中国农业信息
第 30 卷第 4 期 2018 年 8 月　　2018，30（4）：101-108　　China Agricultural Informatics
Vol.30，No.4 Aug.，2018

专题报告

苹果糖度无损检测模型研究*

王风云[1]，沈　宇[1,2]，张　琛[1,2]，刘炳福[1]，郑纪业[1※]
（1. 山东省农业科学院科技信息研究所，济南 250100；2. 山东科技大学计算机科学与工程学院，青岛 266000）

摘要：【目的】对套袋和不套袋苹果分别建立反射光谱与糖度预测模型，并对模型的精度进行比较分析，为构建苹果品质分级系统提供理论支撑。**【方法】**采用美国 ASD 公司的便携式光谱仪和数显折光计分别测量套袋和不套袋烟富 3 号红富士苹果，以苹果赤道位置 4 个取样点的反射率光谱和对应位置的糖度为数据源，原始光谱经多元散射校正后，与糖度数据一同用偏最小二乘回归算法，分别建立套袋和不套袋苹果的反射率光谱糖度模型，进行糖度预测。**【结果】**（1）套袋苹果校正集相关系数 Rc=0.76，均方根误差 RMSEP=0.837 5 Brix；预测集相关系数 Rv=0.72，均方根误差 RMSEP=0.870 2 Brix；（2）不套袋苹果校正集相关系数 Rc=0.69，均方根误差 RMSEP=0.904 0 Brix；预测集相关系数 Rv=0.63，均方根误差 RMSEP=0.913 4 Brix。**【结论】**不套袋苹果的模型精度低于套袋苹果模型精度。相对复杂的表面情况导致不套袋苹果模型精度较差，不套袋苹果的无损检测误差会高于套袋苹果。

关键词：苹果；套袋；不套袋；偏最小二乘回归；多元散射校正；糖度；无损检测
DOI：10.12105/j.issn.1672-0423.20180409

0　引言

中国是世界上最大的苹果生产国[1]。清脆多汁、酸甜可口的苹果一直深受人们的喜爱。尽管苹果产量居世界第一，我国苹果产业仍存在一味追求产量、果品质次价低的问题，即“好的不多，多的不好”。根据农业农村部发表的《苹果品质指标评价规范》[2]，苹果的品质指标包括外观品质指标和内在品质指标。苹果的糖度是内在品质指标的重要组成部分，也是消费者选购苹果的主要依据之一。因此，找到能够无损、快速检测糖度的方法对我国苹果产业具有重要意义。

测量苹果糖度的传统方法是将苹果取样榨汁后用折光计测量，会损坏被测量苹果，加之测量速度较慢，只适用于小规模抽样，不能满足消费者对糖度的差异化需求。近年来，近红外光谱技术被应用于苹果糖度的无损测量，可以对苹果进行快速、大

收稿日期：2018-08-10
第一作者简介：王风云（1974—），女，汉族，山东肥城人，研究员。研究方向：农业信息化。Email：wfylily@163.com
※ **通信作者简介：**郑纪业（1982—），男，汉族，山东泰安人，博士、助理研究员。研究方向：农业信息化。Email：jiyezheng@163.com
* **基金项目：**山东省农业科学院农业科技创新工程（CXGC2017B04）；中国农业科学院与山东省农业科学院科技创新工程协同创新任务（CAAS-XTCX2018023）

批量的无损检测，目前已投入生产实践。该技术的缺点是检测指标较为单一，且不能检测苹果的外观品质指标。高光谱成像技术的研究与应用逐步扩展到农业领域，并应用于苹果品质无损检测中。2011 年，单佳佳等[3]结合高光谱图像处理和光谱分析方法，通过图像扫描对苹果的表面摔伤和糖分含量进行检测，实现了苹果内部品质和外部品质的同时检测。2012 年，郭俊先等[4]采用一阶微分进行光谱预处理，基于多元线性回归（Multivariable Linear Regression，MLR）方法建立苹果糖度的预测模型。2013 年，黄文倩等[5]采用遗传算法（Genetic Algorithm，GA）、连续投影算法（Successive Projections Algorithm，SPA）和 GA-SPA 算法分别从 400~1 000 nm 的苹果高光谱图像中提取特征波长，利用偏最小二乘法（Partial Least Square，PLS）、最小二乘支持向量机（Least Squares Support Veotor Machine，LS-SVM）和多元线性回归（MLR）建模进行苹果可溶性固形物含量（Soluble Solids Content，SSC）的定量分析并进行了综合比较，指出可用连续投影算法（SPA）来进行光谱数据的筛选。2014 年，郭志明等[6]采用偏最小二乘法建立苹果糖度定量分析模型，结果表明提取圆形感兴趣区域建立的苹果糖度模型精度最高，预测能力最强。2015 年，刘文涛等[7]用 BP 神经网络建立了糖度预测模型。2016 年，张晋宝等[8]用偏最小二乘回归（Partial Least Square Regression，PLSR）建立了糖度模型。2017 年，冯迪等[9]用 SPA 算法找到预测苹果糖度和硬度的最佳波长。2018 年，管晓梅等[10]采用优化偏最小二乘因子数的方法，提高模型的预测能力，同时降低了模型的复杂度。

目前为止，还没有套袋苹果与不套袋苹果糖度无损检测的对比研究，但是消费者对套袋与不套袋苹果糖度有着不同的需求，为此文章以烟富 3 号红富士苹果为对象，采用高光谱成像技术采集苹果的反射光谱信息，经多元散射校正后采用偏最小二乘回归算法对套袋和不套袋苹果分别建立反射光谱与糖度预测模型，对模型的精度进行比较，并分析了精度不同的原因，为构建苹果品质分级系统提供理论支撑。

1 实验数据

1.1 研究对象

本研究实验对象为矮化烟富 3 号红富士苹果，来自山东某集团栖霞官道镇姚庄村碑通达王太后基地，北纬 37° 09′ 46.56″，东经 120° 38′ 24.38″。剔除损伤及采样过程中发现的内部腐烂苹果后，最终获得 90 个套袋苹果和 118 个不套袋苹果作为试验样本。

1.2 实验仪器

实验器材主要包括高光谱仪和糖度计。高光谱图像采集系统如图 1 所示。为了避免周围环境光照的影响，保证目标样本光照的均匀性，将整个图像采集系统（除计算机外）置于暗箱中运行。实验选用美国 ASD 公司设计制造的 FieldSpec Hand-Held 便携式地物光谱仪，其主要组成包括光谱仪本体、光纤、探头以及用来做光强校正的白板等。测量光谱的范围是 350~1 000 nm，波长精度为 +/-1 nm，光谱分辨率是 3 nm@700 nm。糖度计使用陆恒生物公司的 LH-B55 数显糖度计。数显糖度计可以快速测定含糖溶液的糖浓度和折射率。该糖度计的量程是 0.0~55% Brix，分辨率是 0.1% Brix，精度是 +0.2 Brix。

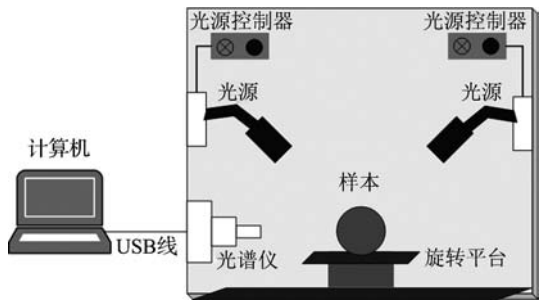

图 1　高光谱数据采集平台

Fig.1　Hyperspectral data acquisition platform

1.3　高光谱数据采集

高光谱图像的采集和处理也称成像光谱学，鉴于所采集数据的形式是高光谱立方体，有时也被称为 3D 光谱学。高光谱成像本质是收集和处理来自电磁频谱的信息。高光谱成像目的是通过获取场景图像中每个像素的光谱来寻找物体、识别材料或检测特定过程[11]。形象地说，高光谱传感器收集信息作为一组“图像”，每幅图像代表电磁频谱中的一个波段，一个波段也被称为一个光谱带。我们可以将这些“图像”组合起来，形成用于处理和分析的三维（X，Y，λ）高光谱数据立方体，其中 X 和 Y 代表场景的 2 个空间维度，λ 表示光谱维度。

研究表明，感兴趣区域选取的形状会影响苹果糖度模型的精度。根据郭志明等人的研究，圆形感兴趣区域精度最好[6]。在苹果赤道位置选取 4 个均匀分布的、直径约为 3 cm 的圆形区域作为感兴趣区域，以每个感兴趣区域的平均光谱作为样品的高光谱，共采得 208 个苹果 832 条光谱样本。

1.4　糖度数据采集

用小刀剜削苹果赤道位置 4 个感兴趣区域挖取长宽各 3 cm、厚 2 cm 的立方体果肉，将榨好的苹果汁涂布折光棱镜的镜面上，连续按测量按钮多次，当最后液晶显示屏 3 次显示值一致时记录该值，共采得 832 个数据，与感兴趣区域的高光谱数据一一对应。

2　数据处理方法

2.1　样本划分

套袋苹果中抽取 66 个苹果样本数据作为校正集，24 个作为预测集。不套袋苹果中抽取 87 个苹果样本数据作为校正集，31 个作为预测集。校正集：预测集约等于 3∶1。

2.2　光谱数据预处理

采集到的反射率光谱首先经过多元散射校正（Multiple Scattering Correction，MSC），消除散射对光谱的影响，提高信噪比，增强光谱与糖度的相关性。

MSC 处理方法：首先通过式（1）求得所有感兴趣区域光谱的平均光谱，将其作为“理想光谱”。将每条光谱与“理想光谱”按式（2）作一元线性回归运算，求得相对于标

准光谱的数值差（回归常数 b_i）和斜率倍数（回归系数 m_i），最后根据式（3）在每条原始光谱中减去数值差同时除以回归系数，原始光谱的各波段上数值及曲线斜率都得到修正[12]。

$$\overline{A}_{i,\ j}=\frac{\sum_{i=1}^{n}A_{i,\ j}}{n} \tag{1}$$

$$A_i=m_i\overline{A}+b_i \tag{2}$$

$$A_{i\ (MSC)}=\frac{(A_i-b_i)}{m_i} \tag{3}$$

上式中，A 表示 $n\times p$ 维定标光谱数据矩阵，n 为定标样品数，p 为光谱采集所用的波长点数，$A_{i,\ j}$ 表示所有样品的原始光谱在各个波长点处求平均值所得到的平均光谱矢量，A_i 是 $1\times p$ 维矩阵，表示单个样品光谱矢量，m_i 和 b_i 分别表示各样品光谱 A_i 与平均光谱 $\overline{A}_{i,\ j}$ 进行一元线性回归后得到的相对偏移系数和平移量。

2.3 模型建立

偏最小二乘回归是一种使用包含相关预测变量数据的技术，是通过将预测变量和可观察变量投影到新空间来找到一个线性回归模型，而主成分回归是寻找响应和自变量之间最大方差的超平面。因为数据 X 和 Y 都投影到新的空间，所以 PLS 系列方法也被称为双线性因子模型。

PLSR、多元线性回归与主成分分析（Principal Component Analysis，PCA）之间的交叉点：多元线性回归可找到符合响应值的预测变量的组合；主成分分析发现具有较大方差的预测变量组合，减少相关性，PCA 不使用响应值。PLS 发现具有较大协方差的预测变量与响应值的组合。因此，PLS 结合了关于预测变量和响应变量的信息，同时也考虑了它们之间的相关性。

PLS 用于找出两个矩阵（X 和 Y）之间的基本关系，例如使用某种潜在变量方法来模拟这 2 个空间中的协方差结构。PLS 模型目标是在 X 空间中找到解释 Y 空间中最大多维方差方向的多维方向。PLS 回归特别适用于预测变量矩阵比观测变量多以及 X 值之间存在多重共线性的情况。

PLS 的一般基础模型：

$$\begin{aligned}X&=TP^{\mathrm{T}}+E\\Y&=UQ^{\mathrm{T}}+F\end{aligned} \tag{4}$$

其中 X 是 $n\times m$ 的预测矩阵，Y 是 $n\times p$ 的响应矩阵；T 和 U 分别是 X 的投影（X 分数，分量或因子矩阵）和 Y（Y 分数）的投影；P 和 Q 分别是 $m\times 1$ 和 $p\times 1$ 的正交载荷矩阵；矩阵 E 和 F 是误差项，假设它们是独立且均匀分布的随机正态变量。分解 X 和 Y 是为了使 T 和 U 之间的协方差最大化。

2.4 模型评判

偏最小二乘回归模型的主要评判指标是校正集和预测集的相关系数 R 和均方根误差 RMSEP。相关系数越接近于 1，均方根误差越小，则模型精度越好。

3 结果与分析

3.1 光谱数据预处理

图 2 展示的是部分由 ASD 光谱仪采得的原始光谱数据，但原始光谱并不适合直接用于建模，主要因为：（1）虽然光谱的整体趋势一致，但不同光谱反射率数值的大小却不尽相同，原因是苹果的形状不规则，每个苹果的 4 个面形状不同，不同苹果的外形有较显著的差异，这就导致卤素灯照到每处采样点的光强不同，反射率的数值自然不同；（2）光谱两端有较多较大的噪声，这是由采集反射光谱的硅光电二极管的特性决定的，光谱仪和其他许多仪器一样，量程中间精度好，两端差。为解决光谱两段噪声多的问题，裁掉两端的光谱，保留中间 420~1 019 nm 波段的光谱。

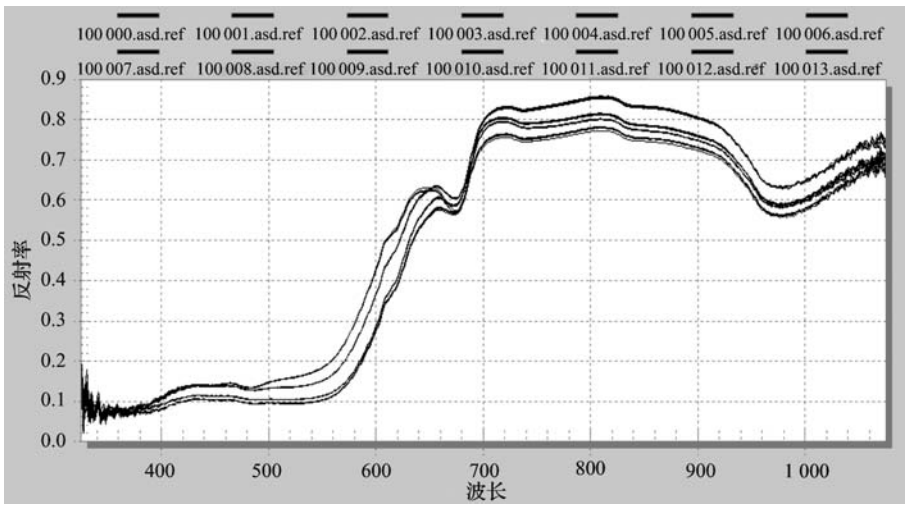

图 2 部分原始光谱数据

Fig.2 Part of original spectral data

之后，使用多元散射校正（MSC）算法对光谱进行处理，目的是减少光照强度不均对苹果表面反射率的影响。图 3 和图 4 分别是多元散射校正前后的光谱图。可以看出经过多元散射校正，光谱向平均光谱（即 MSC 的“理想光谱”）靠拢。

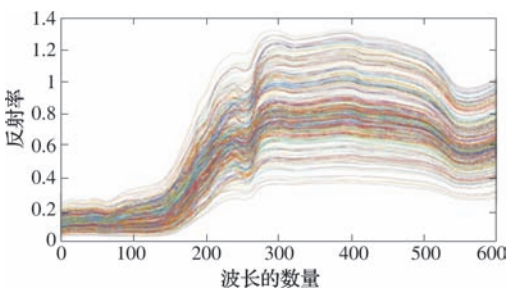

图 3 多元散射校正前的光谱

Fig.3 Spectrum before multiplicative scatter correction

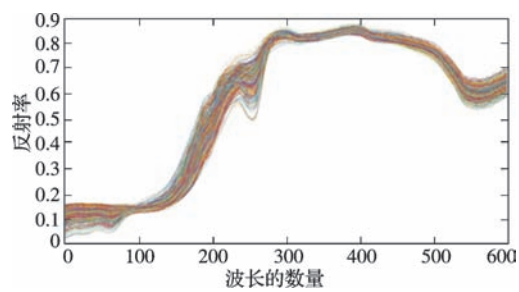

图 4 多元散射校正后的光谱

Fig.4 Spectrum after multiplicative scatter correction

3.2 模型建立

如上所述，使用偏最小二乘回归（PLSR）算法建立苹果反射率光谱 – 糖度模型，首先要确立光谱数据主成分的个数，目的是将 600 个波段的光谱信息压缩为一定数量的主成分信息，方便建模。

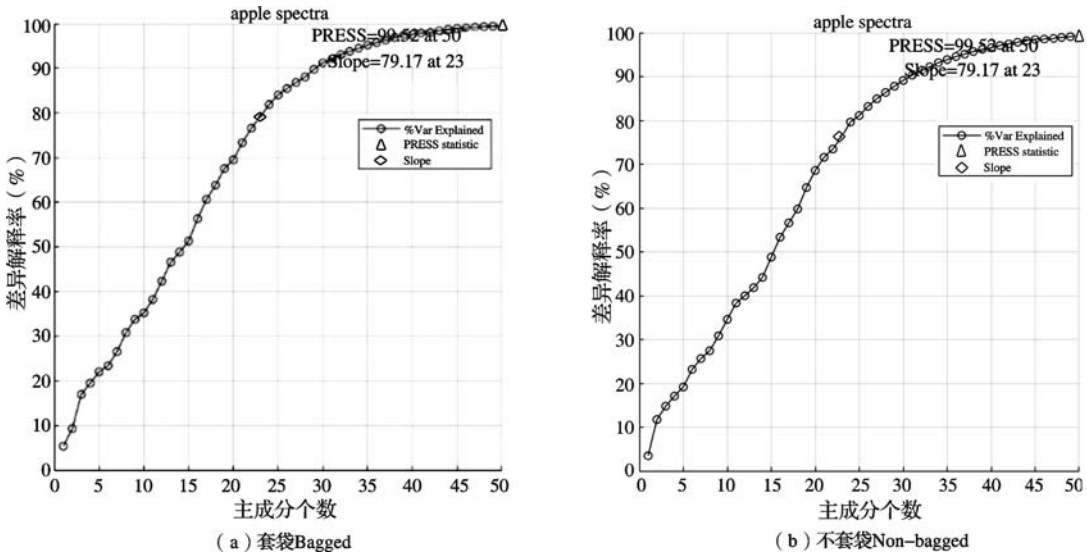

图 5 前 50 个主成分对光谱差异的累积解释率

Fig.5 Cumulative interpretation rate for spectral difference of ahead 50 principal components

由图 5 可以看出，前 50 个主成分对光谱差异的累积解释率达 99.52%，满足建模要求。用偏最小二乘回归（PLSR）建立的苹果反射率光谱 – 糖度模型，其预测结果的残差分布如图 6 所示。

用偏最小二乘回归（PLSR）建立的苹果反射率光谱预测糖度模型，对于套袋苹果，校正集相关系数 Rc=0.76，均方根误差 RMSEP=0.837 5 Brix；预测集相关系数 Rv=0.72，均方根误差 RMSEP=0.870 2 Brix。对于不套袋苹果，校正集相关系数 Rc=0.69，均方根误差 RMSEP=0.904 0 Brix；预测集相关系数 Rv=0.63，均方根误差 RMSEP=0.913 4 Brix。

3.3 结果分析

根据实验结果，不套袋苹果建立的模型精度低于套袋苹果。苹果套袋与不套袋对模型精度的影响是由不同的表面状况造成的。所有实验样品在实验开始时均未经过清洗，套袋苹果表面较为干净，除极个别苹果表面在运输过程中碰伤外，其余苹果表面均无伤痕，苹果各个面颜色基本一致，手感光滑；不套袋苹果表面灰尘较多，苹果表面有大量的斑点以及在成长过程中的伤疤，且向阳面与背阴面颜色相差较大，手感粗糙。不套袋苹果复杂的表面情况会在一定程度上对光谱采集带来不利影响，光谱中较多的噪声导致了较低的建模精度。

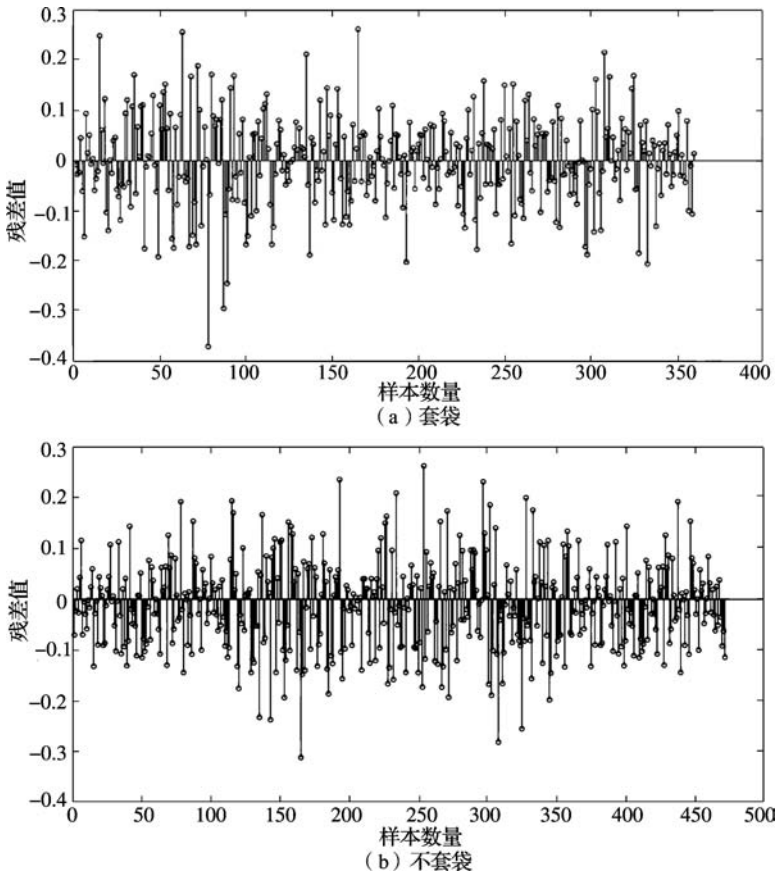

图 6　模型预测结果：残差分布

Fig.6　Residual distribution for prediction results

4　讨论

本文选择烟台栖霞红富士苹果为试验材料，测得反射率光谱及其糖度，用多元散射校正（MSC）处理原始光谱，最后用偏最小二乘回归（PLSR）分别建立了套袋和不套袋苹果的反射率光谱 – 糖度模型。

本研究结果与郭俊先[4]、郭志明[6]、张晋宝[8]等已有研究结果一致，可以利用偏最小二乘回归（PLSR）来建立苹果糖度定量分析预测模型。但由于套袋与不套袋苹果差异，使用同一种方法建立的模型，其结果还是存在差异。不套袋苹果模型精度低于套袋苹果，原因是不套袋苹果表面状况较为复杂，影响了光谱采集的精度，增加了无损检测的难度，不套袋苹果的无损检测误差要高于套袋苹果。今后套袋苹果与不套袋苹果光谱数据差异性，需通过不同模型与方法研究，以提高糖度检测模型的预测精度。

参考文献

[1] 陈磊 . 苹果价格下滑 , 果农果商如何应对 ?—今年我国苹果产量和价格走势预测分析 . 果农之友 , 2015(12): 43~45.

[2] 中华人民共和国农业部 . 苹果品质指标评价规范（NY/T2316–2013）, 2013.
[3] 单佳佳 , 彭彦昆 , 王伟 , 等 . 基于高光谱成像技术的苹果内外品质同时检测 . 农业机械学报 , 2011(3): 140~144.
[4] 郭俊先 , 饶秀勤 , 程国首 , 等 . 基于高光谱成像技术的新疆冰糖心红富士苹果分级和糖度预测研究 . 新疆农业大学学报 , 2012(1): 78~86.
[5] 黄文倩 , 李江波 , 陈立平 , 等 . 以高光谱数据有效预测苹果可溶性固形物含量 . 光谱学与光谱分析 , 2013(10): 2843~2846.
[6] 郭志明 , 黄文倩 , 彭彦昆 , 等 . 高光谱图像感兴趣区域对苹果糖度模型的影响 . 现代食品科技 , 2014(8): 59~63.
[7] 刘文涛 . 基于高光谱成像技术的苹果品质无损检测研究 . 保定 : 河北农业大学 , 2015.
[8] 张晋宝 . 高光谱技术在苹果检测中的应用 . 吉林 : 东北电力大学 , 2016.
[9] 冯迪 , 纪建伟 , 张莉 , 等 . 基于高光谱成像提取苹果糖度与硬度最佳波长 . 发光学报 , 2017(6): 799~806.
[10] 管晓梅 , 杜军 , 张立人 , 等 . 基于高光谱技术的果糖检测优化算法和可视化方法 . 光电子 • 激光 , 2018, 29(2): 173~180.
[11] Hans Grahn, Paul Geladi. Techniques and applications of hyperspectral image analysis. *John Wiley & Sons*, 2007.
[12] Gao, L L, Zhu X C, Li C, et al. Improve the prediction accuracy of apple tree canopy nitrogen content through multiple scattering correction using spectroscopy. *Agricultural Sciences*, 2016(7): 651~659.

Sugar content nondestructive testing study for apple

Wang Fengyun[1], Shen Yu[1, 2], Zhang Chen[1, 2], Liu Bingfu[1], Zheng Jiye[1※]
（1. S&T Information Institute of Shandong Academy of Agricultural Sciences, Jinan 250100, China;
2. College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266000, China）

Abstract: [**Purposes**] Regarding the different sugar content in bagged and unbagged apples, reflectance spectra and sugar prediction models were established respectively. Moreover, the accuracy of these models were compared and analyzed, in order to provide theoretical support for the construction of apple quality classification system. [**Methods**] This research applied the ASD's portable spectrometer and digital refractometer to measure the reflectance spectra of the same batch of Yanfu 3 Hao Red Fuji apples and the sugar content of the corresponding locations. [**Results**] After the original spectrum was corrected by multiple scatter, the partial least-squares regression algorithm was used together with the sugar content data to establish the reflectance spectra of the bagged and non-bagged apples-the sugar content model. The bagged apple model correction set correlation coefficient R_c is 0. 76, root mean square error RMSEP is 0. 837 5 Brix; prediction set correlation coefficient R_v is 0. 72, root mean square error RMSEP is 0. 870 2 Brix. Uncorked apple model correction set correlation coefficient R_c is 0. 69, root mean square error RMSEP=0. 904 0 Brix; prediction set correlation coefficient R_v is 0. 63, root mean square error RMSEP is 0. 913 4 Brix. [**Conclusion**] The relatively complex surface conditions lead to poorer accuracy of the non-bagged apple model. Non-bagged apples have a higher NDT error than bagged apples.

Key words: apple; bagged; non-bagged; partial least square regression; multiple scatter correction; sugar content; nondestructive testing