基于紫外光谱的水产养殖水质总氮含量快速检测研究

李鑫星[1]，周 婧[1]，唐 红[2]，孙龙清[1]，曹霞敏[3]，张小栓[4*]

1. 中国农业大学信息与电气工程学院食品质量与安全北京实验室，北京 100083
2. 中国农业大学烟台研究院，山东 烟台 264000
3. 苏州大学基础医学与生物科学学院，江苏 苏州 215200
4. 中国农业大学工学院，北京 100083

摘 要 应用紫外(Ultraviolet，UV)光谱技术对水产养殖水质总氮含量进行快速检测。为了消除各种系统误差与偶然误差对模型预测性能造成的影响，将88个水样的总氮浓度实测值数据和光谱吸光度数据作为原始数据，将模型建立分为样本集划分、数据预处理、特征波段提取、模型选择与LV数量选择5个阶段，以求达到最优预测效果，其中前4个阶段分别使用多种方法进行比较。结果证明每个阶段都是必不可少的，只有通过对比其优劣才能找到最适合总氮含量测定的建模过程及方法。首先用浓度梯度(CG)法对原始数据进行相同的样本集划分处理，然后在此基础上分别建立主成分回归(PCR)、逐步回归(SR)和偏最小二乘回归(PLSR)三种模型，选择预测效果最好的PLSR作为本文的预测模型。PLSR的建模效果会在很大程度上受到潜在变量(LVs)数量的影响，通常选取模型预测均方根误差RMSEP值最小时所对应的LV个数为最优LV个数。其次，选用CG法、随机抽样(RS)法、Kennard Stone(KS)法和SPXY法4种样本集划分算法对样本进行处理，并对所建立的PLSR模型预测效果进行比较，最终选择SPXY算法作为最优样本划分算法。然后在对样本集进行SPXY法划分的基础上，运用多种预处理算法对光谱吸光度数据进行预处理，包括小波变换(WT)、一阶导数法(Der1st)与二阶导数法(Der2nd)三种单一算法和小波变换与两种导数法的组合预处理算法WT-Der1st和WT-Der2nd。然后在预处理的基础上分别使用连续投影变换(SPA)和逐步回归(SR)两种特征波段提取方法，对比可知，SPA特征提取方法比SR的提取效率高且建模效果好。SPA算法既可以大大地简化模型，又可以在一定程度上提升模型的预测精度。基于WT-Der1st-SPA提取的特征波段为218 nm，与总氮特征波段区间相一致，由此说明该方法比较科学。综合上述建立的10个PLSR模型，考虑到预测精度与模型复杂度2个因素，最终选择基于WT-Der1st-SPA建立的PLSR模型作为最优模型，该模型预测决定系数r^2为0.996，预测均方根误差RMSEP为0.042 mg·L^{-1}。由此可见，所建立的模型预测效果非常好，可以快速准确地测定水体的总氮含量，为实现光谱技术在水产养殖其他水质监测指标的在线检测以及快速测定提供了经验。

关键词 紫外光谱；总氮；小波变换；连续投影变换；潜在变量；偏最小二乘回归

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2020)01-0195-07

引 言

近年来，我国水产养殖在各类养殖行业中占有十分突出的地位，为保障食物供给、改善饮食结构、促进经济增长做出了巨大贡献。水产品的脂肪含量较低，蛋白质的含量较为丰富，深受大众的喜爱，但水产品一旦出现质量安全问题，就可能带来极其严重的后果。水产品质量与水产养殖水环境密切相关，而水中总氮(TN)含量是评价水体受污染严重程度的关键性指标之一[1]。随着经济的迅猛增长，工业废水、生活污水中的含氮物质有可能进入养殖池塘，造成其水质污染；在集约化养殖过程使用的是高蛋白饵料，只有一少部分被鱼虾类等摄入体内转换为蛋白质，更大一部分则是在池塘中残留，造成养殖池塘含氮营养物质水平过高。如果超标排

收稿日期：2018-11-01，**修订日期**：2019-03-07

基金项目：国家重点研发计划项目(2017YFE0111200)资助

作者简介：李鑫星，1983年生，中国农业大学信息与电气工程学院食品质量与安全北京实验室副教授 e-mail：lxxcau@cau.edu.cn

*通讯联系人 e-mail：zhxshuan@cau.edu.cn

放则会破坏水体自身的平衡，导致水体中含氮物质的富集，造成水体中的水藻和浮游生物的快速生长，致使水体中的溶解氧浓度下降，鱼虾类等生物大面积死亡，最终使得水质恶化而不适合于水产养殖。

紫外(Ultraviolet，UV)光谱技术能够实现快速无损检测，且检测成本较低，近年来已在水质检测领域得到广泛应用。郑一力等[2]以金镶玉竹叶片为样本，建立了 4 种竹叶片氮含量高光谱估测模型，最后比较发现基于主成分分析的 BP 神经网络构建的竹叶片氮含量估测模型效果较好。赵进辉等[3]运用小波分析方法对原始光谱数据进行数据预处理，建模效果明显改善。刘思伽等[4]、Wang 等[5]采用 SPA 算法对农产品光谱数据进行特征波段的选择，简化了模型并提升了其预测性能。Chen 等[6]、Guo 等[7]基于 UV 光谱技术分别建立了水质 COD 与重金属离子的 PLS 模型，结果显示其预测值十分接近于真实值。传统的总氮浓度检测方法通常与化学分析方法结合紧密，但是这类方法通常过程繁复，耗时长，不适宜大范围使用。目前，光谱技术在水质 COD 含量和重金属离子的浓度检测方面应用较多，且较为成熟，但在水质氮磷含量的检测方面应用偏少。光谱分析技术是水质监测领域一个十分有前景的应用方向，与传统方法相比，它的操作简易、试剂消耗量小、重复性良好、测定精度较高且检测速度快，非常适合于水质的快速在线检测[8]。本文基于紫外光谱技术，设计出一种快速无损，又相对精确的水产养殖水质总氮含量快速检测方法。

1 实验部分

1.1 仪器设备

本实验采用 UV-2450 UV/Vis 光谱仪，其测定波长范围为 190～900 nm，吸光度的测量范围为－4～5 Abs，光谱分辨率为 1 nm。通过查阅文献并结合水样所测吸光度值绘制的光谱曲线可知，测量 190～350 nm 波段的吸光度即可满足需要。每个样本测量 5 次，计算 5 次的平均值为最后的光谱测量值。本文的光谱数据处理与数据建模软件为 Excel 2010、Matlab R2016a。

图 1 UV-2450 紫外-可见光光谱仪

Fig. 1 UV-2450 spectrograph

1.2 样本采集及总氮含量测定

本实验于 2018 年 3 月到 6 月进行实验数据的采集工作，于 2018 年 4 月到 9 月集中进行实验数据处理工作。实验样本采集地点是苏州大学试验养殖基地。试验水样主要成分为硝酸盐水样总氮含量由专业检测机构检测，水样光谱吸光度数据由实验室光谱仪采集。采集的样本量共 90 个，首先测定 90 个样本的总氮浓度值，然后用光谱仪分别测定其在 190～350 nm 的吸光度值，绘制光谱曲线图。结合样本光谱数据与实测浓度值，剔除 2 个明显不合规律的样本后剩余 88 个样本。

1.3 光谱总氮含量预测模型构建

模型的建立需要经过样本集划分、数据预处理、特征波段提取、LV 数量选取与模型构建 5 个步骤，每个步骤都是必不可少的环节，其处理效果的好坏与否可能会影响到最终模型的预测效果。

1.3.1 样本集划分

样本集划分是依照特定方法把研究所用样本集分成建模集和预测集，以分别用于模型的构建与验证，样本集划分是建立模型的过程中必不可少的一步，其划分是否科学将直接影响模型的预测效果。随机抽样(random sampling，RS)法、浓度梯度(concentration gradient，CG)法、Kennard Stone (KS)算法和 SPXY 算法是常见的 4 种样本划分方法。这 4 种方法各有特色，通过比较不同方法所建立模型预测效果的优劣以确定最适合的方法[9-10]。

1.3.2 光谱数据预处理

为了消除光谱仪在扫描过程中引发的噪声问题，减轻多种外界干扰，并简化数据处理过程，因此对原始光谱进行数据预处理是一项十分有必要的操作。小波变换(wavelet transform，WT)、一阶导数法(first-derivative，Der1st)和二阶导数法(second-derivative，Der2nd)是常见的 3 种光谱数据预处理算法[11]。在某些情况下，小波变换与导数法相结合的方法小波-一阶导数法(WT-Der1st)和小波-二阶导数法(WT-Der2nd)会较 3 种单一预处理算法效果好。比较经 5 种预处理算法处理后所建模型的效果，选出效果最好的预处理算法。

1.3.3 特征波段提取

光谱数据是连续采集的多波段吸光度数据，光谱多波段数据之间相关性非常高，存在数据冗余现象。运用特征波段提取算法能够减少数据冗余，提升模型的运算效率。连续投影变换(successive projection algorithm，SPA)与逐步回归(stepwise regression，SR)两种特征波段提取算法均能实现从所有采集的光谱波段中选择具有代表性的几个波段作为模型的输入，从而使构建的模型更加简单，并在一定程度上提升模型预测性能[12]。

1.3.4 潜在变量(latent variables，LVs)

LV 数量是 PLSR 模型的一个内部参数，它的数量会直接影响到 PLSR 模型的预测性能。用于建模的 LV 个数过多或者过少，均会降低 PLSR 的模型精度。因此确定合适数量的 LV 至关重要，通常会选择 RMSEP 值最小时所对应的 LV 个数为 PLSR 模型的最佳 LV 数量。

1.3.5 建模方法

逐步回归(stepwise regression，SR)、主成分回归(principal component regression，PCR)和偏最小二乘回归(partial least squares regression，PLSR)是建立光谱定量模型中常见的 3 种线性模型。当因变量与自变量之间存在显著的线性关

系时，这 3 种方法均较为适用[13-14]。比较分别运用 3 种方法所建模型的预测效果，选择最适用的方法。

1.3.6　模型评价指标

通常用于评价定量模型精度的指标有决定系数与均方根误差。R^2 表示建模集的决定系数，r^2 表示预测集的决定系数，决定系数越接近于 1，模型精度越高。RMSEC 表示建模集的均方根误差，RMSEP 表示预测集的均方根误差，均方根误差越小，模型精度越高。决定系数越高，均方根误差越小，建立的模型越理想，反之表示模型越不理想。

2　结果与讨论

2.1　养殖水体的 UV 光谱图

图 2 为水产养殖水体总氮样本的 UV 原始光谱曲线，从图 2 中可以看出不同样本的光谱曲线走势大致相同，没有呈现出显著性的差异。由于不同样本的总氮浓度不同，相对应的吸光度曲线会出现整体细微平移的现象。由图 2 可知，水体总氮浓度与吸光度之间的线性关系较为显著，水体总氮浓度越高，吸光度值越大。

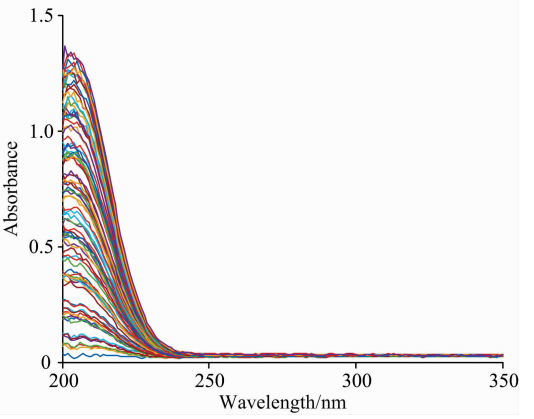

图 2　不同水样的原始吸收光谱曲线

Fig. 2　Original absorption spectra of different aquaculture water samples

2.2　样本划分与建模方法选取

由于本文的样本数据量有限，因此建模集与预测集的比例选取会大大地影响模型的预测效果，本文对两者比例的确定过程中将会十分谨慎。通常建模集与预测集的比例在 2∶1～3∶1之间比较合适[10]，因此本文选用了建模集∶预测集＝3∶1 和建模集∶预测集＝2∶1 两种方式，通过比较选择最适合的比例进行建模。第一种方式选择建模集与预测集为建模集 66 个，预测集 22 个，建模集∶预测集＝3∶1，其划分方式是从第 2 个样本起，每隔 3 个样本取样为预测集，剩余的为建模集。第二种方式选择建模集 59 个，预测集 29 个，建模集∶预测集≈2∶1，其划分方式是从第 2 个样本起，每隔 2 个样本取样为预测集，剩余的为建模集。CG 法对样本分别进行两种比例的划分，其样本集统计结果如表 1 所示。

表 1　CG 法对样本集进行两种比例的划分结果

Table 1　Two sample division results of TN concentration of aquaculture water

Data set	No.	Mean/(mg·L^{-1})	Minimum/(mg·L^{-1})	Maximum/(mg·L^{-1})	Standard deviation/(mg·L^{-1})
Calibration	66	1.031	0.03	2.08	0.601
Prediction	22	1.047	0.07	2.05	0.610
Calibration	59	1.040	0.03	2.08	0.606
Prediction	29	1.024	0.06	2.02	0.598

2.2.1　建模方法选取

为了确定小比例的样本集划分方法能否达到所需预测精度，本文用两种比例的样本划分方式分别建立 PCR，SR 和 PLSR 3 种线性模型。表 2、表 3 分别为建模集∶预测集＝3∶1 与建模集∶预测集≈2∶1 的模型预测效果，并通过对比找出最适合本文的建模方法。

表 2　PCR，SR 和 PLSR 的预测效果

（建模集＝66，预测集＝22）

Table 2　Prediction results of PCR，SR and PLSR models（calibration set number＝66，prediction set number＝22）

建模方法	建模集		预测集	
	R^2	RMSEC	r^2	RMSEP
PCR	0.983	0.077	0.974	0.100
SR	0.992	0.054	0.980	0.093
PLSR	0.993	0.049	0.980	0.088

注：RMSEC 和 RMSEP 的单位均为 mg·L^{-1}

表 3　PCR，SR 和 PLSR 的预测效果

（建模集＝59，预测集＝29）

Table 3　Prediction results of PCR，SR and PLSR models（calibration set number＝59，prediction set number＝29）

建模方法	建模集		预测集	
	R^2	RMSEC	r^2	RMSEP
PCR	0.978	0.089	0.985	0.088
SR	0.987	0.068	0.985	0.077
PLSR	0.990	0.061	0.993	0.052

注：RMSEC 和 RMSEP 的单位均为 mg·L^{-1}

由表 2 可知，当模型建模集与预测集比例为 3∶1 时，R^2 与 r^2 均超过 0.97，R^2 较高，且二者差别为 0.01 左右，RMSE 均小于等于 0.1 mg·L^{-1}，且 RMSEP 与 RMSEC 相差不超过 0.04 mg·L^{-1}，因此说明数据的质量比较好。当模型建模集与预测集比例近似为 2∶1 时，同样建立三种模型，建模结果如表 3。由表 3 可知，R^2 与 r^2 仍均超过 0.97，r^2 较

高，且二者差别小于 0.01，RMSE 仍均小于 0.1 mg·L^{-1}，RMSEP 与 RMSEC 的差距缩小为 0.01 mg·L^{-1}。比较可知，运用第二种比例的样本划分方法所建模型 RMSEP 较第一种整体下降，且 R^2 比第一种整体上升。因此，当模型建模集与预测集比例近似为 2∶1 时，模型的预测效果更好。除此之外，由表 2 和表 3 对比可知，PLSR 的建模效果是 3 种方式中最好的，因此，本文选择 PLSR 模型进行建模。

通常 PLSR 模型的 LV 数量会在很大程度上影响模型精度。在模型较为理想时，RMSEP 一般会随 LV 个数的增加而降低并达到最小值，然后随着 LV 个数的增加反而再升高。在极少数情况下，LV 个数为 1 时就可以使模型最理想。图 3 是在对样本集进行 SPXY 划分后，没有经过数据处理的情况下，所建 PLSR 模型 LV 个数为 1～10 时 RMSEP 的变化曲线。当 LV 个数等于 2 时，RMSEP 最小，模型最理想。但是不同 PLSR 模型理想的 LV 个数不一定相同，因此需要针对每一个 PLSR 模型分别确定其最优 LV 个数。

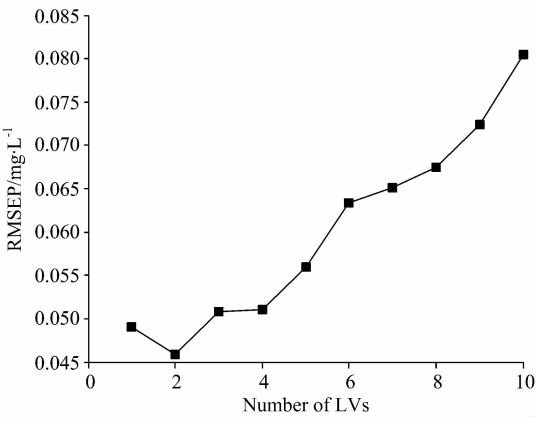

图 3 LV 数量对模型 RMSEP 的影响

Fig. 3 Relationship of latent values and RMSEP

2.2.2 样本划分方法比较

选择合适的样本划分方法有助于提升模型的预测性能，相反如果方法不当，极有可能无法获得理想的预测效果。本文使用 CG 法、KS 算法和 SPXY 算法三种样本划分方法进行样本集划分。三种方法如前文介绍，均有自己的优势。本文欲根据所采集到数据样本的特征，选出适合本文的样本集划分方法。样本总量共 88 个，取建模集∶预测集≈2∶1，即建模集为 59 个，预测集为 29 个。运用 CG 法分两种方法取样，第一种方法是从第 2 个样本起，每隔 2 个样本取样作为预测集。第二种方法是从第 3 个样本起，每隔 2 个样本取样为预测集。4 种样本集划分统计结果如表 4 所示。

表 4 四种样本集划分统计结果

Table 4 Statistics description of sample selection for the aquaculture sample based on four different algorithms

Sample selection	Data set	Sample No.	Mean/ (mg·L^{-1})	Minimum/ (mg·L^{-1})	Maximum/ (mg·L^{-1})	Standard deviation/ (mg·L^{-1})
CG 1	Calibration	59	1.040	0.03	2.08	0.606
	Prediction	29	1.024	0.06	2.02	0.598
CG 2	Calibration	59	1.029	0.03	2.08	0.605
	Prediction	29	1.047	0.07	2.05	0.601
RS1	Calibration	59	0.922	0.03	2.08	0.556
	Prediction	29	1.263	0.21	2.01	0.630
RS2	Calibration	59	0.988	0.03	2.08	0.600
	Prediction	29	1.130	0.06	2.05	0.601
KS	Calibration	59	1.174	0.03	2.08	0.589
	Prediction	29	0.752	0.07	2.00	0.527
SPXY	Calibration	59	1.094	0.03	2.08	0.593
	Prediction	29	0.915	0.07	2.02	0.600

表 5 基于四种样本划分方法建立的 PLSR 模型的预测效果

Table 5 Prediction results of PLSR models based on four different subsets partition algorithms

建模方法	LVs	建模集		预测集	
		R^2	RMSEC	r^2	RMSEP
CG 1	4	0.990	0.061	0.993	0.052
CG 2	2	0.986	0.058	0.990	0.072
SR1	5	0.990	0.056	0.991	0.060
SR2	2	0.985	0.072	0.991	0.057
KS	6	0.993	0.049	0.984	0.077
SPXY	2	0.983	0.077	0.996	0.046

注：RMSEC 和 RMSEP 的单位均为 mg·L^{-1}

由表 5 所建立的 6 种 PLSR 模型的预测效果可知，CG 法 1 的预测效果比 CG 法 2 好，两种 SR 法所建模型的预测效果差别较小，CG 法和 SR 法所建模型的建模效果与预测效果的差别较其他两种方法相对较小，但这两种方法的随机性较强，需要经过多次试验才能得到较好的建模效果。KS 法与 SPXY 法的划分结果是确定的，相对来说建模效果比较稳定。KS 算法的建模集 R^2 最高，RMSEC 最小，而预测集的预测效果比建模集差。SPXY 算法情况恰好相反，建模集 r^2 最高，RMSEP 最小，预测集的预测效果较建模集更好。本文以预测集预测精度的高低作为最终评价模型优劣的指标，因此选择 SPXY 为本文的样本集划分方法。

2.3 最优数据处理方法

本文运用的预处理算法包括小波变换(WT)、一阶导数

(Der1st)、二阶导数(Der2nd)、小波-一阶导数(WT-Der1st)以及小波-二阶导数(WT-Der2nd)5 种预处理算法。运用了 SPA 和 SR 两种特征波段提取算法，通过比较运用不同预处理及特征波段提取算法的 PLSR 模型的预测效果，以确定最优预处理及特征波段提取算法。表 6 综合比较了不同数据处理算法所对应的模型预测效果，分析可选出最优数据处理方法。

表 6　基于多种预处理和特征提取方法建立的 PLSR 模型效果对比

Table 6　PLSR models' effect of TN concentration based on different pretreatment and feature extraction methods

模型序号	预处理	特征提取	变量数量	LVs	建模集		预测集	
					R^2	RMSEC	r^2	RMSEP
1	—	—	161	2	0.983	0.077	0.996	0.046
2	WT	—	161	2	0.982	0.080	0.996	0.043
3	Der1st	—	161	1	0.989	0.062	0.991	0.058
4	Der2nd	—	161	8	0.940	0.144	0.924	0.166
5	WT-Der1st	—	161	2	0.982	0.078	0.996	0.043
6	WT-Der2nd	—	161	1	0.987	0.067	0.994	0.055
7	WT-Der1st	SPA	1	1	0.978	0.087	0.996	0.042
8	WT-Der2nd	SPA	5	5	0.986	0.071	0.990	0.063
9	WT-Der1st	SR	4	2	0.986	0.069	0.995	0.045
10	WT-Der2nd	SR	8	4	0.982	0.078	0.992	0.060

注：RMSEC 和 RMSEP 的单位均为 mg·L^{-1}

由表 6 可以得出如下 7 个结论：(1)除了模型 4 使用 Der2nd 预处理的模型预测效果较差之外，其他 9 个模型的预测效果均较好；(2)模型 1，2，5 和 7 的决定系数均为最高，但模型 7 的 RMSEP 最低，由此可知，预处理算法、LV 变量个数和特征波段提取算法的选择，对于模型精确度的提升均有影响；(3)由模型 1，2，3 和 4 对比可知，WT 在一定程度上可以改善模型的预测效果，主要表现在 RMSEP 的降低，而单纯的导数预处理则没有使模型的预测效果得到改善；(4)由模型 2，5 和 6 对比可知，WT-Der1st 的预处理后的模型预测效果比单一的 WT 好，WT-Der2nd 预处理后的模型预测效果则不如 WT 好；(5)由模型 5 和 7 对比可知，经过同样的预处理方法，然后经 SPA 特征提取后，模型的预测效果比只经过 WT-Der1st 预处理后的建模预测效果略好；(6)由模型 6 和 8 对比可知，基于 WT-Der2nd 预处理，经 SPA 特征提取后，模型的预测效果比只经过预处理后的建模预测效果差；(7)由模型 7 和 9 对比，模型 8 和 10 对比可知，运用 SPA 进行特征提取的波段数量小于 SR，其特征提取效率比 SR 高，基于 WT-Der1st 预处理后，再经过 SPA 特征波段提取后建模的效果比 SR 好，基于 WT-Der2nd 预处理后，再经过 SPA 特征波段提取后建模的效果却不如 SR 好。

综上表明，本文以模型预测精度尽可能高且模型复杂度尽可能低为模型选择原则，SPA 特征提取算法能够大大地简化模型。因此选择模型 7 作为最优模型，即 WT-Der1st 预处理方法作为最优预处理方法，SPA 作为最优特征提取方法。

2.4　基于小波-一阶导数-连续投影变换的 PLSR 模型

通过对比不同预处理方法、不同特征提取方法所建立的模型预测效果，最终通过比较 10 个模型的 r^2 与 RMSEP，发现基于小波－一阶导数－连续投影变换的 PLSR 模型的预测效果与运行效率是最好的。首先通过小波变换对原始光谱曲线进行噪声去除，使得曲线变得更加平滑，其次再通过求一阶导数，可以消减光谱数据的基线漂移，提高数据质量等。然后通过 SPA 进行特征提取，提取到的特征波段是 218 nm，该特征波段与查阅文献所获知的总氮的吸收峰位置相一致。

基于上述数据处理方法，建立了总氮 PLSR 模型，所建模型为 $y=-31.75x-0.035$，该模型的 r^2 为 0.996，模型的 RMSEP 为 0.042 mg·L^{-1}，图 4 为总氮预测值和实测值的比较结果，发现所有数据点均在 $y=x$ 直线左右分布，非常靠近直线。由此可见，本文所使用的一系列数据处理方法相对科学，所得模型的预测效果十分理想，能够满足要求。

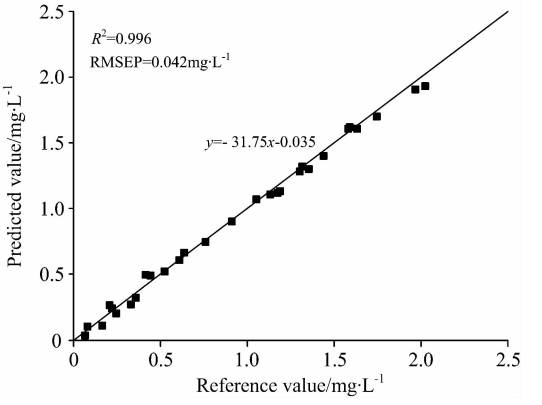

图 4　基于 WT-Der1st-SPA 的 PLSR 模型预测值与实测值对比图

Fig. 4　Relation between reference value and predicted value of TN by WT-Der1st-SPA combined with PLSR method

3　结　论

使用紫外光谱技术进行总氮含量快速检测方法与传统的化学试剂检测法相比是一种快速、有效、简便、环保的检测方法。本文大量地应用算法的对比方法，比较运用不同数据

处理方法建立的 10 个 PLSR 模型，得出结论可知，SPXY 法是最合适的样本集划分算法，在此基础上建立的基于 WT-Der1st-SPA 的总氮 PLSR 模型的预测效果最好，其预测结果 $r^2=0.996$，RMSEP=0.042 mg·L^{-1}。相比原始光谱数据的全波段建模效果，本文所建模型仅使用 1 个波段，而 RMSEP 却降低了 8.7%。由此可见，光谱预处理算法能够降低原始光谱数据的噪声，提高数据精确度；特征波段提取算法与 LV 变量选择能够解决光谱的数据冗余问题，简化模型。由此可见使用紫外－可见光谱技术能够快速准确地预测出总氮含量，为实现光谱技术在水产养殖其他水质监测指标的在线检测以及快速测定提供了经验。

References

[1] ZHANG Yan, LI Chang-you, SHEN Hung-tao, et al(张 岩, 李畅游, SHEN Hung-tao, 等). Advances in Water Science(水科学进展), 2013, 24(5): 728.

[2] ZHEN Yi-li, ZHAO Yan-dong, DONG Wei, et al(郑一力, 赵燕东, 董 玮, 等). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2018, 49(s1): 393.

[3] ZHAO Jin-hui, YUAN Hai-chao, LIU Mu-hua, et al(赵进辉, 袁海超, 刘木华, 等). Chinese Journal of Analytical Chemistry(分析化学), 2013, 41(4): 546.

[4] LIU Si-jia, TIAN You-wen, ZHANG Fang, et al(刘思伽, 田有文, 张 芳, 等). Food Science(食品科学), 2017, 38(8): 277.

[5] Wang Lu, Pu Hongbin, Sun Dawen. Talanta, 2016, 147: 422.

[6] Chen B, Wu H, Li S F Y. Talanta, 2014, 120: 325.

[7] Guo Y, Liu X, Han Y, et al. Water, Air, and Soil Pollution, 2017, 228(8): 317.

[8] LI Xin-xing, ZHU Chen-guang, ZHOU Jing, et al(李鑫星, 朱晨光, 周 婧, 等). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2018, 34(19): 184.

[9] CHEN Yi-yun, ZHAO Rui-ying, QI Tian-ci, et al(陈奕云, 赵瑞瑛, 齐天赐, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2017, 37(7): 2133.

[10] MAO Fu-hui, SUN Hong, LIU Hao-jie, et al(毛博慧, 孙 红, 刘豪杰, 等). Transactions of the Chinese Society for Agricultural Machinery(农业机械学报), 2017, (s1): 160.

[11] Shi Haitao, Yu Peiqiang. Food Control, 2017, 82: 57.

[12] YOU Shi-bing, YAN Yan(游士兵, 严 研). Stastics & Decision(统计与决策), 2017, (14): 31.

[13] Shi T, Luan X L, Liu F. Vibrational Spectroscopy, 2017, 92: 302.

[14] Abdi H, Williams L J. Methods in Molecular Biology, 2013, 930: 549.

Rapid Determination of Total Nitrogen in Aquaculture Water Based on Ultraviolet Spectroscopy

LI Xin-xing[1], ZHOU Jing[1], TANG Hong[2], SUN Long-qing[1], CAO Xia-min[3], ZHANG Xiao-shuan[4]*

1. Beijing Laboratory of Food Quality and Safety, College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China
2. Yantai Institute of China Agricultural University, Yantai 264000, China
3. School of Biology and Basic Medical Sciences, Soochow University, Suzhou 215200, China
4. College of Engineering, China Agricultural University, Beijing 100083, China

Abstract The paper is intended to achieve rapid determination of total nitrogen (TN) concentration by using Ultraviolet (UV) spectroscopy technology, which was one of the most important indicators to measure the pollution degree in aquaculture water. The original dataset used in the paper contains 88 samples data with actual concentration value and spectral absorbance value. It is helpful to select the optimal model through the five stages that include sample set division algorithms, data preprocessing algorithms, feature band extraction algorithms, model selection algorithms and latent values (LVs) selection method. In the first four stages, the comparison results of different methods show that each stage is necessary, and only by comparing the advantages and disadvantages of modeling results with various algorithms can we find the most suitable modeling process and method. First of all, the original sample set is processed by the concentration gradient (CG) method, then three models are built which respectively are principal component regression (PCR), stepwise regression (SR) and partial least squares regression(PLSR), and it proves that the PLSR is the best prediction model. The number of LVs can greatly influence the accuracy of model, and usually

when the value of the model root mean square error of prediction (RMSEP) is the minimum, the LV number is optimal. Secondly, it is testified that the SPXY algorithm is the best by comparing the effect of random sampling (RS) algorithm, concentration gradient (CG) algorithm, kennard stone (KS) algorithm and SPXY algorithm. Thirdly, based on SPXY algorithm, the paper uses five preprocessing algorithms which are wavelet transform (WT) method, first derivative (Der1st), and second derivative (Der2nd) three single preprocessing algorithms, WT-Der1st and WT-Der2nd. Fourthly, according to the results of data processing, using successive projections algorithm (SPA) and stepwise regression (SR) for feature band extraction algorithms, the results show that the extraction efficiency of SPA not only can greatly reduce the complexity of model, but also improve the prediction accuracy. The feature band extracted based on WT-Der1st-SPA is 218 nm, which is consistent with the characteristics of total nitrogen band range, indicating the method was relatively scientific. Finally, considering the prediction accuracy and complexity of model, the PLSR based on WT-Der1st-SPA with the best results with the determination coefficient (r^2) and RMSEP being 0.996 and 0.042 mg • L^{-1} for the prediction set in 10 models. In short, the prediction model established could be applied to the rapid and accurate determination of total nitrogen concentration. Moreover, this study laid the foundation for further implementation of online analysis of aquaculture water and rapid determination of other water quality parameters.

Keywords Ultraviolet spectroscopy; Total nitrogen; Wavelet transform; Successive projections algorithm; Latent values (LVs); Partial least squares regression (PLSR)

(Received Nov. 1, 2018; accepted Mar. 7, 2019)

* Corresponding author

《光谱学与光谱分析》对来稿英文摘要的要求

来稿英文摘要不符合下列要求者，本刊要求作者重写，这可能要推迟论文发表的时间。

1. 请用符合语法的英文，要求言简意明、确切地论述文章的主要内容，**突出创新之处**。

2. 应拥有与论文同等量的主要信息，包括四个要素，即研究目的、方法、结果、结论。其中后两个要素最重要。有时一个句子即可包含前两个要素，例如“用某种改进的 ICP-AES 测量了鱼池水样的痕量铅”。但有些情况下，英文摘要可包括研究工作的主要对象和范围，以及具有情报价值的其他重要信息。在结果部分最好有定量数据，如检测限、相对标准偏差等；结论部分最好指出方法或结果的优点和意义。

3. 句型力求简单，尽量采用被动式，建议经专业英语翻译机构润色，与中文摘要相对应。用 A4 复印纸单面打印。

4. 摘要不应有引言中出现的内容，换言之，摘要中必须写进的内容应尽量避免在引言中出现。摘要也不要对论文内容作解释和评论，不得简单重复题名中已有的信息；不用非公知公用的符号和术语；不用引文，除非该论文证实或否定了他人已发表的论文。缩略语、略称、代号，除相邻专业的读者也能清楚地理解外，在首次出现时必须加以说明，例如用括号写出全称。