第31卷 第4期 华 中 农 业 大 学 学 报 Vol. 31 No. 4
2012年 8月 Journal of Huazhong Agricultural University Aug. 2012,518～523

基于可见-近红外光谱的砂糖橘总酸无损检测

代 芬 洪添胜 罗 霞 洪 涯 李 岩

华南农业大学工程学院/南方农业机械与装备关键技术省部共建教育部重点实验室/
国家柑橘产业技术体系机械研究室,广州 510642

摘要 以砂糖橘为对象,建立基于可见-近红外光谱的砂糖橘总酸含量的无损检测方法。试验采集170个完整砂糖橘的500～2 500 nm漫反射光谱,然后采用滴定法测定总酸含量。采用Sym8小波变换对光谱进行去噪预处理,并采用连续投影算法(successive projections algorithm,SPA)结合间隔偏最小二乘法(interval partial least squares,iPLS)优选波长,最终建立BPNN和偏最小二乘法(partial least squares method,PLS)总酸预测模型。结果表明:砂糖橘光谱的小波去噪方法产生的信噪比均值SNR=175.291 1,去噪信号与原始信号间的均方根误差均值RMSE=0.000 13,性能优于常规去噪方法。SPA与iPLS相结合构成的反向偏最小二乘法(backward interval partial least squares,BiPLS)_SPA波长选择法能将光谱变量从2 001个压缩到14个,能简化模型并提高建模精度和稳定性。BPNN模型具有更好的非线性映射能力,基于这14个变量的BPNN总酸预测模型的预测相关系数R_p=0.867,预测均方根误差RMSEP=0.061 6,性能优于线性的PLS模型。

关键词 近红外光谱;小波去噪;连续投影算法;砂糖橘;总酸含量;无损检测

中图分类号 S 24 **文献标识码** A **文章编号** 1000-2421(2012)04-0518-06

砂糖橘又名“十月橘”,口感细腻,极甜无渣,是广东省特色水果之一,其总酸含量是衡量砂糖橘品质的主要理化指标,对于砂糖橘的成熟度、口味和营养都有重要的影响[1]。传统的实验室滴定法测量总酸含量,不仅耗时繁琐、具有破坏性,而且试剂和人工成本很高,因此寻求快速无损的检测方法显得十分必要。

近红外光谱分析具有样品处理简单、分析速度快、可以同时测定多种组分、非破坏性和无污染性等优点,已经被越来越多地应用于食品、石油化工、制药等领域[2-3]。国内外很多学者利用近红外光谱技术对苹果、梨、番茄等水果内部品质、绿茶品质以及土壤有机质含量等进行了评价研究[4-8]。

在近红外光谱技术的研究中,国内外学者主要集中于对近红外光谱预处理、波长选择和建模方法等[9-11]三方面研究。砂糖橘成份复杂,其样品的光谱信号严重重叠,吸收较弱,易受各种噪声干扰,这些噪声信号的存在会影响最终分析结果的准确性,必须进行预处理。本研究采用小波变换对原始光谱进行去噪预处理,相对于常规去噪预处理能提高信噪比。进一步进行波长选择,这样一方面可以简化模型,更主要的是由于不相关或非线性变量的剔除,可以得到预测能力更强的校正模型。笔者采用新颖的变量提取方法——连续投影算法(successive projections algorithm,SPA)结合间隔偏最小二乘法(interval partial least squares,iPLS),能将光谱变量压缩到原来的0.70%,最终提取出与总酸测量相关的特征波长,显著提高模型预测精度和稳健性,为实现砂糖橘品质在线无损检测系统产业化奠定了理论基础。

1 材料与方法

1.1 试验材料

试验使用的砂糖橘样本产自广东省四会市,将其洗净、擦干,最后共得到完好的样本170个。

1.2 试验设备

试验使用FieldSpec 3光谱仪(美国,ASD公

收稿日期:2011-07-20

基金项目:国家自然科学基金项目(30871450)、现代农业(柑橘)产业技术体系建设专项(农科教发[2011]3号)和华南农业大学校长基金项目(2009K005)

代 芬,博士,讲师. 研究方向:基于光谱分析的农产品无损检测. E-mail:sunflower@scau.edu.cn

通讯作者:洪添胜,博士,教授. 研究方向:机电一体化和信息技术在农业中的应用. E-mail:tshong@scau.edu.cn

司），其光谱分辨率为 3 nm（350～1 000 nm）和 10 nm（1 000～2 500 nm），光谱测量范围 350～2 500 nm，采样间隔为 1.4 nm（350～1 000 nm）和 2 nm（1 000～2 500 nm），扫描次数 10 次，数据间隔 1 nm，裸光纤探头前视场角为 25°，并配有自带光源的接触式反射探头，光源是与光谱仪配套的 14.5 V 卤素灯。光谱数据以 ASCII 码格式导出进行处理，分析软件采用 Matlab、unscrambler 9.8 和 ASD View Spec Pro。

1.3　检测方法

1）光谱采集。研究表明，相对于较远距离，近距离测量得到的定量分析模型效果较好[6]，本试验采用反射探头接触式采集砂糖橘表面反射光谱。在每个砂糖橘的赤道部位取 3 点进行测量，每点相隔 120°。将每个点扫描 30 次的 30 个光谱数据进行平均，作为该点的光谱，然后将 3 个点的光谱平均作为整个砂糖橘的漫反射光谱。共采集 170 个完整果实的漫反射光谱。

2）总酸测定。实验室总酸含量的测定依据 GB/T 12293－1990《水果、蔬菜制品可滴定酸度的测定》，采用 NaOH 中和滴定法测定砂糖橘总酸度（滴定酸度），测定结果见表 1。

表 1　实验室测定砂糖橘样本总酸含量

Table 1　The total acidity measured by titration method

样本 Sample	样本数目 Number of samples	总酸/% Total acidity 范围 Range	平均 Mean	标准偏差 Standard deviation
校正集 Calibration set	136	0.120～0.700	0.340	0.123 00
预测集 Prediction set	34	0.180～0.560	0.301	0.088 08

1.4　小波变换消噪原理

设信号在某一尺度 2^L 上的离散逼近 $f(n)$ 被加性噪声 $w(n)$ 污染，观测数据 $x(n)=f(n)+w(n)$。将 $x(n)$ 在下述正交规范基上分解：

$$B=[\{\Phi_{J,m}(n)\}_{m\in Z},\{\Psi_{j,m}(n)\}_{L<j\leqslant J,m\in Z}]$$

其中，$\{\Phi_{J,m}(n)\}_{m\in Z}$ 是尺度函数 $\Phi(n)$ 经过伸缩和平移之后得到的函数族；$\{\Psi_{j,m}(n)\}_{L<j\leqslant J,m\in Z}$ 是正交小波函数 $\Psi(n)$ 经过二进伸缩和平移之后得到的函数族。小波去噪是对分解系数取阈值后进行重构，即对 f 的估计可写成

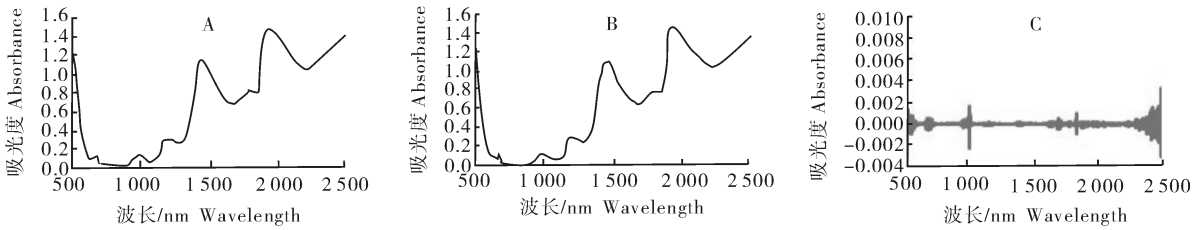
$$\tilde{F}=\sum_{j=L+1}^{J}\sum_{m}\rho_T(\langle x,\Psi_{j,m}\rangle)\Psi_{j,m}+\sum_{m}\rho_T(\langle x,\Phi_{J,m}\rangle)\Phi_{J,m}$$

其中，ρ_T 表示对分解系数取硬阈值或软阈值。

通常而言，噪声能量一般集中在高频部分，所以应当对小波系数取阈值，当阈值幅度以很大的概率高于噪声幅度时，将低于阈值幅度的小波系数置零，这样就能在很大程度上滤除噪声，信号能量主要集中在低频部分，所以通常应当保留逼近系数。砂糖橘原始光谱、小波消噪光谱如图 1 所示。由于两者比较相似，为了更好地观察去噪效果，将二者之差也绘制在图中，由图 1 可知，在光谱的两端噪声较大，这是由于接近测量仪器的线性测量边界，故而有较大噪声。同时在1 000 nm 和 1 830 nm 处噪声幅度较大，这是由于 FieldSpec 3 光谱仪本身预热不足等原因容易在 1 000 nm 和 1 830 nm 处产生断层，从而产生仪器噪声。

1.5　连续投影算法

连续投影算法（SPA）是一种比较新颖的波长选取方法，能寻找含有最低限度冗余信息的变量组，尽可能消除众多波长变量之间的共线性影响。连续投影算法步骤如下：



A：原始光谱 Original spectrum；B：小波去噪光谱 De-noising spectrum by wavelet；C：噪声光谱 Noise spectrum.

图 1　砂糖橘原始光谱（A）、小波去噪光谱（B）以及噪声光谱（C）

Fig. 1　Original spectrum（A），de-noising spectrum（B）by wavelet and noise spectrum（C）

1)初始化:$n=1$(第1次迭代),在光谱矩阵中任选一列向量 x_j,记为 $x_{k(0)}$。

2)集合 S 定义为:$S=\{j,1\leqslant j\leqslant K,j$ 不属于 $\{k(0),\cdots,k(n-1)\}\}$,即还没有被选进波长链的列向量,分别计算 x_j 对 S 中向量的投影向量:

$$Px_j=x_j-(x_j^T-x_{k(n-1)})x_{k(n-1)}$$
$$(x_{k(n-1)}^T-x_{k(n-1)})^{-1}$$

3)记录最大投影的序号为:$k(n)=\arg(\max\|Px_j\|,j\in S)$。

4)将最大的投影作为下轮的投影向量 $x_j=Px_j,j\in S$。

5)$n=n+1$,如果 $n<N$,回到第2)步循环计算。

这样得到 $N\times K$ 对波长组合,对每对 $x_{k(0)}$ 和 N 所决定的组合分别建立多元回归模型,使用交互验证均方根误差(root mean square error of cross-validation,RMSECV)来决定所建模型的优劣。选出最小 RMSECV 所对应的 $x_{k(0)}$ 和 N 即为最佳波长组合[11]。

1.6 间隔偏最小二乘法的变量区间选择算法

间隔偏最小二乘算法(iPLS)将全光谱区间分割成若干个等宽的子区间,并基于每个子区间的光谱进行 PLS 建模,然后通过比较每个子区间建立模型的交互验证均方根误差(RMSECV),选择得到最小 RMSECV 的子区间作为最优。反向间隔偏最小二乘算法(backward interval PLS,BiPLS)则在 iPLS 的基础之上,从全波段光谱区间开始,每次将1个区间去除,直到剩下最后1个区间为止。每次去除的区间是当该区间被去除之后建立的模型所得到的 RMSECV 最低。组合间隔偏最小二乘算法(interval iPLS,siPLS)是计算由 iPLS 得到的所有每2、3和4个子区间组合之后的光谱区间建立的 PLS 模型,通过比较得到 RMSECV 最小的一个模型。

2 结果与分析

2.1 基于小波变换的砂糖橘光谱去噪

为了去除光谱信号中的噪声,分别采用移动平均平滑法、Savitzky2Golay 卷积平滑法和小波变换法对原始光谱进行去噪预处理。通过多次前期比较,决定采用 sym8 小波的3层分解,对细节系数选用 sure 阈值模式。对于去噪效果的评定,一般采用信噪比作为标准。将原始光谱信号定义为 $a(n)$,消噪后的光谱定义为 $\hat{a}(n)$,则消噪后估计信号的信噪比为:

$$\mathrm{SNR}=20\lg(\mathrm{norm}(a(n))/\mathrm{norm}(a(n)-\hat{a}(n)))$$

式中 $\mathrm{norm}(a(n))$ 是 $a(n)$ 的欧几里德长度。原始信号与消噪信号之间的均方根误差(RMSE)定义为:

$$\mathrm{RMSE}=\sqrt{\frac{1}{m}\sum_m(a(n)-\hat{a}(n))^2}$$

信号的信噪比越高,噪声消除越充分,原始信号与消噪信号的均方根误差越小,则消噪信号就越接近于原始信号,消噪效果越好[12-13]。表2列出了3种消噪方法的 SNR 和 RMSE 值。

表2 3种消噪方法的 SNR 和 RMSE 对比

Table 2 The SNR and RMSE of 3 de-noising methods

项目 Item	小波去噪(sym8,3层) Wavelet de-noising(sym8,3)		移动平均平滑法(3点) Average smoothing (3)		Savitzky2Golay 卷积平滑法 Savitzky2Golay smoothing	
	SNR	RMSE/%	SNR	RMSE/%	SNR	RMSE/%
1	177.083 6	0.000 12	170.773 3	0.000 17	60.176 8	0.041 9
2	174.189 3	0.000 14	166.349 8	0.000 21	59.902 5	0.042 1
3	176.948 2	0.000 12	171.383 1	0.000 15	58.715 8	0.043 0
4	171.868 1	0.000 16	168.739 2	0.000 19	60.200 6	0.042 8
5	176.366 2	0.000 12	169.763 0	0.000 17	59.191 0	0.041 8
平均值 Mean	175.291 1	0.000 13	169.401 7	0.000 18	59.637 3	0.042 3
标准差 S_d	2.238 6	0.000 019	1.980 6	0.000 021	0.657 1	0.000 545

从表2中可以看到小波去噪和移动平均平滑法2种处理方式分析精度较高,且小波去噪方法略高于移动平均平滑法。小波消噪方法的信噪比明显优于卷积平滑法,SNR 从 60 dB 左右提高到 175 dB 左右,原始信号与消噪信号之间的均方根误差由 0.042 3降到 0.000 13,表明小波消噪效果最理想。

2.2 优选波长

采用校正集 136 个样本建立模型,然后用检验

集中的 34 个样本进行预测。首先用全波段光谱 2 001个变量建立 PLS 全谱模型，得到预测相关系数为 0.855 5，预测均方根误差为 0.050 4。但全波段建模数据量太大，故用 iPLS 方法进行变量区间选择。经过分析比较，将全波段 2 001 个变量等间距分为 10 个子区间，每个区间有 200 个变量（最后 1 个子区间有 201 个变量，范围是 2 300～2 500 nm），siPLS 优化效果较好。分别用每个区间光谱建立 PLS 模型。图 2 为基于各子区间所建立 PLS 模型的 RMSECV 分布情况，可见在 700～899 nm 区间的 200 个变量所建立的 PLS 模型的 RMSECV 值最小，而 2 300～2 500 nm 范围的 PLS 模型的 RMSECV 值最大。这表明全谱范围中各个区间对待测成份的敏感程度不同，存在部分信息冗余。将最佳的 1 个、2 个、3 个和 4 个区间组合，即采用 iPLS、siPLS_2、siPLS_3 和 siPLS_4 波长选择方法，建立 PLS 模型，其预测精度均低于全谱模型，结果如表 3 所示。

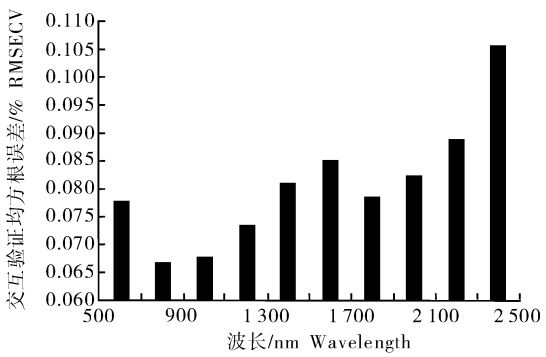


图 2　基于 10 个子区间的 PLS 模型的 RMSECV 分布

Fig. 2　RMSECV of PLS models based on 10 intervals

表 3　基于不同变量的完整果总酸 PLS 预测模型比较

Table 3　Comparison of PLS models based on different variables

变量选择方法 Method of variable selection	变量个数 Number of variables	R_c	RMSEC	R_p	RMSEP
无 Non	2 001	0.881 0	0.058 2	0.855 5	0.050 4
iPLS	200	0.867 2	0.061 2	0.754 5	0.059 8
siPLS_2	400	0.872 1	0.060 2	0.797 2	0.054 8
siPLS_3	600	0.890 0	0.056 0	0.802 5	0.055 3
siPLS_4	800	0.878 0	0.058 5	0.730 6	0.063 3
SPA	21	0.853 1	0.064 1	0.821 8	0.055 2
biPLS _SPA	14	0.863 1	0.062 1	0.858 9	0.054 6

在全谱范围中使用 SPA 算法，进行变量压缩，最终得到 21 个变量，以此建立全谱 SPA 模型。该模型的预测相关系数为 0.821 2，预测均方根误差为 0.055 2。相对于全谱模型而言，该模型所使用的波长变量大大减少，但同时预测性能也有所下降。

分析连续投影算法，是通过计算吸光度矩阵中某一波长对其他波长的投影，选取投影量最大的波长作为该波长序列中的下个波长，序列中的每个波长都与其前 1 个波长相关性最小，最大程度消除共线性对模型的影响，以此降低模型复杂度，却没有考虑到总酸含量矩阵的影响。故此借助 biPLS 的思想，根据图 2，从建模 RMSECV 最大的区间开始，每次从全谱中去掉 1 个区间，将剩下的波长变量使用 SPA 进行压缩，然后建立 PLS 模型，直到剩下最后 1 个区间为止，从中选出最优的模型，并把这种波长选择方法称为 biPLS_SPA 法。使用该方法的波长选择结果如图 3 所示，当波长数为 14 个时，PLS 模型的验证均方根误差达到最小。此时去掉了第 10 个区间，即变量压缩在 500～2 299 nm 范围内进行，提取出 547、581、602、711、786、908、1 204、1 302、1 344、1 389、1 894、2 110、2 200、2 299 nm 共 14 个变量，波长数量只占 2 001 个全部波长变量的 0.70%。

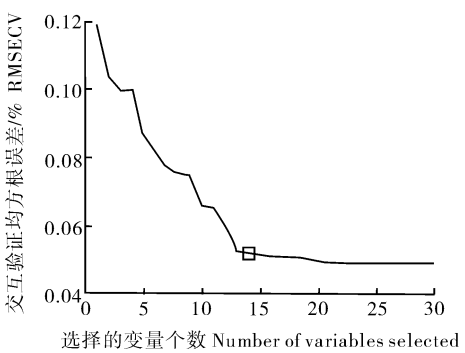


图 3　SPA 选取波长数目与对应的 RMSE 关系

Fig. 3　Relation of variables distilled by SPA and RMSE

2.3　建模结果

使用 14 个优选波长建立 3 层 BPNN 模型和 PLS 模型，并用检验集中的 34 个样本进行预测，预测结果如表 4 所示，预测值与实验室测定值之间的回归曲线如图 4 和图 5 所示。

由表 4 可见，PLS 模型的预测相关系数达 0.859，不仅高于全谱 PLS 模型的预测精度，而且高于全谱 SPA 压缩后 21 个波长变量建立的 PLS 模

表4 PLS模型和BPNN模型预测结果比较

Table 4 Comparison of PLS and BPNN models

变量个数 Number of variables	建模方法 Method of modeling	R_p	RMSEP
14	BPNN(8个隐层神经元, 8 hidden layer neurons)	0.867	0.061 6
14	BPNN(9个隐层神经元, 9 hidden layer neurons)	0.856	0.055 7
14	BPNN(10个隐层神经元, 10 hidden layer neurons)	0.830	0.061 6
14	PLS	0.859	0.052 6

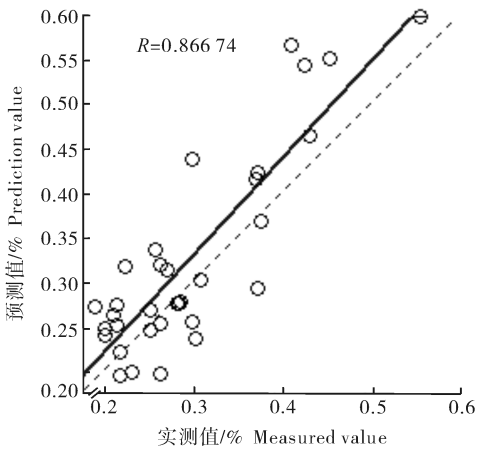


图4 基于14个变量最佳BPNN网络的预测回归图

Fig. 4 Regression curve of sample's real value and prediction value of BPNN models based on 14 variables

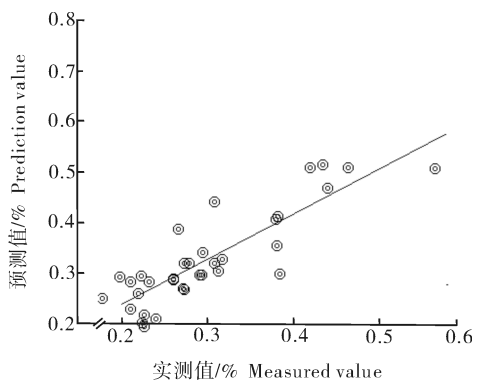


图5 基于14个变量的完整果预测集样本预测值与真实值的回归图

Fig. 5 The regression picture of intact fruits prediction set's prediction value and real value of PLS model established by 14 wavelengths

型精度。非线性的人工神经网络BPNN模型,在选择合适隐层节点数之后,产生了0.867的预测相关系数,表明BPNN模型具有很好的非线性映射能力,适合于基于近红外光谱建模的分析。同时也表明biPLS_SPA方法不仅能最大程度消除共线性对模型的影响,而且能剔除与待测成份不敏感的波长,有效提取特征变量,简化模型结构,提高模型性能。

3 讨 论

利用连续投影算法和小波去噪对砂糖橘样品的光谱进行处理,进而得出砂糖橘的总酸预测模型。相对于平滑去噪法,本文采用小波去噪,能提高信号的信噪比和降低原始信号与消噪信号之间的均方根误差,去噪性能更为理想。本文使用连续投影算法能比较有效地对海量光谱数据进行压缩。全谱SPA模型使用21个波长变量,所得的预测相关系数及均方根误差为 $R_p=0.821\ 8$ 和 RMSEP = 0.055 2。相比同类研究中单独采用一种波长选择方法,本文将biPLS与SPA有机结合,预先缩小光谱范围,剔除对待测成份不敏感的波长,能简化模型结构,并提高模型性能。使用biPLS_SPA方法后,PLS模型使用14个波长数据,得到的预测相关系数及均方根误差为 $R_p=0.858\ 9$ 和 RMSEP = 0.054 6。建模效果优于全谱模型和全谱SPA模型。BPNN模型具有很好的非线性映射能力,适合于近红外光谱的建模工作,在选择合适的隐层节点后,其建模效果优于PLS模型。

参 考 文 献

[1] 牛森.作物品质分析[M].北京:农业出版社,1992.

[2] 严衍禄,赵龙莲,韩东海,等.近红外光谱分析基础与应用[M].北京:中国轻工业出版社,2005.

[3] 洪添胜,乔军,NING W,等.基于高光谱图像技术的雪花梨品质无损检测[J].农业工程学报,2007,23(2):151-155.

[4] 吴泽鑫,李小昱,王为,等.基于近红外光谱的番茄农药残留无损检测方法研究[J].湖北农业科学,2010,49(4):961-963.

[5] 耿响,陈斌,叶静,等.茶叶咖啡碱近红外光谱模型简化方法[J].农业工程学报,2009,25(10):345-349.

[6] 高志奎,王梅,任士福,等.定植密度和剪叶处理对日光温室番茄冠层光截获性能的影响[J].华中农业大学学报,2011,30(2):161-166.

[7] 蔡剑华,王先春,胡惟文,等.基于EMD的土壤有机质含量近红外光谱检测[J].农业机械学报,2010,41(9):182-186.

[8] 黄凌霞,吴迪,金航峰,等.基于变量选择的蚕茧茧层量可见-近红外光谱无损检测[J].农业工程学报,2010,26(2):231-236.

[9] 李东华,纪淑娟,重滕和明.南果梨糖、酸度近红外光谱模型适用的贮藏期研究[J].农业工程学报,2009,25(4):270-275.

[10] 于海燕,应义斌,刘燕德.农产品品质近红外光谱分析结果影响因素研究综述[J].农业工程学报,2005,21(11):160-163.

[11] 洪涯，洪添胜，代芬，等. 连续投影算法在砂糖橘总酸无损检测中的应用[J]. 农业工程学报，2010，26(12)：380-384.
[12] 郝勇，陈斌，朱锐. 近红外光谱预处理中几种小波消噪方法的分析[J]. 光谱学与光谱分析，2006，26(10)：1838-1841.
[13] 史波林，赵镭，刘文，等. 苹果内部品质近红外光谱检测的异常样本分析[J]. 农业机械学报，2010，41(2)：132-137.

Nondestructive test of total acidity in 'shatangju' (*Citrus reticulatablanco*) with near-infrared spectroscopy

DAI Fen　HONG Tian-sheng　LUO Xia　HONG Ya　LI Yan

College of Engineering, South China Agricultural University/ Key Laboratory of Southern Agricultural Machinery and Equipment Key Technology, Ministry of Education/Mechanical Laboratory of National Citrus Industry Technology System, Guangzhou 510642, *China*

Abstract　Near-infrared spectroscopy was used to measure total acidity in 'shatangju' (*Citrus reticulatablanco*). The diffuse reflection spectra of 170 intact samples within 500-2 500 nm were collected. The total acidity in intact samples were measured by titration method. After that, the spectra were de-noised using the orthogonal wavelet functions sym8 (level=3). And then the spectra variables were optimized by successive projections algorithm (SPA) and interval partial least squares (iPLS). Finally, the PLS calibration models of intact samples were established and compared. As a result, wavelet de-noising can produce higher SNR and lower RMSE than that of routine method. Wavelength variables were decreased from 2 001 to 14 by biPLS_SPA, and this can help to make the models more concise and robust. The BPNN model produced R_P=0.867 and RMSEP=0.061 6 with 14 variables as inputs.

Key words　near-infrared spectroscopy; wavelet de-noising; successive projections algorithm (SPA); 'shatangju' (*Citrus reticulatablanco*); total acidity; nondestructive examination

(责任编辑：陆文昌)