基于高光谱的柑橘叶片氮素含量多元回归分析

黄双萍[1,2,3]，洪添胜[1,2,3]，岳学军[1,2,3,4※]，吴伟斌[1,2,3]，蔡坤[1,2,3]，徐兴[1,2,3]

（1. 华南农业大学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，广州 510642；
2. 国家柑橘产业技术体系机械研究室，广州 510642；3. 华南农业大学工程学院，广州 510642；
4. 南昆士兰大学工程与测绘学院，图文巴 QLD4350）

摘 要：快捷、准确、无损地检测柑橘叶片氮（N）素含量，对柑橘树 N 肥施用的精准动态管理有重大现实意义。以 117 株园栽罗岗橙为试验研究对象，在不同生长期用 ASD 公司的 FieldSpec3 采集柑橘树健康叶片的高光谱反射值，以高光谱反射数据或其变换形式作为柑橘树样本多元矢量描述；用凯氏定氮法同期检测出柑橘树叶的真实 N 素含量值；在用 PCA 对高维光谱矢量降维的基础上，利用支持矢量回归算法（SVR）建立高光谱多元表达和 N 素含量间的映射关系，以实现任意柑橘树 N 素含量的预测分析。试验结果表明，测试集上预测值和真实值间的平方决定系数 R^2 为 0.9730，平均相对误差为 0.9033%，均方误差 MSE 为 0.090343，证明了该方法的有效性，为利用高光谱技术进行柑橘树 N 素含量的无损检测提供了参考。

关键词：N 元素，回归分析，高光谱，柑橘叶片，SVR
doi：10.3969/j.issn.1002-6819.2013.05.018
中图分类号：S24 **文献标志码**：A **文章编号**：1002-6819(2013)-05-0132-07

黄双萍, 洪添胜，岳学军，等. 基于高光谱的柑橘叶片氮素含量多元回归分析[J]. 农业工程学报，2013，29(5)：132－138.
Huang Shuangping, Hong Tiansheng, Yue Xuejun, et al. Multiple regression analysis of citrus leaf nitrogen content using hyperspectral technology[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2013, 29(5): 132－138. (in Chinese with English abstract)

0 引 言

氮（N）元素是柑橘（citrus reticulata banco）生长所需的一种重要养分，被称之为柑橘体内的蛋白质；虽然如此，但过度施用 N 肥，会造成柑橘树长叶不长果，果实酸度增高等问题，同时造成种植成本上升、环境污染、土壤肥力下降等负面影响。因此，必须根据柑橘树 N 素水平和生长状况，按需施用 N 肥，对 N 肥用量进行精准管理，达到提高产量、节约肥料、节能减排目的。为此，必须快速、准确、无损地检测柑橘树 N 素含量，为柑橘果园定量按需施用氮肥提供依据。

近年来，高光谱技术作为一种农作物 N 素含量无损定量检测方法而备受关注[1-11]。这些研究工作可分为一元回归分析和多元回归分析两大类。前者以单一敏感波长点或多个特征波长点组合光谱信息（各种光谱植被指数）为自变量，基于训练数据建立自变量与预测变量（各种作物营养元素）间的映射关系，进而反演新的作物样本 N 素含量值[1-7]。例如，Dennis 等[1]建立包括 NDVI，GNDVI 等植被指数与小麦 N 含量间的二次多项式模型，模型决定系数（R^2）在 0.52～0.80 间。Daniela 等[3]通过测试水稻光谱反射值在可见光/短波红外区内所有波长组合，找到与植物氮浓度性最相关的归一化指数（NDI），建立对数回归模型，R^2 为 0.53。

另一种较为普遍采用的技术是多元回归分析法，用多点光谱信息构成多元描述,建立该描述与营养成分间的多元回归模型，取得了更好的预测效果[8-10]。文献[8]利用原始光谱数据或其变形，分别用多元线性回归[20-21]和 BP 神经网络（BP，back propagation）[18-19]建立估计油菜 N 素含量模型，取得最高 R^2 为 0.8689。Valentina[8]用偏最小二乘法（PLS，partial least square）[17]，建立西红柿叶片的可见光-近红外反射谱倒数的对数与其 N 素含量间的函数关系，模型决定系数为 0.94。这些研究成果说明利用作物光谱信息进行 N 素营养诊断的可行性，但这些研究成果主要集中在水稻、油菜、冬小麦等大田作物。

收稿日期：2012-09-15 修订日期：2013-02-27
基金项目：国家自然科学基金（30871450）；广东省自然科学基金项目（S2012010009856）；省部产学研结合项目（2011B090400359)资助
作者简介：黄双萍（1972－），女，湖南邵阳人，讲师，博士，农业工程学会高级会员（E041200596S），主要从事农业智能信息处理和数据挖掘，计算机视觉等方面的研究。广州 华南农业大学工程学院，510642。Email: huangshuangping@scau.edu.cn
※通信作者：岳学军（1971－），女，重庆南岸人，副教授，博士，硕士生导师，农业工程学会高级会员（E041200598S），主要从事农业工程、机电一体化和信息技术应用研究。广州 华南农业大学工程学院，510642。Email：yuexuejun@scau.edu.cn, Xuejun.Yue@usq.edu.au

近年来，仅有少量利用高光谱技术进行柑橘树 N 素含量定量分析建模的相关研究工作[11-13]。P.Menesatti[11]等以塔罗科血橙为试验对象，用 PLS 建立叶片与 N，P，K，Ca，Mg，Fe 等元素含量间的模型，其中 N 模型决定系数为 0.909。易时来等[12]以锦橙为研究对象，研究叶片钾含量与可见近红外反射光谱的相关性。

本文以广州市萝岗区蟹庄村柑橘园 117 株罗岗橙树为试验对象，用 ASD 公司高光谱仪 Fieldspec3 采集不同生长时期罗岗橙叶片高光谱数据，用 350～2 500 nm 波段内光谱反射值构成柑橘树多元矢量描述。用凯氏定氮法[21]测试同期同批柑橘叶样本的 N 素含量，作为预测目标真实值。高光谱多元矢量和 N 素含量值构成建模橘树样本训练数据对。对高维光谱矢量进行 PCA[14]降维以去除数据噪声和冗余，利用基于核的非线性支持矢量回归（support vector regression，SVR）[15]分析法建立柑橘树叶高光谱信息与 N 素含量间的映射关系，以实现新样本 N 素含量预测。

1　样本种植管理和训练数据采集

1.1　柑橘树样本的种植管理

广州市萝岗区蟹庄村柑橘园是广东省无公害农产品生产基地，是萝岗甜橙种植示范场。试验选择果园中 4 年生树 117 株作为柑橘树样本，柑橘树位于园内同一区域（坡度 20°，树高 2 m，间距 4×3 m），生长状况基本一致。对选为样本的柑橘树进行规范化植株管理：参照岭南柑橘品种生长发育程度，分别在萌芽期、稳果期、壮果促梢期与采果期施入不同水平的 N 肥，培养试验所需样本；选择纯度为 46%的尿素 $CO(NH_2)_2$ 作为 N 肥，分别在 2011 年 2 月、4 月、6 月、8 月上中旬施用全年 N 肥总量的 30%、30%、25%、15%。柑橘树根部 N 肥施肥条件变化，样本营养状态相异，氮素含量的变化会在叶片光谱特性上有所反应。柑橘树生长温度低于−3℃时就采取喷灌防冻措施；柑橘园内保证良好的光照条件；在生长期内灌溉频度为每周 2 次；加配富含有机质, 含水率高，pH 值为 4～5 的疏松泥炭藓或共生苔藓的土壤。

1.2　训练数据采集

柑橘叶片采集时间为每次施肥 15 天后，共采集了 4 次：2011 年 2 月 20 日、4 月 24 日、6 月 26 日和 8 月 23 日。每株树分东南西北和上下层共采集 8 片大小均匀的健康叶片，选位于顶梢起向下数的第 3 至第 4 片鲜叶。用 ASD Fieldspec3 采集柑橘鲜叶片的光谱反射值，8 片叶子光谱反射值的均值作为柑橘样本的光谱描述。该仪器光谱范围为 350～2 500 nm，光谱分辨率在 350～1 000 nm 范围内为 3nm，在 1 000～2 500 nm 范围内为 7 nm；采样间隔在 350～1 000 nm 范围内为 1.4 nm，在 1 000～2 500 nm 范围内为 2 nm；输出数据间隔为 1 nm。采用全自动方式进行灵敏度调整。由于 ASD Fieldspec3 输出数据间隔为 1 nm，因此，每个柑橘样本的光谱描述是 2151 维的高维矢量。每株树的 8 片鲜叶均匀混合后作为一个样本，用传统的化学测量法凯氏定氮法测量柑橘树叶片的 N 含量。不同批次 117 株树样本的叶片平均高光谱反射谱图见图 1，N 素含量统计数据见表 1。

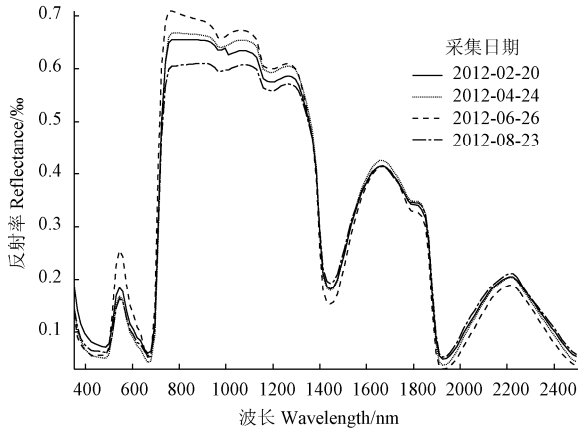

图 1 不同生长阶段 117 株柑橘样本的平均反射光谱

Fig.1 Mean reflectance spectra of 117 citrus samples during the four different growing stages with different N fertilizer

表 1 实验室内柑橘叶片 N 素含量统计描述表

Table 1 Descriptive statistics of the nitrogen content measured in the laboratory at four different periods

‰

N 素含量 Nitrogen content	2012-02-20 Feb.20 2012	2012-04-24 Apr.24 2012	2012-06-26 Jun.26 2012	2012-08-23 Aug.23 2012
最小值 Min	22.6887	23.3182	24.5453	22.5844
最大值 Max	28.2784	28.1674	28.0582	25.7624
平均值 Average	26.6328	26.9818	25.2578	23.9114

2　试验方法及结果分析

2.1　PCA 降维

物质光谱响应有突出的波段性，相邻波段的反射数据有较大相关性。波长 350～2 500 nm 范围内每隔 1 nm 的光谱反射量构成的柑橘树高光谱矢量描述，必定包含大量冗余信息。由于仪器或者测量环境影响，该描述数据同时不可避免地存在较大噪声。因此，在进行模型训练之前，进行去冗余和降噪的高维矢量降维处理。主成分分析（principle component analysis，PCA）[14]是一种通用高维数据

降维工具，通过线性变换将高维数据投影到低维空间，投影的基本原则是：找出最能够代表原始数据的投影方法，即投影降维后数据不失真，降维过程去除数据噪声或冗余信息，从而降低后续建模复杂度，提高多元回归分析精度。

2.2 支持矢量回归分析建模

支持矢量回归（support vector regression，SVR）[15]是一个专门针对有限样本的学习机器，其优化的基本思想是结构风险最小化，即在数据逼近精度与逼近函数复杂性之间寻求折衷，以期获得最好的模型泛化能力。SVR 最终转化为凸二次规划问题，从理论上说，得到的将是全局最优解，解决了神经网络等方法中无法避免的局部极值问题。SVR 优化中巧妙地利用核函数，将复杂实际问题通过非线性变换转换到高维特征空间，在高维空间中构造线性决策函数来实现原空间中的非线性决策。核函数的引进，巧妙地规避高维映射定义和高维空间内积运算问题，并保证模型有较好的推广能力。

考虑到柑橘高光谱数据和预测变量 N 素含量之间映射关系的复杂性和非线性性，利用基于核函数的新的通用学习算法 SVR 建模，在一定程度上规避了过拟风险，用核函数代替线性方程中的线性项，使原来的线性算法“非线性化”，从而完成非线性回归分析。

2.3 试验结果及分析

为提高模型的鲁棒性，适当增加训练样本建立柑橘高光谱数据和 N 素含量的关系模型。试验选定 1，2 两批数据，每批数据 117 个，即总样本数为 234 个，每个样本对应一个数据对(x_i,y)，其中 x_i 是每个柑橘样本（不同发育时期对同一株柑橘树的测试值被认为是 2 个独立样本）高光谱反射数据构成的矢量描述，每个样本有 2151 维光谱数据, 是在波长 350～2 500 nm 范围内的光谱反射量的测定结果。y 是该样本 N 素含量真实值。试验过程中，随机选取 80%的样本，即 187 个数据对构成训练集，剩下 47 个样本数据构成测试集，以评估模型回归预测性能，验证方法的有效性。

为评估回归模型的有效性，试验选定“模型决定系数 R^2”、“最大相对误差值（max relative error）”，“平均相对误差值（average relative error）”，“均方误差 MSE（mean square error）”等评估指标。每种试验方案独立运行 15 次，取性能指标 15 次的均值和标准差作为试验结果记录。

2.3.1 独立主成分数目的确定

试验选择 PCA 对原始光谱数据降维处理，并按照协方差矩阵分解后特征值所占能量比来确定独立主成分数目。试验比较不同主成分能量比时，独立主成分数目及对应模型的性能，试验结果见表 2。

表 2 不同能量比情况下主成分数及模型性能评估

Table 2 Model evaluation with different energy ratio corresponding to different number of principle components

能量比 Energy ratio/%	主成分数 Num. of principle components	模型决定系数（R^2） Coefficient of determination	均方误差（MSE） Mean square error	相对误差（RE） Relative error	
				最大值 Max	平均值 Average
98	3	0.965±0.0100	0.130±0.0312	0.044±0.0202	0.010±0.0008
99	5	0.971±0.0075	0.099±0.0224	0.037±0.0120	0.009±0.0010
99.9	15	**0.973±0.0053**	**0.090±0.0168**	**0.030±0.0074**	**0.009±0.0010**
99.99	39	0.968±0.0064	0.108±0.0237	0.034±0.0137	0.009±0.0010

从表 2 看出，降维后能量比为 99.9%，独立主成分数为 15 时，柑橘样本 N 素含量预测结果最佳，模型决定系数 R^2 高达 0.973，均方误差 MSE 为 0.090，最大相对误差（max relative error）为 3.0%，平均相对误差（average relative error）为 0.90%。相比能量比为 99%时，独立主成分数为 5，R^2 略低，为 0.971，MSE 略高，为 0.099，最大相对误差略高，为 3.7%，这 3 个性能指标的标准差偏高。这表明：当独立成分数目较大时，对任意训练测试数据分割，模型鲁棒性好。从表 2 中还可以看出，当能量比为 99.99%，主成分为 39 时，模型性能反倒差，表明有些能量由数据噪声引起，不是真正的数据信息，这些噪声成分的存在反倒降低模型的回归预测性能。因此，试验最终选定 15 维数据，进行柑橘叶片 N 素含量预测。

2.3.2 SVR 核函数种类的确定

SVR 核函数一般有多项式核（polynomial，公式 1）、径向基函数（radial basis function，RBF,公式 2）、线性核（linear，公式 3）等。SVR 回归分析中，核函数的选择决定映射特征空间的结构，因此往往对拟合结果有较大影响。另一方面，SVR 回归模型的推广性能取决于一组好的参数，包括模型参数 C、核参数 $gamma$、$coef_0$ 等。模型参数（正则化参数 C）能够在模型复杂度和训练误差之间取一个折衷，使模型有较好推广能力，不同数据的子空间中最优的 C 值不同。RBF 核参数 $gamma$ 反映了训练样本数据的分布或范围特性，它确定了局部邻域的宽度。本试验采用格子搜索（grid search）和交

叉验证的方法（5-cross validation），搜索最佳模型参数 C 和核函数参数。试验结果见表 3。

$$(gamma \times u' \cdot v + coef_0)^{degree} \quad (1)$$

$$\exp(-gamma \times |u - v|^2) \quad (2)$$

$$u' \cdot v \quad (3)$$

其中 u，v 是任意两个样本的多元矢量表达，$gamma$，$coef_0$，$degree$ 是多项式核和径向基核参数。

表 3　不同核函数情况下的模型性能评估

Table 3　Model evaluation with different kernel function

核函数 Kernel function	模型决定系数（R^2） Coefficient of determination	均方误差（MSE） Mean square error	相对误差（RE） Relative error	
			最大值 Max	平均值 Average
Linear	0.923±0.0266	0.280±0.0831	0.067±0.0202	0.016±0.0015
RBF	**0.973±0.0053**	**0.090±0.0168**	**0.030±0.0074**	**0.009±0.0010**
polynomial	0.869±0.0847	0.434±0.2364	0.116±0.0764	0.018±0.0023

注：Energy Ratio：99.9%。

从表 3 中可以看出，RBF 核函数情况下 SVR 回归模型性能最好，模型最稳定，而多项式核模型性能最差，模型性能指标标准差最大，意味着模型随不同训练和测试样本的分割性能指标变化很大。

2.3.3　不同光谱数据变形的建模结果

光谱数据由于其多重共线性、反射率的非线性、基线变动和附加散射变动以及随机噪声的叠加等原因而呈现复杂性[22]。为获得良好的建模效果，必须对光谱数据进行合适的预处理。

试验利用原始光谱反射值或其变换形式：一阶导数光谱（R）、二阶导数光谱（(R)）、倒数变换（$1/R$）、对数变换（$\log(R)$）以及倒数的对数（$\log(1/R)$）等作为自变量，以柑橘树叶 N 含量作为因变量，用 SVR 建立多元回归估算模型。试验过程中 PCA 保持能量比取 99.9%，核函数为 RBF。试验结果见表 4。

表 4　反射光谱不同数据变换形式下的模型性能评估

Table 4　Model evaluation with different transformation of hyperspectral reflectance data

光谱数据形式 Transformation of hyperspectral reflectance	主成分数 Num.of principle components	模型决定系数（R^2） Coefficient of determination	均方误差（MSE） Mean square error	相对误差（RE） Relative error	
				最大值 Max	平均值 Average
R	**15**	**0.973±0.0053**	**0.090±0.0168**	**0.030±0.0074**	**0.009±0.0010**
R'	70	0.865±0.0264	0.448±0.0650	0.079±0.0005	0.020±0.0203
(R')	56	0.756±0.0533	0.833±0.1733	0.106±0.0311	0.028±0.0038
$(1/R)$	25	0.833±0.0434	0.602±0.1254	0.030±0.0074	0.097±0.0126
$\log(R)$	29	0.952±0.0143	0.166±0.0471	0.055±0.0243	0.012±0.0010
$\log(1/R)$	29	0.950±0.0146	0.173±0.0445	0.057±0.0212	0.012±0.0009

注：能量比：99.9%；核函数：RBF；原始光谱记为 R

Note: Energy ratio：99.9%；Kernel function：RBF；Denote the original hyperspectral data as 'R'

从表 4 可以看出，利用原始光谱反射值建立 SVR 模型，对柑橘树叶 N 素的预测性能最好。其对数变换、倒数的对数变换也依次取得了较好的预测性能。性能最差的变换形式是二阶导数，其 R^2 仅为 0.756，最高相对误差达 10.6%。

2.3.4　与其他建模方法的比较

将本文方法与其他几种主流多元建模方法：BP 神神经网络方法（back propagation artificial neural networks，BPANN）[18-19]、偏最小二乘法（partial least square，PLS）[17]、逐步多元线性回归法（stepwise multiple linear regression，SMLR）[20-21]进行试验比较，试验结果见图2。图中斜线是1∶1比例线，a－d 4个子图所示是各种建模方法下，分别以测试集中样本N素含量真实值和模型预测值为横、纵轴画出的散点图。各种模型均通过数据变换、调整参数和PCA降低到合适维度以达到最优模型预测状态。图2中，BPANN建模过程中，根据模型最佳预测性能的原则，选取98%的能量比用PCA对原始数据降维，即数据降为3维。PLS根据MSE最小的优化原则，在(0, min(训练样本数−1，样本维度)]的半闭区间内找到最优降维后的维度。根据试验结果，图中寻找到最优参数ncomp=9（Matlab R2011b，plsregress命令中的参数），用原始光谱可以获得最佳的PLS预测模型；SMLR建模亦采用PCA降维，保持99.9%的能量，用原始反射光谱取得最佳模型性能。

试验结果表明，本文提出的方法预测性能最佳，其模型决定系数最高，MSE 远远低于其他 3 个模型。相比之下，取得较好结果的是 BP 神经网络模型，其 R^2 为 0.943927，MSE 为 0.1627。4 个模型中预测性能最差的是 SMLR，其 R^2 仅为 0.74807，MSE 高达 0.862371。PLS 性能在 4 种多元回归分析法中位居第 3，这从其在测试集上的 MSE 和 R^2 的取值可以看出。试验表明本文方法的有效性。

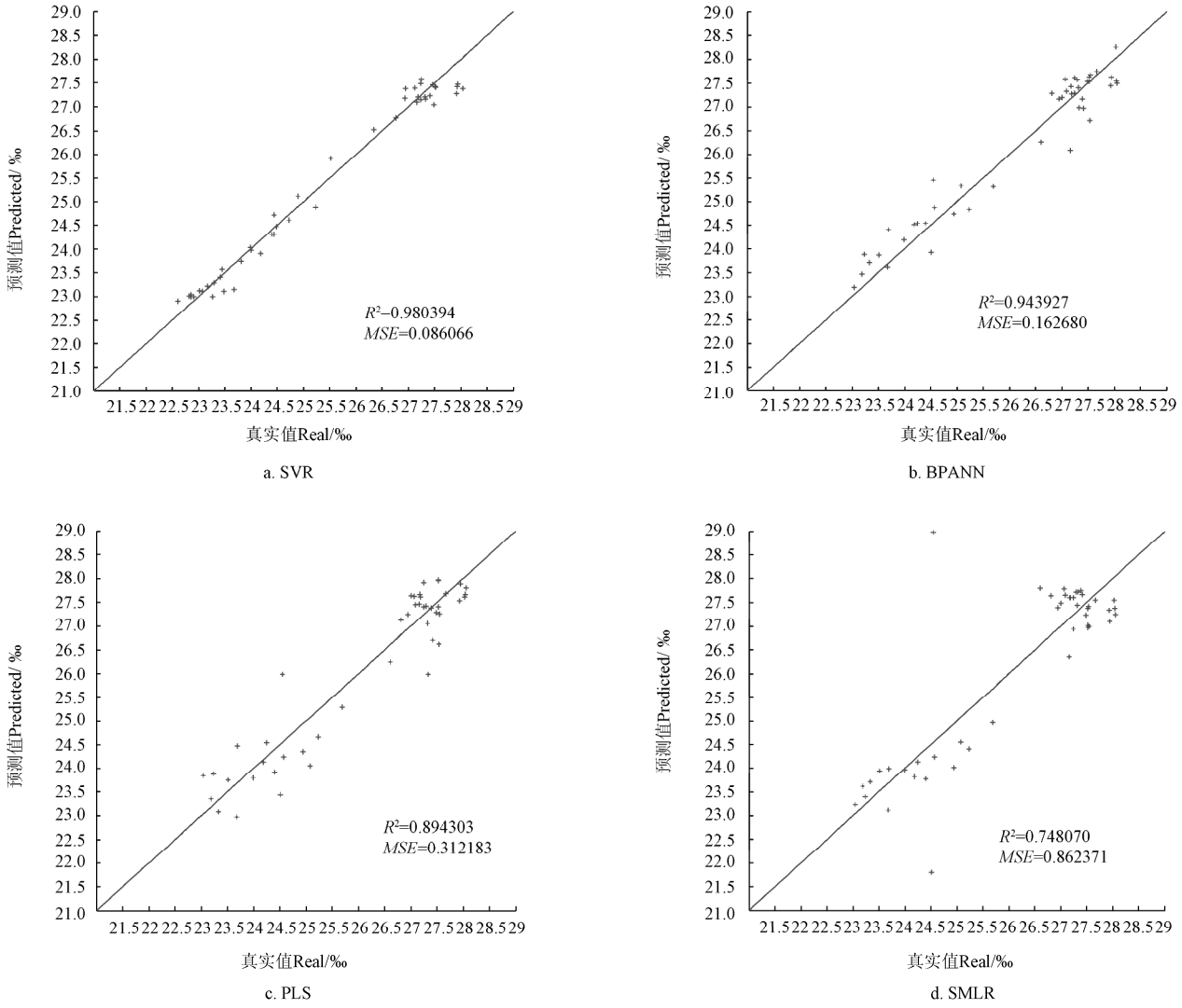

图 2 不同建模方法下氮素含量实测值和预测值的比较

Fig.2 Measured versus estimated nitrogen content using different modeling method

3 讨 论

农业环境时空变异性大，复杂多变，环境因子严重影响了叶片中 N 含量，而各种因素间还可能会复合作用。论文虽然在试验过程中已经按照农业学科试验处理方法做了一定简单化与理想化，但毕竟样本培植等试验过程都在实际果园环境下完成，导致样本数据的复杂化。因此，数据处理和建模方法要充分考虑数据中噪声的客观存在，设计性能更加鲁棒，容噪性能更好的分析手段。

论文试验过程中采集了 4 批数据，但建模分析过程中只采用了 2 批，主要考虑规避 SVR 回归分析算法出现模型过拟现象。在后续研究过程中，应进一步分析其它批次数据的噪声情况，并用不同的 2 批次数据组合进行建模分析，保证更加充分的试验力度。

通常情况下，在可见光波段范围，随着氮含量的增加，柑橘叶片反射率降低；而在近红外波段范围，随着氮含量的增加，反射率增加。论文图 1 出现了局部与常规规律不太一致的现象。这表明农业科研中大田试验与果园试验的常规规律体现方式与理想试验中体现方式有所不同。另一方面，117 株柑橘树样本在 1 年期完整生长全过程的规范化种植过程中，除施用氮肥，还施用了磷肥，钾肥，多因子复合出现了协同或胁迫作用，也是可能的原因之一。

4 结 论

1）用原始高光谱反射数据作为柑橘树样本的矢量表达，在对数据进行 PCA 降维处理的基础上建立支持矢量回归分析（SVR）模型，其模型决定系数高达 0.9730，MSE 仅为 0.090。

2）当 SVR 回归分析法选取径向基函数（RBF）、PCA 维度根据能量比为 99.9%选择独立主成分数，

模型对 N 素含量预测性能最佳且最稳定。

3）与其他主流多元回归分析的比较试验表明，PCA 加 SVR 的建模方法更有效，明显优于 BP 神经网络、偏最小二乘法 PLS 和逐步线性回归 SMLR 法。

［参　考　文　献］

[1] Dennis L Wright, V Philip Rasmussen, R Douglas Ramsey, et al. Baker. Canopy Reflectance Estimation of Wheat Nitrogen Content for Grain Protein Management[J]. GIScience and Remote Sensing, 2004, 41(4): 287－300.

[2] Thenkabail P S, Smith R B, De Pauw E. Hyperspectral vegetation indices and their relationships with agricultural crop characteristics[J]. Remote Sens. Environ. 2000, 71(2): 158－182.

[3] Daniela Stroppiana, Mirco Boschetti, Pietro Alessandro Brivio Stefano Bocchi. Estimation of Plant Nitrogen Concentration in paddy rice from field canopy spectra[J]. Rivista italiana di Telerilevamento, 2009, 41(1): 45－57.

[4] 薛利红，曹卫星，罗卫红，等. 基于冠层反射光谱的水稻群体叶片氮素状况监测[J]. 中国农业科学，2003，36(7)：807－812.
Xue Lihong, Cao Weixing, Luo Weihong, et al. Diagnosis of nitrogen status in rice leaves with the canopy spectral reflectance[J]. Scientia Agricultura Sinica, 2003, 36(7): 807－812. (in Chinese with English abstract)

[5] 蒋金豹，陈云浩，黄文江，等. 条锈病胁迫下冬小麦冠层叶片氮素含量的高光谱估测模型[J]. 农业工程学报，2008，24(1)：35－39.
Jiang Jinbao, Chen Yunhao, Huang Wenjiang, et al. Hyperspectral estimation models for LTN content of winter wheat canopy under stripe rust stress[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2008, 24(1): 35－39. (in Chinese with English abstract)

[6] 张俊华，张佳宝. 不同生育期冬小麦光谱特征对叶绿素和氮素的响应研究[J]. 土壤通报，2008，39(3)：586－592.
Zhang Junhua, Zhang Jiabao. Response of winter wheat spectr al reflectance to leaf chlorophyll, total nitrogen of above ground[J], Chinese Journal of Soil Science, 2008, 39(3): 586－592. (in Chinese with English abstract)

[7] Yan Zhu, Wei Wang, Xiao Yao. Estimating Leaf Nitrogen Concentration (LNC) of Cereal Crops with Hyperspectral Data[M]. Hyperspectral Remote Sensing of Vegetation Alfredo Thenkabail, Prasad S. Lyon and John G. Huete CRC Press 2011: 187－206.

[8] Yuan Wang, Fumin Wang, Jingfeng Huang, et al. Validation of artificial neural network techniques in the estimation of nitrogen concentration in rape using canopy hyperspectral reflectance data[J]. International Journal of Remote Sensing, 2009, 30(17): 4493－4505.

[9] Valentina Ulissi, Francesca Antonucci, Paolo Benincasa, et al. Nitrogen concentration estimation in tomato leaves by VIS-NIR non-destructive spectroscopy[J]. Sensors, 2011, 11: 6411－6424.

[10] Hansena P M, Schjoerring J K. Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indices and partial least squares regression[J]. Remote Sensing of Environment, 2003, 86(4): 542－553.

[11] Menesatti P, Antonucci F, Pallottino F, et al. Estimation of plant nutritional status by Vis-NIR spectrophotometric analysis on orange leaves[J]. Biosystems Engineering, 2010, 105(4): 448－454.

[12] 李震，洪添胜，曾洁媚. 基于柑橘树冠层光谱信息的土壤营养元素含量预测[J]. 农业工程学报，2011.
Li Zhen, Hong Tiansheng, Zeng Jiemei. Soil nutrient content estimation based on citrus tree canopy spectral information[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2011. (in Chinese with English abstract)

[13] 易时来，邓烈，何绍兰，等. 锦橙叶片钾含量光谱监测模型研究[J]. 中国农业科学，2010，43(4)：780－786.
Yi Shilai, Deng Lie, He Shaolan, et al. A spectrum based models for monitoring leaf potassium content of citrus sinensis (L) cv. jincheng orange[J]. Scientia Agricultura Sinica, 2010, 43(4): 780－786. (in Chinese with English abstract)

[14] Abdi H, Williams L J. Principal Component Analysis[J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2010, 2: 433－459.

[15] Vapnik V N. Statistical Learning Theory[M]. New York, Wiley, 1998: 1－50.

[16] Pearson K. On lines and planes of closest fit to systems of points in space[J]. Philosophical Magazine, 1901, 2(6): 559－572.

[17] Wold S, Sjostrom M, Erikssonn L. PLS-regression: a basic tool of chemometrics[J]. Chemometrics and Intelligent Laboratory System, 2001, 58: 109－130.

[18] Stuart Russell, Peter Norvig. Artificial Intelligence A Modern Approach[M]. 2009.

[19] Arthur Earl Bryson, YuChi Ho. Applied optimal control: optimization, estimation, and control[M]. Blaisdell Publishing Company or Xerox College Publishing. 1969: 481.

[20] Hocking R R. The analysis and selection of variables in linear regression[J]. Biometrics, 1976, 32(1): 1－49.

[21] Draper N, Smith H. Applied Regression Analysis [M]. 2d Edition, New York: John Wiley and Sons, Inc. 1981.

[22] Julius B. Cohen. Practical Organic Chemistry[M]. DISC, Indian Institute of Science, Bangalore, 2003.

[23] 李民赞. 光谱分析技术及其应用[M]. 北京：科学出版社，2006.

[24] Li Minzan. Spectroscopy and its Applications[M]. Beijing: Science Press, 2006.

Multiple regression analysis of citrus leaf nitrogen content using hyperspectral technology

Huang Shuangping[1,2,3], Hong Tiansheng[1,2,3], Yue Xuejun[1,2,3,4※], Wu Weibin[1,2,3] , Cai Kun[1,2,3], Xu Xing[1,2,3]

(1. *Key Laboratory of Key Technology on Agricultural Machine and Equipment* (*South China Agricultural University*), *Ministry of Education*, *Guangzhou* 510642, *China*; 2. *Division of Citrus Machinery, China Agriculture Research System*, *Guangzhou* 510642, *China*; 3. *College of Engineering, South China Agricultural University, Guangzhou* 510642, *China*; 4. *Faculty of Engineering and Surveying, University of Southern Queensland, Toowoomba QLD* 4350, *Australia*.)

Abstract: In order to evaluate the nitrogenous status of citrus trees, non-destructively, accurately and rapidly, the modeling of the nitrogen (N) content prediction based on the reflectance spectra is studied in this paper. Field experiments were conducted on 117 planted Luogang citrus trees in the Crab Village of Guangzhou. The citrus trees were divided into several groups and 1-year standardized management was performed on them. Nitrogenous fertilizer was applied to the citrus trees only during four phenological periods in the year, and each group was treated with various levels of N-fertilization in order to cultivate differentiation samples with varied nitrogenous content. 15 days after each fertilization, fresh and healthy citrus leaves were collected to gather training samples from different growth stages. Hyper-spectrometer ASD FieldSpec was used to detect spectral reflectance while the Kjeldahl method was used to measure the N-content of citrus leaves from the same batch. In this way, each sample is described as an instance-label pair, where a multi-variable vector was used as the descriptor and the ground truth of the nitrogen level was used as the label. The collected samples were used to construct a large-scale dataset，80% of which were used as the train set and the remaining 20% were used as the test set. PCA (Principle Component Analysis) was applied to the original vectors for dimension reduction and noise removal and SVR (Support Vector Regression) was adopted to build the regression analysis model for predicting the nitrogen level of the citrus trees. The model relied on a training set and was created by mapping the multi-variable vectors to the related ground truths label through SVR. The test set was used to evaluate the performance of the model. The experiment on the test set resulted in reaching a square correlation coefficient (R2) of 0.9730, a mean relative error of 0.9033%, and a mean square error (MSE) of 0.090343. Conclusions can be drawn from the experimental results: First, compared with various deformations of spectral data, e.g. first derivative spectrum, second derivative spectrum, reciprocal spectrum, logarithmic spectrum, logarithm of reciprocal spectrum, the original high spectral reflectance data, as the vector-descriptor of the samples, achieved the best experimental result when using the approach in this paper. Second, when the Radial Basis Function (RBF) is used as the kernel for SVR and PCA determines the principal components with the cumulative contribution rate set to 99.9%, the model will achieve the best performance and be the most robust. Third, comparative experiments between our method and other mainstream multivariate regression analysis algorithms demonstrate the validity of using SVR and PCA to do modeling. Experimental results show our method is obviously superior to Partial Least Squares (PLS), Back Propagation (BP) and Stepwise Multiple Linear Regression (SMLR). Finally, using SVR to build the regression model based on PCA-processed data successfully achieved the ideal performance index, which indicates the effectiveness of the proposed method and provides a theoretical basis for the applications of high spectral reflectance in non-destructive nitrogen level detection.

Key words: nitrogen, regression analysis, hyperspectral, citrus tree leaves, SVR