烟碱含量近红外光谱预测模型的评价

张优茂，沈光林，孔浩辉，程志颖，张心颖
广东中烟工业公司技术中心，广州荔湾区中山七路 333 号 510145

摘 要：以烟碱为例，建立了近红外光谱的预测模型，采用回归分析法和假设检验法对该模型的预测效果进行了系统的评价。结果表明，该模型对烟草样品烟碱含量的预测效果良好，模型预测值与化学测定值之间不存在显著性差异。

关键词：近红外光谱 预测模型 烟草 烟碱 评价
中图分类号：TS411.1 文献标识码：A 文章编号：1004－5708(2007)05－0006－04

Evaluation on near infrared prediction model of nicotine in tobacco

ZHANG You-mao，SHEN Guang-lin，KONG Hao-hui，CHENG Zhi-ying，ZHANG Xin-ying
Technology Center of China Tobacco Guangdong Industrial Corporation，Guangzhou 510145，China

Abstract：Near infrared prediction model was established to predict content of nicotine in tobacco. Evaluation on the prediction results was made by regression analysis and hypothesis testing. Results indicated that the prediction performance of this model to nicotine was satisfactory，and there was no significant difference between prediction value and measured value.

Key words：near infrared；prediction model；tobacco；nicotine；evaluation

1961 年，Crowell[1] 首次应用近红外(Near Infrared，NIR)技术测定焦油中的水分，此后，McClure 等[2-3]、Hamid 等[4]先后采用近红外技术测定烟草中的还原糖、植物碱、多酚及无机元素等，Luzio 等[5] 及 Williamson 等[6]采用近红外技术对烟气成分进行分析，近红外技术在烟草分析测试中的应用也越来越广泛。在我国，最早的公开文献报道是王文真等[7]于 1995 年采用近红外技术测定烟草中的总氮含量。目前，近红外技术已基本普及了每一个烟草研究单位，现已成功建立了烟草中总糖、还原糖、总氮、烟碱、硫、磷、氯、钾、钙、镁、挥发酸、挥发碱、淀粉、石油醚等[8-12]化学成分的近红外光谱预测模型，并取得了较好的应用。本文以烟碱为例，建立了近红外光谱的定量预测模型，以回归分析法和假设检验法对该预测模型进行了内部检验和外部检验，并对其预测效果进行了评价。

作者简介：张优茂(1980—)，男，硕士，广东中烟工业公司助理工程师，主要从事近红外光谱仪及自动电位滴定仪的分析研究。
Tel：020-81812508-782，E-mail：zhyoum@163.com

收稿日期：2007-01-26

1 材料与方法

1.1 仪器与设备

傅立叶变换 MPA 型近红外光谱分析仪(内置镀金漫反射积分球)及 OPUS 定量分析软件包(德国 Brucker 公司)；SKALAR 连续流动分析仪(荷兰 SKALAR 公司)；旋风精密粉碎机(Foss-1093，Sweden)；天平、烧杯等常规分析仪器。

1.2 样品

2005 年云南、贵州、四川、广东等 10 个省市的烤烟 246 种，其中 221 种作为建模样品，25 种作为检验样品。

1.3 试验方法

将 246 种烟叶样品用同一切丝机切成烟丝，在 40 ℃下烘干至一定程度(含水量 5.0% ~ 7.0%)，再用旋风精密粉碎机充分粉碎、研磨均匀，得到过 40 目筛的粉末样品。

取 5 g 左右烟末样品放入石英旋转样品杯，加上压样器，放在旋转台内进行光谱扫描。扫描条件为：分辨率：8 cm^{-1}，扫描次数：64 次，扫描范围：4000 ~ 12000 cm^{-1}，检测器：RT-PbS，扫描频率：10 KHz。

采用 SKALAR 连续流动分析仪测定样品的烟碱含量，称为化学测定值(下同)。

近红外光谱预测模型采用 OPUS 定量分析软件建立。选择 4000～9000 cm^{-1}范围的光谱数据，采用一阶导数、二阶导数、矢量归一法、消除常数偏移量、最大最小归一法、多元散射校正等方法对光谱数据进行前处理，前处理后与样品烟碱含量的化学测定值进行关联，采用偏最小二乘法建立近红外光谱数据与样品烟碱含量的数学模型，然后用该模型预测样品的烟碱含量，称为模型预测值(下同)。

2　结果与讨论

当一种新的方法提出后，常用做法是通过对已知样的测定，与标准测定方法进行比较，以确证该方法的可行性[13]。模型内部检验，是指用近红外光谱模型预测建模所用样品的烟碱含量，再与烟碱的化学测定值进行比较；而模型外部检验，是指用近红外光谱模型预测未知样品(非建模所用样品)的烟碱含量，再与样品的化学测定值进行比较。

2.1　模型内部检验

2.1.1　回归分析检验

通过对近红外光谱数据的一系列前处理，剔除 30 个异常值，对剩余的 191 个样品采用偏最小二乘法建立校正模型。在理论上，样品烟碱含量的化学测定值 x 与近红外光谱模型预测值 y 应该相等，即有 y = x。为了考察样品烟碱含量的化学测定值与其近红外光谱模型预测值之间的差异性，采用回归分析法，以化学测定值为 x 轴，近红外光谱模型预测值为 y 轴，建立两者之间的示意图，如图 1 所示。根据回归结果和相关数据，可得：

截距 a = 0.0180，斜率 b = 0.9933，r = 0.9949，n = 191

则 $F = \dfrac{r^2(n-2)}{1-r^2} = 18340.47$

查 F 分布表可知 $F_{0.05}(1,189) = 3.90$，则 $F > F_{0.05}(1,189)$，这表明线性回归效果显著，即烟碱含量化学测定值与模型预测值之间线性相关关系显著。

一元回归分析标准偏差

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = 0.0730$$

截距 a 的标准偏差

$$s_a = s\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2}} = 0.0195$$

斜率 b 的标准偏差

$$s_b = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = 0.00735$$

当置信度 $\alpha = 0.95$，自由度 $f = n - 2 = 189$ 时，查 t 分布表得 $t(0.95,189) = 1.97$，则截距和斜率的置信区间分别为：

$a = a \pm ts_a = 0.0180 \pm 0.0384$，$b = b \pm ts_b = 0.9933 \pm 0.0145$

由此可见，截距 a 和斜率 b 与理论值 0 和 1 差异很小。

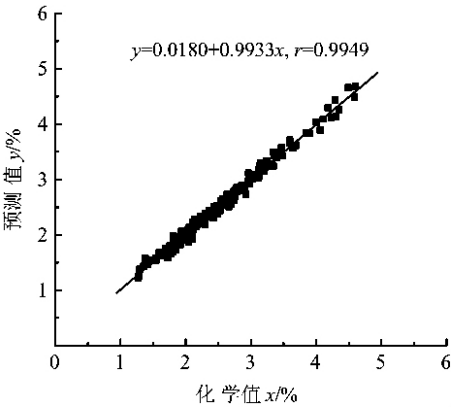

图 1　烟碱含量模型预测值与化学测定值关系图

同时，通过计算可得到该模型的交叉验证均方根(root of mean squared error of cross－validation，RMSECV)为：

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n-1}} = 0.0729$$

上述数据表明，烟碱含量的化学测定值与模型的预测值之间存在非常好的线性相关关系，截距 a = 0.0180，斜率 b = 0.9933，表明两者之间非常接近，即化学测定值与模型预测值之间的差异很小。

2.1.2　假设检验

假设检验是采用统计学方法，通过对样品烟碱含量化学测定值与模型预测值 2 组数据的比较，以考察

2 组数据之间是否存在显著性差异。

根据统计学中的成对结果的 t 检验法(paired t-test)[13],选取 $z_i = y_i - x_i$ 建立一组新的数据,则 z 的数学期望值 $\mu = 0$,再用 t 检验法进行检验,根据获得的实验数据可计算得:

n = 191,自由度 f = n − 1 = 190

算术平均值 $\bar{z} = \frac{\sum_{i=1}^{n}(y_i - x_i)}{n} = -0.00105$,标准偏差 $s = \sqrt{\frac{\sum_{i=1}^{n}(z_i - \bar{z})^2}{n-1}} = 0.0729$

则 $t = (\bar{z} - \mu)\frac{\sqrt{n}}{s} = -0.1991$

查 t 分布表得 $t(0.05, 190) \approx 1.97$,$|t| = 0.1991 < t(0.05,190)$,这表明 2 组数据不存在显著性差异,即模型内样品烟碱含量的化学测定值与近红外光谱模型预测值之间不存在显著性差异。

2.2 模型的外部检验

模型的外部检验是采用已建好的近红外光谱预测模型对未知样品的烟碱含量进行预测,并与化学测定值进行比较。实验选取 25 种不同产地的烤烟,先用近红外光谱仪进行扫描,根据获得的近红外光谱数据,采用近红外预测模型对样品的烟碱含量进行预测,并与化学测定值进行比较,实验结果如表 1 所示。

从表 1 可知,相对偏差超过 10.0% 的样品仅有 2 个,比率为 8.0%,相对偏差在 5.0% ~ 10.0% 的样品为 4 个,比率为 16%,而相对偏差低于 5.0% 为 19 个,比率为 76%,总体平均相对偏差为 3.586%,近红外模型的预测标准差(root of mean squared error of prediction, RMSEP)为 0.162。

理论上 y = x,则相对偏差 η 的数学期望值为 0,采用 Grubbs 准则,设残差 $v_i = \eta_i - \bar{\eta}$,当 $|v_i| > \lambda(\alpha,n)s$ 时(取 $\alpha = 0.05$,λ 查 Grubbs 系数表可得,s 为 η 的标准偏差),即可将该数值当作异常值予以剔除。通过 2 次使用 Grubbs 准则,剔除相对偏差为 13.059% 和 10.176% 2 个异常值,对剩余的 23 个样品烟碱含量的化学测定值与模型预测值进行回归分析,结果如图 2 所示,可得:

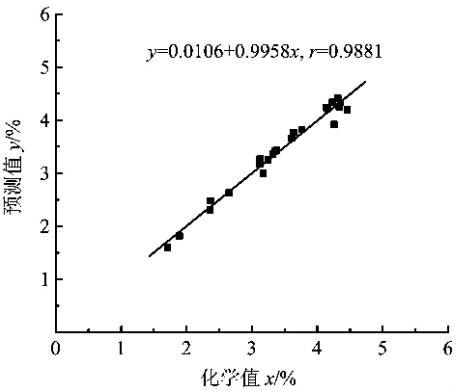

图 2 烟碱含量模型预测值与化学值关系图

表 1 检验样品烟碱化学测定值与近红外模型预测值的比较

编号	化学测定值 x/%	模型预测值 y/%	偏差 /%	相对偏差 η/%
1	3.606	3.657	0.051	1.414
2	3.118	3.170	0.052	1.668
3	3.128	3.267	0.139	4.444
4	3.321	3.360	0.039	1.174
5	4.226	4.334	0.108	2.556
6	2.651	2.633	0.018	0.679
7	3.638	3.770	0.132	3.628
8	3.377	3.433	0.056	1.658
9	3.175	2.996	0.179	5.638
10	3.246	3.247	0.001	0.031
11	3.351	3.415	0.064	1.910
12	4.354	4.327	0.027	0.620
13	3.764	3.823	0.059	1.567
14	4.132	4.238	0.106	2.565
15	4.337	4.251	0.086	1.983
16	4.314	4.411	0.097	2.248
17	4.260	3.916	0.344	8.075
18	4.457	4.197	0.260	5.834
19	3.844	4.346	0.502	13.059
20	1.892	1.814	0.078	4.123
21	1.651	1.483	0.168	10.176
22	1.711	1.597	0.114	6.663
23	4.154	4.214	0.060	1.444
24	2.360	2.310	0.050	2.119
25	2.373	2.477	0.104	4.383
平均值			0.116	3.586
RMSEP				0.162

y = 0.0106 + 0.9958x,r = 0.9881,s = 0.1288,s_a = 0.1190,s_b = 0.0338

其中 a = 0.0106 ∈ (0.0180 ± 0.0384),

$b = 0.9958 \in (0.9933 \pm 0.0145)$，

则　$F = \frac{r^2(n-2)}{1-r^2} = 866.5$

查 F 分布表可知 $F_{0.05}(1,21) = 4.32$，即 $F > F_{0.05}(1,21)$，这表明烟碱含量化学测定值与模型预测值之间线性相关关系显著。

当置信度为 95% 时，自由度 $f = n - 2 = 21$，t 值为 2.08，则截距与斜率的置信区间为

$a = 0.0106 \pm 0.2475$，$b = 0.9958 \pm 0.0703$

由此可见，截距 a 和斜率 b 与理论值 0 和 1 差异很小，且在预测模型回归方程的截距与斜率范围内。

根据成对结果的 t 检验法（paired t-test），选取 $z_i = y_i - x_i$ 建立一组新的数据，z 的数学期望值 $\mu = 0$，再用 t 检验法进行检验，根据获得的实验数据可得：$n = 25$，自由度 $f = 24$，算术平均值 $\bar{z} = -0.010$，标准偏差 $s = 0.162$。

则　$t = (\bar{z} - \mu)\frac{\sqrt{n}}{s} = -0.309$

查 t 分布表得 $t_{(0.05,24)} = 2.06$，$|t| = 0.309 < t_{(0.05,24)}$，这表明 2 组数据在显著性水平 $\alpha = 0.05$ 的条件下不存在显著性差异，即化学测定法与模型预测法测得的烟碱含量没有显著性差异。

模型外部检验结果表明，近红外光谱模型对样品烟碱含量的预测效果良好，模型预测值与化学测定值非常接近，两者之间不存在显著性差异。

3　结论

（1）近红外光谱模型检验结果表明，该模型烟碱含量的预测值与其化学测定值之间的数学关系为 $y = 0.0180 + 0.9933x$，相关系数 $r = 0.9949$，交叉验证均方根为 0.0729，模型预测值与其化学测定值之间不存在显著性差异。

（2）模型外部检验结果表明，该模型对样品烟碱含量的预测效果良好，模型预测值与化学测定值差异很小，平均相对偏差为 3.586%，预测标准差为 0.162，两者之间不存在显著性差异。

回归分析法和假设检验法为近红外光谱预测模型的评价提供了一个量化的指标，解决了以往单纯依靠决定系数 R^2 及预测标准差 RMSEP 等评价方法的单一性，使评价结果更具说服力，可用于烟草其它化学成分的近红外光谱预测模型的评价中。

参考文献

[1] 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，2005：470.

[2] McClure W F，Norris K H，Weeks W W. Rapid spectrophotometric analysis of the chemical composition of tobacco. Part 1：Total reducing sugars[J]. Beitr Tabakforsch，1977(9)：13-17.

[3] McClure W F，Williamson R E，Hamid A. Measurement of polyphenols in tobacco by computerized near infrared spectrophotometry[J]. Tob Chem Res Conf，1978(32)：27-28.

[4] Hamid A，McClure W F，Weeks W W. Rapid spectrophotometric analysis of the chemical composition of tobacco. Part2：Total alkaloid[J]. Beitr Tabakforsch，1978(9)：267-274.

[5] Di Luzio C，Morzilli S，Cardinale E. Rapid Near Infrared reflectance analysis of mainstream smoke collected on cambridge filter pades [J]. Beitrage zur Tabakforschung International，1995，16(4)：171-184.

[6] Williamson R E，Chaplin J F，McClure W F. Near infrared spectrophotometry of tobacco leaf for estimating tar yield of smoke [C]//Tobacco Chem Res Conf . Knoxville Tenn，1986，40：26-27.

[7] 王文真，张怀宝. 利用 IA450 近红外分析仪快速测定烟草中的总氮含量[J]. 仪器仪表与分析监测，1995(20)：53-55.

[8] 何智慧，练文柳，吴名剑，等. 声光可调—近红外光谱技术分析烟草主要化学成分[J]. 分析化学，2006，34(5)：702-704.

[9] 王家俊，罗丽萍，李辉，等. FT-NIR 光谱法同时测定烟草根、茎、叶中的氮、磷、氯和钾[J]. 烟草科技，2004(12)：24-27.

[10] 付秋娟，王树声，窦玉青，等. 近红外定量分析青烟叶中 K、Ca、Mg 含量的研究[J]. 中国烟草学报，2006，12(2)：17-19.

[11] 蒋锦锋，赵明月. 近红外光谱法快速测定烟草中的总挥发酸与总挥发碱[J]. 烟草科技，2006(3)：33-37.

[12] 邱军，王允白，张怀宝，等. 烟草中淀粉、石油醚提取物的近红外光谱分析模型研究[J]. 分析化学，2006，34(4)：588.

[13] 许禄，邵学广. 化学计量学方法[M]. 2 版，北京：科学出版社，2004：14-32.