

基于可见-近红外光谱比较主成分回归、偏最小二乘回归和反向传播神经网络对土壤氮的预测研究

李 硕¹ 汪善勤^{1,2} 张美琴¹

(¹ 华中农业大学资源与环境学院, 湖北 武汉 430070
² 中国科学院土壤与农业可持续发展国家重点实验室, 江苏 南京 210008)

摘要 建模方法是影响可见-近红外光谱定量结果的主要因素之一。在 470~1000 nm 波段的 12 个土壤剖面对 48 个剖面样经过风干、研磨、过筛后进行光谱采集。经一阶微分变换及 Savitzky-Golay 平滑处理后, 分别应用主成分回归(PCR)、偏最小二乘回归(PLSR)和反向传播神经网络(BPNN)3 种方法建立土壤全氮(TN)的定量模型。PCR 与 PLSR 两线性模型的决定系数(R^2)分别为 0.74 和 0.8, 其剩余预测偏差(RPD)分别为 2.23 和 2.22, 但两模型仅能用于 TN 的粗略估计。由 PCR 提供主成分数, PLSR 提供潜变量(LV)数分别作为 BPNN 的输入所构建的两个非线性模型均明显优于线性模型 PCR 和 PLSR。其中以 4 个 LV 作为输入的 BPNN-LV 模型预测性能最优, R^2 以及 RPD 分别达到 0.9 和 3.11。实验结果表明, 提取可见-近红外光谱的 PLSR LV 因子作为 BPNN 的输入, 所建定量模型可用于土壤氮纵向时空分布的快速准确预测。

关键词 光谱学; 土壤光谱; 可见近红外; 主成分回归; 偏最小二乘回归; 反向传播神经网络

中图分类号 TP79; S151.9 **文献标识码** A **doi:** 10.3788/AOS201232.0830001

Comparison Among Principal Component Regression, Partial Least Squares Regression and Back Propagation Neural Network for Prediction of Soil Nitrogen with Visible-Near Infrared Spectroscopy

Li Shuo¹ Wang Shanqin^{1,2} Zhang Meiqin¹

(¹ College of Resources and Environment, Huazhong Agricultural University, Wuhan, Hubei 430070, China
² State Key Laboratory of Soil and Sustainable Agriculture, Chinese Academy of Sciences, Nanjing, Jiangsu 210008, China)

Abstract The selection of modeling method is one of the main factors influencing the quantitative accuracy with visible-near infrared (Vis-NIR) spectroscopy. We compare the performance of three calibrations methods, i. e., principal component regression (PCR), partial least squares regression (PLSR), and back propagation neural network (BPNN) based on Vis-NIR reflectance spectra of soil total nitrogen (TN) quantitative forecast results. Covered in the 470~1000 nm wavelength range, spectroscopy of 48 soil samples selected from 12 profiles are air-dried, screened and mused, then processed by the first order derivative and Savitzky-Golay smoothing methods. Leave-one-out cross validation is also adopted to determine the optimal factor numbers. The results indicate that PCR and PLSR linear models are able to meet general prediction and with little difference, where coefficients of determination (R^2) are 0.74 and 0.8, respectively, and residual predictive deviation (RPDs) are 2.23 and 2.22. The two nonlinear models built by BPNN in combination with PCR and PLSR, respectively, are superior to the linear models of PCR and PLSR in the precision of prediction. BPNN, principal components (PCs) whose input is the PCs resulted from the PCR, while the BPNN latent variables (LVs) whose input is the first 4 LV results obtained from PLSR has the best performance ($R^2 = 0.9$, RPD is 3.11). It is recommended to adopt BPNN-LV model to rapidly

收稿日期: 2012-02-14; 收到修改稿日期: 2012-03-14

基金项目: 国家自然科学基金(40801082)资助课题。

作者简介: 李 硕(1983—), 男, 博士研究生, 主要从事土壤高光谱定量遥感方面的研究。

E-mail: shuoguoguo@webmail.hzau.edu.cn

导师简介: 汪善勤(1974—), 男, 副教授, 主要从事农业地理遥感信息处理与应用等方面的研究。

E-mail: sqwang@mail.hzau.edu.cn

predict the vertical spatial and temporal distribution of TN with Vis-NIR spectroscopy.

Key words spectroscopy; soil spectroscopy; visible-near infrared; principal component regression; partial least squares regression; back propagation neural network

OCIS codes 300.6340; 300.6490; 300.6550; 070.4790; 120.5700

1 引 言

氮是土壤肥力的重要指标,传统的土壤全氮(TN)测试方法费时、繁琐、成本高,难以满足快速监测土壤氮含量的需求。近年来利用光谱测定技术快速、简便、无损地对样品进行定量分析,已在各个领域广泛应用^[1~3]。

国内外很多相关研究表明漫反射光谱可用于土壤氮的定量预测。Chang 等^[4]利用主成分回归(PCR)引入 7 个主成分(PC)数建立了 400~2500 nm 土壤光谱 TN 的预测模型,其决定系数 R^2 为 0.85,剩余预测偏差(RPD)为 2.5; Kusumo 等^[5]对 350~2500 nm 反射光谱建立了土壤氮的偏最小二乘回归(PLSR)预测模型, $R^2 = 0.81$, RPD 为 2.32; 郑立华等^[6]使用 PCR 提供的 PC 作为反向传播神经网络(BPNN)的输入,建立了基于 833~2500 nm 近红外(NIR)光谱的土壤 TN 含量分析模型,预测精度 R 为 0.808,模型精度有待提高。虽然线性方法 PCR 和 PLSR 在光谱定量分析中较常用,但非线性方法 BPNN 在可见-近红外(Vis-NIR)反射光谱定量估计土壤氮方面的应用并不多见。另外,氮含量在各土壤层间差异显著,随着土壤层次的加深含量逐步降低^[7]。从国内外研究现状来看,大都选用深 0~15 cm 或 0~20 cm 的耕层土样^[8~11],鲜有针对剖面样的相关报道。

本文以 0~100 cm 剖面样为研究对象,基于 Vis-NIR 反射光谱建立 PCR、PLSR 和 BPNN 三种方法对土壤 TN 的定量预测模型。其中,由 PCR 提供 PC 数,PLSR 提供潜变量(LV)数分别作为

BPNN 的输入,建立反向传播神经网络-主成分(BPNN-PC)和反向传播神经网络-潜变量(BPNN-LV)预测 TN 模型。比较分析了以 PCR、PLSR 为代表的线性模型和以 BPNN 为代表的非线性模型预测土壤 TN 含量的精度,以期快速估算土壤氮纵向时空分布提供参考。

2 材料与方 法

2.1 土样的采集与制备

供试土样来自湖北省潜江后湖农场的 12 个土壤剖面,土壤类型为潮土。剖面深达 1 m,在同一采样区域按土壤层次采集 3~5 个点状样品,共计 48 个。所有土样经风干、磨碎、过直径 2 mm 筛后再过 100 目筛并分成两份,分别用于光谱测定和化学分析。

2.2 仪器设备

供试土样的光谱均使用成像光谱仪采集。仪器为美国 Headwall Photonics 公司 HS-VNIR N 系列,配备 150 W 光纤式卤素灯及 25 μm 狭缝,35 mm 焦距镜头的视场角为 13.1°。电荷耦合器件(CCD)为 1004 pixel \times 1002 pixel,其波长 400~1000 nm,采样间隔 0.8 nm,共 753 个波段。拍摄平台为自行设计的三轴向土壤高光谱成像实验台^[12]。

2.3 土样的化学测定

采用凯式定氮法(GB/T 5009.5-2010)得到土样的 TN 含量。按照含量梯度法以 2:1 的比例从总样本中挑选 32 个用于建模,其余 16 个用于检验。所有土样 TN 统计特征如表 1 所示。

表 1 土壤全氮含量的统计特征

Table 1 Statistics of soil TN content

Data set	Number of samples	Range	Mean	Standard deviation	Coefficient of variation /%
Calibration	32	0.25~2.13	0.96	0.46	48
Validation	16	0.4~1.85	0.95	0.44	46
All	48	0.25~2.13	0.96	0.44	46

2.4 土样光谱测定及预处理

将土样填满直径 56 mm、深 9 mm 的铝盒。在光谱测定前,为减弱成像光谱仪暗电流和光源对数据的影响,分别采用遮盖镜头和采集标准参考白板图像的方法获取黑、白影像。经多次实验最终确定物距为 0.8 m,光源垂直距样品 0.23 m。光谱仪预

热 30 min 后在暗室条件下进行拍摄。数据采集软件为 Hyperspec-N。

为保留高光谱影像中的土壤部分,以每个铝盒中心为圆心对各样品逐一进行感兴趣区(ROI)选取,每个 ROI 约 24000 pixel,以其平均值作为该土样的反射光谱。在所有 753 个波段范围内首端信噪

比较低(图 1),结合前期预实验的分析,最终确定保留 470~1000 nm(图 2)进行后续分析建模。

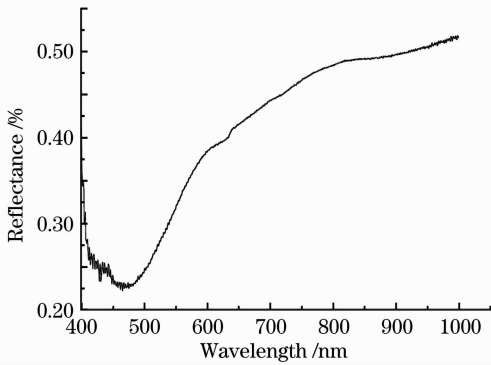


图 1 原始 Vis-NIR 土壤反射光谱

Fig. 1 Original Vis-NIR reflection spectrum

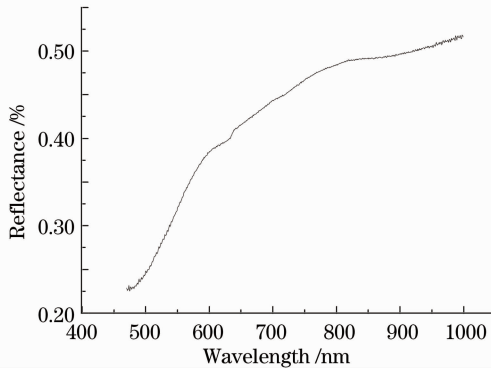


图 2 保留的 Vis-NIR 土壤反射光谱

Fig. 2 Remained Vis-NIR reflection spectrum

2.5 建模方法及评价参数

采用 PCR、PLSR 和 BPNN 算法分别建立 TN 定量模型,并分析比较模型精度,各算法原理为:1) PCR 先利用主成分分析(PCA)将变量重新组合成一组新的互相无关的综合变量即 PC,从中取出几个较少的 PC 尽可能多地反映原有变量。再通过关联光谱与化学值,建立二者之间的校正模型并预测光谱以获得相应化学值。2) PLSR 是在普通多元回归的基础上,借鉴 PCA 和典型相关分析的思想,很好地解决了自变量间多重共线性问题,成为光谱分析中建立线性定量校正模型的通用方法^[13],在预测变量高度相关或预测变量数超过样本数时特别有效。3) BPNN 是一种由输入层、隐含层和输出层组成的前向网络,学习过程由正向传播及反向训练组成。正向传播时,输入的样本以 S 型传递函数从输入层经各神经元按层进行处理,通过所有的隐含层再传向输出层。输出层给出预测值并与期望值进行比较。如果现行输出不等于期望输出,则进入反向

训练过程;反向训练时,又把误差信号按误差最大的方向沿原来正向传播的通路反向传回,同时对每层各个神经元间的权系数进行调整,从而使误差信号趋于最小,满足预先设定的限差要求。4)采用 R^2 , 预测均方根误差(RMSEP)和 RPD 对模型进行评价。预测能力好的模型应同时具有较高的 R^2 ,较小的 RMSEP 和较大的 RPD。

3 结果与分析

3.1 模型预测结果

光谱数据采用先做一阶微分变换再做 Savitzky-Golay 平滑(2 阶 7 窗口)处理。对校正集数据分别建立 PCR、PLSR 模型,为避免过度拟合,通过留一法进行交叉验证,确定合适的 PC 数和 LV 数。校正模型建好后,对验证集进行检验来完成模型最终的评定。

BPNN 的输入层可以是来自 PCR 的 PC,也可以是来自 PLSR 的 LV,这样既能减少计算量又可提高模型的稳健性^[14]。本研究中作为输入层的 PC 及 LV 的数量均由 PCR 与 PLSR 建模时留一法确定。48 个样本的 70% 做训练,30% 做检验,设置 10 个隐含层。建模结果如表 2 所示。

表 2 四种模型对土壤全氮含量的预测结果

Table 2 Prediction of soil TN content by four models

Model	Factor	R^2	RMSEP	RPD
PCR	5	0.74	0.1976	2.23
PLSR	4	0.80	0.1982	2.22
BPNN-PC	5	0.88	0.1840	2.39
BPNN-LV	4	0.90	0.1414	3.11

当 RPD 小于 2 时,模型不可用;当 RPD 取值在 2.0~2.5 时,模型可用来做粗略的定量预测;当 RPD 取值在 2.5~3.0 时,能够很好地进行变量预测,当 RPD 大于 3.0 时,模型非常优异^[15]。四种模型的散点图如图 3 所示。

3.2 模型的比较

线性模型中 PCR 与 PLSR 的预测精度差异不大,虽然后者的 R^2 为 0.8,略高于前者的 0.74,但前者的 RMSEP 为 0.1982 $\text{g} \cdot \text{kg}^{-1}$ 值稍大于后者的 0.1976 $\text{g} \cdot \text{kg}^{-1}$,而两模型的 RPD 值分别为 2.23 和 2.22,总体而言两者均可用于粗略的土壤 TN 定量估计;而 BPNN 与 PCR、PLSR 相比,既提高了决定系数 R^2 ,同时又减小了 RMSEP 并增大了 RPD,预测精度明显提升。也就是说,对于 TN 的预测,非线性模型 BPNN 优于线性模型 PCR、PLSR。其中以 5 个 PC 作

为输入的 BPNN-PC 模型较 PCR、PLSR 的 R^2 提高到了 0.88 且 RMSEP 降至 $0.184 \text{ g} \cdot \text{kg}^{-1}$, 但其 RPD 为 2.39, 虽有提高但仍旧属于粗略定量范畴。以 4 个 LV 作为输入的 BPNN-LV 在所有模型的预测结果中

表现最好, 对土壤 TN 预测精度最高, R^2 , RMSEP 和 RPD 分别为 0.9, $0.1414 \text{ g} \cdot \text{kg}^{-1}$ 和 3.11, 该模型拥有最大的 R^2 , 最小的 RMSEP 以及最大的 RPD。

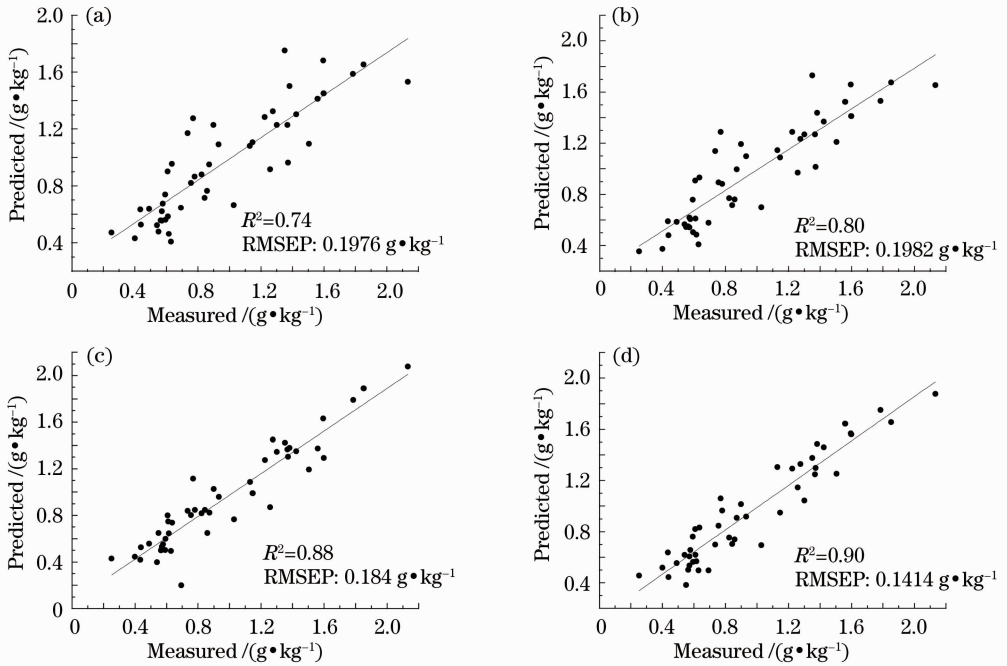


图 3 (a) PCR, (b) PLSR, (c) BPNN-PC 和 (d) BPNN-LV 预测模型的散点图

Fig. 3 Scatter plots of (a) PCR, (b) PLSR, (c) BPNN-PC, and (d) BPNN-LV models

3.3 回归系数分析

图 4 是 470~1000 nm 全波段范围 PCR 与 PLSR 模型预测 TN 的回归系数, 虚线为零相关线。图中呈现了几个明显的峰, 通常认为这些区域对模型有重要影响。观察曲线的总体走势可以看出, 两线性模型的表现相似, 这与建模结果相吻合。两模型在绿光区的 500 nm 及 525 nm 处均有明显的反向峰, 同时在红光区 650 nm 附近均有微弱的反向峰, 而红光区的 620 nm 处正向峰为 PLSR 模型所

独有。文献[16]指出 400~700 nm 波段与土壤母质对于作物、植株以及矿物的吸收联系密切。PLSR 在 750 nm 处的反向峰较 PCR 明显, 该波段被认为与土壤腐殖质以及色素密切相关^[17,18]。

在 NIR 区域 (800~1000 nm), 两模型于 850 nm 处均有一显著正向峰, 该波段受 C-H-O 双键影响^[19]。在 Mouazen 等^[20]对 Mg 所建立的 PLSR 预测模型的回归系数分析中, 在同样的 850 nm 区域有一明显的反向峰, 表明 TN 与 Mg 在该波段可能是协同变量。950 nm 是公认的水分吸收带, 在 950~1000 nm, PCR 有两正向峰而 PLSR 为两反向峰, 但都偏低。土样经过标准制样处理已去除了离子水, 此处表明分子间结合水对模型有一定的贡献。总之, 通过分析 470~1000 nm 全波段的回归系数曲线, 500、520、850 nm 附近与 TN 呈高度相关。PLSR 模型整体表现较 PCR 更加突出, 可能这是造成 BPNN-LV 模型优于 BPNN-PC 模型的原因之一。

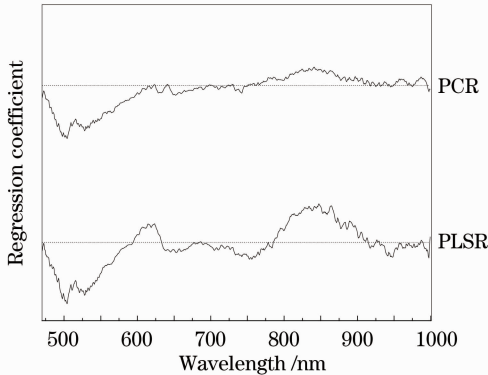


图 4 基于 PCR 和 PLSR 的全波段回归系数

Fig. 4 Regression coefficient distributions over the entire wavelength range based on PCR and PLSR

4 结 论

以 0~100 cm 剖面土样为研究对象, 基于 470~

1000 nm 范围的 Vis-NIR 反射光谱经一阶微分变换和 Savitzky-Golay 平滑后,通过 PCR、PLSR 和 BPNN 三种方法对土壤 TN 含量的预测结果对比分析认为对 0~100 cm 剖面样建立 TN 的线性和非线性光谱定量模型是可行的。非线性模型 BPNN 优于线性模型 PCR 和 PLSR,其中以 BPNN-LV 模型预测精度最高(R^2 为 0.9,RPD 为 3.11),线性模型 PCR 与 PLSR 的预测精度则有待提高。对于剖面样,光谱与 TN 之间存在的非线性关系可能强于线性关系。通过回归系数分析,500、525、850 nm 附近对于 TN 的预测有明显影响,但深层次原因还有待后期深入研究。TN 与 Mg 在 NIR 光谱定量估计中是否有相关联系同样值得考虑。

参 考 文 献

- Sun Guangming, Liu Fei, Zhang Fan *et al.*. Fast determination of proline in herbicide-stressed oilseed rape leaves based on near infrared spectroscopy[J]. *Acta Optica Sinica*, 2010, **30**(4): 1192~1196
孙光明,刘飞,张帆等.基于近红外光谱技术检测除草剂胁迫下油菜叶片中脯氨酸含量的方法[J].*光学学报*,2010,**30**(4): 1192~1196
- Guo Weiliang, Wang Dan, Song Jia *et al.*. Simultaneous and rapid quantitative analysis of four components in *Cordyceps militaris* mycelium powder using near infrared spectroscopy[J]. *Acta Optica Sinica*, 2011, **31**(2): 0230002
郭伟良,王丹,宋佳等.近红外光谱法同时快速定量分析蛹虫草菌丝体中4种有效成分[J].*光学学报*,2011,**31**(2): 0230002
- Liu Yande, Chen Xingmiao, Ouyang Aiguo. Non-destructive measurement of soluble solid content in Gannan navel oranges by visible/near-infrared spectroscopy[J]. *Acta Optica Sinica*, 2008, **28**(3): 478~481
刘燕德,陈兴苗,欧阳爱国.可见/近红外光谱法无损检测赣南脐橙可溶性固形物[J].*光学学报*,2008,**28**(3): 478~481
- C.-W. Chang, D. A. Laird, M. J. Matuschek *et al.*. Near-infrared reflectance spectroscopy: principal components regression analyses of soil properties[J]. *Soil Science Society of America Journal*, 2001, **65**(2): 480~490
- B. H. Kusumo, M. J. Hedley, C. B. Hedley *et al.*. Measuring carbon dynamics in field soils using soil spectral reflectance: prediction of maize root density, soil organic carbon and nitrogen content[J]. *Plant and Soil*, 2011, **338**(1-2): 233~245
- Zheng Lihua, Li Minzan, Pan Luan *et al.*. Estimation of soil organic matter and soil total nitrogen based on NIR spectroscopy and BP neural network[J]. *Spectroscopy and Spectral Analysis*, 2008, **28**(5): 1160~1164
郑立华,李民赞,潘婵等.基于近红外光谱技术的土壤参数 BP 神经网络预测[J].*光谱学与光谱分析*,2008,**28**(5): 1160~1164
- Gao Yingzhi, Wang Shiping, Han Xingguo *et al.*. Soil nitrogen regime and the relationship between aboveground green phytobiomass and soil nitrogen fractions at different stocking rates in the Xilin River Basin, Inner Mongolia [J]. *Acta Phytoecologica Sinica*, 2004, **28**(3): 285~293
高英志,汪诗平,韩兴国等.退化草地恢复过程中土壤氮素状况以及与植被地上绿色生物量形成关系的研究[J].*植物生态学报*,2004,**28**(3): 285~293
- Lu Peng, Wei Zhiqiang, Niu Zheng. Estimate of soil attributes using the method of special band and reflectance inflection difference [J]. *Spectroscopy and Spectral Analysis*, 2009, **29**(3): 716~721
路鹏,魏志强,牛铮.应用特征波段和反射变形差的方法进行土壤属性估算[J].*光谱学与光谱分析*,2009,**29**(3): 716~721
- Lu Yanli, Bai Youlu, Wang Lei *et al.*. Determination for total nitrogen content in black soil using hyperspectral data [J]. *Transactions of the Chinese Society of Agricultural Engineering*, 2010, **26**(1): 256~261
卢艳丽,白由路,王磊等.黑土土壤中全氮含量的高光谱预测分析[J].*农业工程学报*,2010,**26**(1): 256~261
- Xu Yongming, Lin Qizhong, Huang Xiuhua *et al.*. Experimental study on total nitrogen concentration in soil by VNIR reflectance spectrum [J]. *Geography and Geo-Information Science*, 2005, **21**(1): 19~22
徐永明,蔺启忠,黄秀华等.利用可见光/近红外反射光谱估算土壤总氮含量的实验研究[J].*地理与地理信息科学*,2005,**21**(1): 19~22
- R. A. V. Rossel, D. J. J. Walvoort, A. B. McBratney *et al.*. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties[J]. *Geoderma*, 2006, **131**(1-2): 59~75
- 高云,汪善勤,李小昱等.三轴向土壤高光谱成像实验台:中国,200920087238.7 [P]. 2010-07-14
- 褚小立,许育鹏,陆婉珍.偏最小二乘法方法在光谱定性分析中的应用研究[J].*现代仪器*,2007,(5): 13~15
- L. J. Janik, D. Cozzolino, R. Damberg *et al.*. The prediction of total anthocyanin concentration in red-grape homogenates using visible-near-infrared spectroscopy and artificial neural networks[J]. *Analytica Chimica Acta*, 2007, **594**(1): 107~118
- R. Zornoza, C. Guerrero, J. Mataix-Solera *et al.*. Near infrared spectroscopy for determination of various physical, chemical and biochemical properties in Mediterranean soils[J]. *Soil Biology & Biochemistry*, 2008, **40**(7): 1923~1930
- D. Cozzolino, A. Moron. Potential of near-infrared reflectance spectroscopy and chemometrics to predict soil organic carbon fractions[J]. *Soil & Tillage Research*, 2006, **85**(1-2): 78~85
- C. S. T. Daughtry. Discriminating crop residues from soil by shortwave infrared reflectance [J]. *Agronomy Journal*, 2001, **93**(1): 125~131
- K. P. Fabrizzi, A. Moron, F. O. Garcia. Soil carbon and nitrogen organic fractions in degraded versus non-degraded mollisols in Argentina [J]. *Soil Science Society of America Journal*, 2003, **67**(6): 1831~1841
- M. Zimmermann, J. Leifeld, J. Fuhrer. Quantifying soil organic carbon fractions by infrared-spectroscopy [J]. *Soil Biology & Biochemistry*, 2007, **39**(1): 224~231
- A. M. Mouazen, B. Kuang, J. De Baerdemaeker *et al.*. Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy [J]. *Geoderma*, 2010, **158**(1-2): 23~31