可以看出，加入更多的样品后，模型预测性能大幅提高。说明样本实验数据的准确性不是很重要，在样本不是很多的时候，应该努力增加建模样本的数量，以提高模型的预测性能，而不是一味的提高样品基础数据的准确性。

Table 5 Results of calibration models B1-B6 modified with another 10 samples for prediction validation set A

Models	RMSEP	Bias	RPD	R^2	Slope	Intercept
B1x	0.580	0.130	3.53	0.921	0.961	0.705
B2x	0.608	0.174	3.42	0.916	0.955	0.785
B3x	0.553	0.138	3.72	0.931	0.984	0.211
B4x	0.642	0.153	3.20	0.904	0.948	0.949
B5x	0.684	0.139	2.98	0.896	0.984	0.194
B6x	0.627	0.283	3.56	0.924	0.972	0.325
B′1x	0.532	0.093 8	3.81	0.933	0.978	0.374
B′2x	0.573	0.083 5	3.52	0.927	1.014	-0.383
B′3x	0.507	0.089 9	4.00	0.938	0.934	1.310
B′4x	0.545	0.089 7	3.71	0.929	0.970	0.541
B′5x	0.514	0.090 3	3.94	0.935	0.942	1.140
B′6x	0.518	0.106	3.93	0.937	0.976	0.396

3 结 论

样品集样品基础数据实验误差的大小影响着模型的质量以及预测的准确性，样品基础数据两次测量绝对误差越小，所建立的模型越好，对应的预测结果也越好，但不能大幅度提高模型的预测性能。对于参考方法不是很准确的样品，采用近红外技术可以得到更准确的结果。在样品量不是很大的情况下，加入更多的建模样品比增加基础数据的准确性更为重要。

References

[1] LU Wan2zhen, YUAN Hong2fu, XU Guang2tong(陆婉珍, 袁洪福, 徐广通). Modern Near Infrared Spectroscopy Analytical Technology (现代近红外光谱分析技术). Beijing: China Petrochemical Press(北京: 中国石化出版社), 2000.
[2] LI Yong, WEI Yi2min, WANG Feng(李 勇, 魏益民, 王 锋). Acta Agriculturae Nucleatae Sinica(核农学报), 2005, 19(3): 236.
[3] WU Jing2zhu, WANG Yi2ming, ZHANG Xiao2chao(吴静珠, 王一鸣, 张小超). Modern Scientific Instruments(现代科学仪器), 2006, (1): 69.
[4] WANG Yi2bing, WANG Hong2yu, ZHAI Hong2ju(王一兵, 王红宇, 翟宏菊). Chinese Journal of Analytical Chemistry(分析化学), 2006, 34(5): 699.
[5] XIA Ba2yang, REN Qian(夏柏杨, 任 芊). Chinese Journal of Spectroscopy Laboratory(光谱实验室), 2005, 22(3): 629.
[6] FU Xia2ping, YING Yi2bin, LU Hu2shan(傅霞萍, 应义斌, 陆辉山). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2007, 27(5): 911.

The Influence of Reference Data Noise on the NIR Prediction Results

YAO Sheng, WU Guo2feng, ZHOU Shu2ke, JIANG Yi2fei, JIN Xiao2juan, ZHAO Qiang, PU Jun2wen*
College of Material Science and Technology, Beijing Forestry University, Beijing 100083, China

Abstract This article used hemicelluloses content in acacia spp. wood as a case study to demonstrate the influence of noise in the reference data on the results of NIR calibration model. The results indicated that the accuracy of NIR calibration model was affected by the reference data noise. The less noisy data was used in calibration model, the better result could be obtained. But when the noise was larger, NIR calibration model which was built by using regression mathematics methods can perform better than using primary reference data.

Keywords Near infrared; Noise; Hemicelluloses

(Received Aug. 18, 2010; accepted Nov. 12, 2010)

* Corresponding author