近红外光谱结合新型双集成建模方法快速测定土壤有机质含量

Rapid determination of soil organic matter using near-infrared spectroscopy combined with a novel double ensemble modeling method

李莹霞，黄海平，赵子贞，谭小耀，卞希慧*

（天津工业大学化学工程与技术学院，中国天津 300387）

摘要：近红外光谱技术因快速、无损、环保等优点，常与多元校正方法结合用于复杂样品的定量分析。然而，使用全光谱建模时，会有许多无关变量（例如噪音）对建模效果产生影响，从而降低建模精度，增加计算负担。集成建模可以通过融合多个子模型来提高单个模型的稳定性和预测精度。因此，本文提出一种基于蒙特卡洛（Monte Carlo, MC）-蝴蝶优化算法（butterfly optimization algorithm, BOA）的双集成偏最小二乘（partial least squares, PLS）建模方法，并将其与近红外光谱相结合，用于土壤样本中有机质含量的定量分析。为了验证该方法的有效性，将其应用于土壤有机质成分的定量分析。结果表明，BOA-PLS 在建模性能上优于 PLS，而 MC-BOA-PLS 进一步提高了 BOA-PLS 的性能。因此，MC-BOA-PLS 方法可以显著提高 PLS 模型的预测性能。

关键词：蒙特卡洛；蝴蝶优化算法；偏最小二乘；集成建模；多元校正

1 引言

土壤有机质是土壤养分的主要来源，也是衡量土壤肥力的重要指标[1-3]。它在作物生长发育和土壤质量改善中起着至关重要的作用[4]。因此，快速准确地估算土壤有机质含量对于改良土壤和提高作物产量至关重要。传统的 SOM 含量测量方法耗时、费力、效率低；无法及时、无损地进行大规模的土壤有机质测定[5-7]。由于土壤的空间变异性，少量的土壤采样无法准确反映研究区域土壤养分的空间分布[8-10]。因此，需要建立一种准确、经济、高效、无损的分析技术估算 SOM 含量。

在现有的分析方法中，近红外光谱被广泛用于测量土壤有机质含量。2016 年，Song 等人[11]在测量 SOM 时，使用 PLS 方法得出 RMSEP 为 1.76；2017 年，Lin 等人[12]使用 PLS 方法测得 RMSEP 为 2.2516±0.3543；2023 年，Yang 等人[13]使用同样方法得出 RMSEP 为 2.049。因此提高预测准确度很有必要。

近红外光谱由宽、弱、重叠的波段组成，这增加了模型的复杂性[14,15]。为了克服上述问题，引入了多元校准方法包括一系列统计模型和技术，如偏最小二乘（PLS）[16-19]、人工神经网络（ANN）[20,21]、支持向量回归(SVR)[22,23]、极值学习机（ELM）[24,25]等。PLS 在处理具有线性行为的数据方面的简单性和有效性使其成为多元校准中最受推崇的方法[26]。然而，单模型校准方法的预测性能有时并不令人满意。为此，可以采用集成策略。有几种方法可以生成训练子集，例如从样本方向、变量方向、样本和变量方向以及原始信号的分解生成训练子集[27]。基于样本方向的集成建模可以通过重新选择训练集来增加集成建模的多样性。然而，当样本数量太小时，预测性能很差。基于变量方向的集成建模可以解决小样本问题，删除无关变量，降低模型的复杂性。与单集成建模相比，双集成建模可以进一步提高模型的稳定性和预测精度[26]。结合样本集成和变方向集成的优点，双集成策略在光谱分析中受到越来越多的关注。

为了定量分析土壤中有机质含量，基于蒙特卡洛（MC）采样和 BOA 的优点，提出了一种新的双集成 PLS 方法，称为 MC-BOA-PLS。在该方法中，通过重复的 MC 采样和从原始训练集中进一步选择 BOA 变量来生成多个训练子集。然后，在这些训练子集上构建 PLS 子模型。最终预测是通过简单地对 PLS 子模型的预测进行平均来获得的。通过比较 PLS、BOA-PLS 和 MC-BOA-PLS，验证了所提出方法的有效性。

2 原理与算法

2.1 蒙特卡洛取样

蒙特卡罗取样是一种基于随机数和概率统计的强有力技术，用于研究多变量问题[28,29]。在化学计量学等多个领域中，它都受到了广泛的关注和应用。在本工作中，每次建立 PLS 子模型时，都采用了 MC 采样技术。具体来说，就是随机选取一定数量的样本作为样本子集。这种随机选取的方式有助于捕捉数据的多样性和不确定性，从而更全面地反映数据的特征。通过进行多次 MC 采样，可以显著提高模型的精确度和稳定性，这样也在一定程度上避免了对单一模型的依赖。

2.2 MC-BOA-PLS 算法和原理

在近红外建模中，通过将 MC 采样和 BOA 变量选择方法相结合，并从样本方向和变量方向分别进行整合，结合 MC 采样和 BOA 的优点，去除冗余变量，构建更稳定、准确、可靠的近红外模型。MC-BOA-PLS 的流程图如图 1 所示，具体步骤如下：

（1）将所有样本的 2/3 分为训练集，1/3 分为预测集，分别用于建模和外部验证。

（2）在训练集中使用 MC 采样，随机选择一定数量的样本形成样本子集。

（3）对样本子集进行 BOA 变量选择，以删除不相关的变量并获得训练子集

（4）使用上述训练子集建立 PLS 子模型。重复步骤（2）至（4）K 次，以获得 K 个不同的子模型。

（5）子模型用于预测预测集中的样本，每个子模型提供预测结果。对所有子模型的预测结果进行简单的平均，得到最终的预测结果。

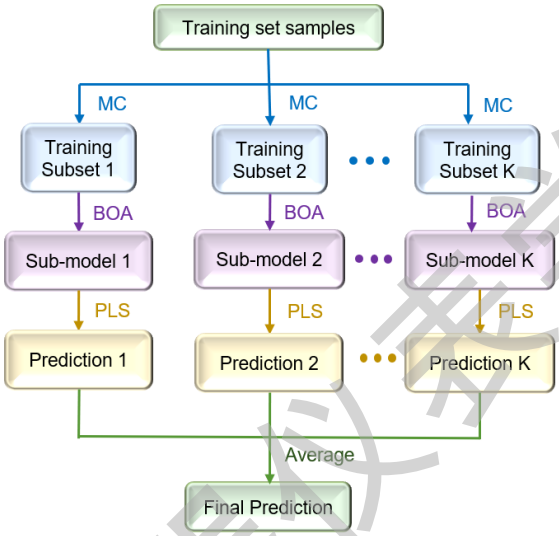

图 1　MC-BOA-PLS 流程图

3 实验

数据集为土壤数据集，是由 Rinnan 等人[30]提供，其中包括 108 个土壤样本，以及有机质和麦角甾醇含量。此数据集可以在网址 http://www.eigenvector.com.上下载。光谱由 NIRSystems 6500 近红外分光光度计（Foss Analytical，Hoganas，Sweden）测量。每个光谱在 400-2500nm 的波长范围内记录了 1050 个变量，间隔为 2nm。以近红外光谱和土壤有机质为研究对象。图 2（a）和（b）分别显示了近红外光谱和土壤浓度。

在计算之前，土壤数据集被分为 72 个和 36 个样本进行训练和预测。训练集和预测集分别用于模型构建和外部验证。在 PLS、BOA-PLS 和 MC-BOA-PLS 模型的比较中，采用了相同的训练集、预测集方法。

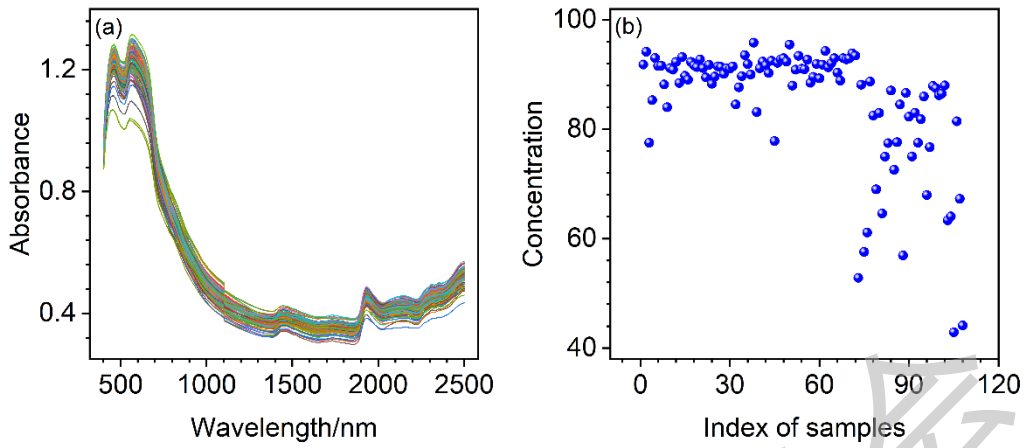

图 2　土壤数据集的 NIR 图（a）和土壤中有机质组分的浓度分布图（b）

4 结果讨论

4.1 蝴蝶优化算法中迭代次数的优化

迭代次数是 MC-BOA-PLS 建模中的一个重要参数。当迭代次数较小时，模型的预测精度较差，当迭代次数较大时，计算时间和模型复杂度会增加。因此，选择最佳迭代次数对确保模型的准确性具有深远的影响。在这项研究中，迭代次数的范围是 1-500。在每次迭代中，建立 MC-BOA-PLS 模型并获得 RMSECV。

图 3 显示了有机质组分的 RMSECV 随迭代次数的变化。从图 3 中可以看出，当迭代次数很小时，RMSECV 相对较大。随着迭代次数的增加，RMSECV 首先显著下降，然后略有波动，最后趋于平稳。当迭代次数为 500 时，RMSECV 几乎不变。因此，500 被认为是土壤中有机物成分的最佳迭代次数。

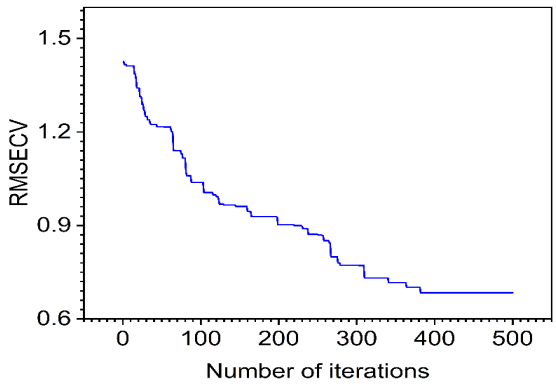

图 3　土壤中有机质组分的 RMSECV 随着迭代次数的变化图

4.2 蝴蝶种群数的确定

蝴蝶数量（N）的大小会影响建模的准确性，因此，N 的确定也至关重要。N 不足可能导致搜索空间覆盖不足，容易陷入局部最优。如果 N 太大，虽然可以提高全局搜索能力，但会延长找到最优解的时间，降低收敛速度和计算速度。为了观察蝴蝶数量与其性能之间的关系，并获得最佳的蝴蝶数量，以 5 的间隔将 N 从 5 增加到 100。该模型预测的 RMSECV 被用作最佳种群规模的评价标准，并得到土壤有机质 RMSECV 随蝴蝶种群数和时间的变化。

图 4 显示了土壤有机质 RMSECV 随蝴蝶种群数和时间的变化。可以看出，随着 N 的增加，整体趋势呈现波动和下降趋势。当 N 为 70 时，RMSECV 达到其最低点。因此，将 N 设置为 70 可以更好地找到最优解。此外，从时间的角度来看，时间随着种群数的增加而不断增加。当 N 达到 100 时，运行时间保持在 30 秒内，表明蝴蝶算法的高效性。

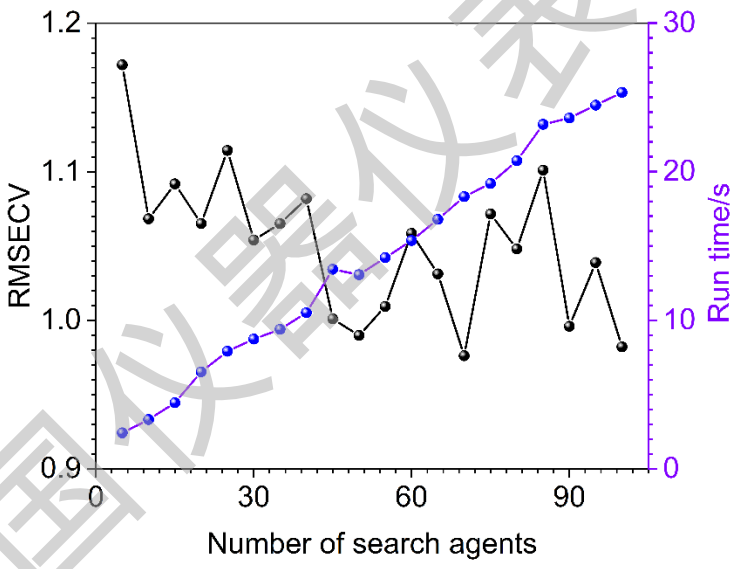

图 4　土壤中有机质的 RMSECV 随蝴蝶种群数以及时间的变化图

4.3 子模型的迭代次数

在 MC-BOA-PLS 模型中，BOA 用于识别和保留对模型性能有重大影响的变量。因此，有必要找到子模型迭代的最佳次数，以提高集成模型的准确性。因此，有必要找到子模型迭代的最佳次数，以提高集成模型的准确性。在这项工作中，对子模型进行了 1 到 200 次迭代的研究，并计算了每次迭代的 RMSECV。

土壤有机质组分 RMSECV 随子模型迭代次数的变化如图 5 所示。一开始，RMSECV 相当大。在子模型迭代次数达到 25 之前，随着子模型迭代数量的增加，RMSECV 急剧下降，

然后曲线在一定范围内波动并稳定。这表明在 MC 进行多次采样后，模型的预测性能有所提高。

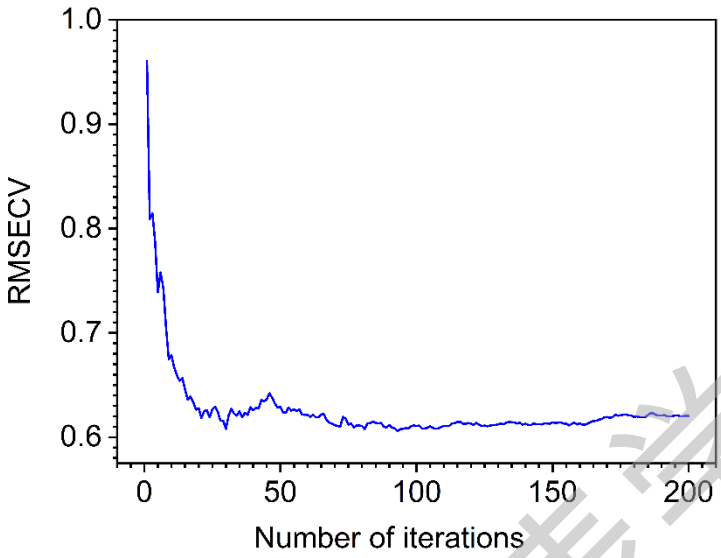

图 5　土壤中的有机质的 RMSECV 随着子模型迭代次数的变化图

4.4 预测结果比较

在确定所有最佳参数的基础上，建立了 MC-BOA-PLS 模型，并将其用于预测土壤中的有机质含量。为了验证该方法的稳定性和预测能力，对 PLS、BOA-PLS 和 MC-BOA-PLS 进行了比较。在这些方法中，PLS 是一种简单且广泛使用的单一模型，BOA-PLS 在为 BOA 选择变量后使用 PLS 构建模型。图 6 显示了土壤数据集的 PLS（a）、BOA-PLS（b）和 MC-BOA-PLS（c）预测结果，并计算了数据集的 R 和 RMSEP。

根据图 6，PLS、BOA-PLS 和 MC-BOA-PLS 的 R 值均在 0.9 以上。然而，在 BOA 变量选择后建立的 PLS 模型的 R 值大于全谱 PLS 模型，达到 0.99 以上。与 PLS 相比，BOA-PLS 的 RMSEP 值也显著降低。这表明 BOA-PLS 方法可以提高模型的预测性能。其次，与 BOA-PLS 和 PLS 相比，MC-BOA-PLS 的预测值与真实值之间的拟合度最高，R 值大于 BOA-PLS 或 PLS。MC-BOA-PLS 的 RMSEP 值也低于 BOA-PLS 和 PLS 模型。结果表明，MC-BOA-PLS 具有最佳的预测性能，从而证明了该方法的优越性。

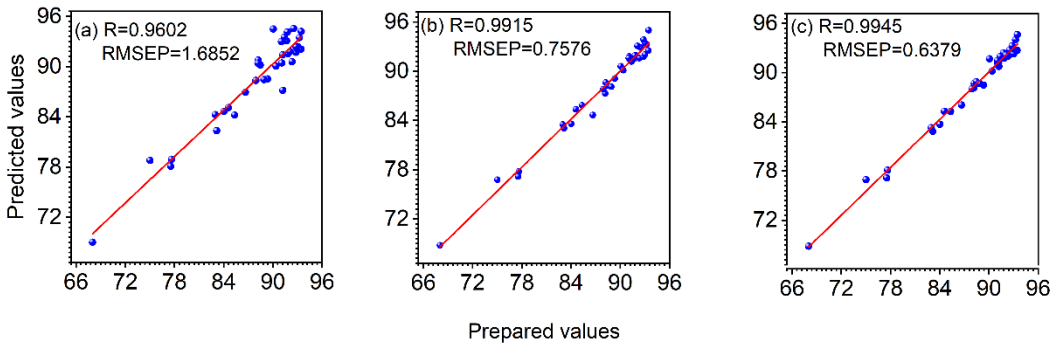

图 6　土壤数据集 PLS（a）、BOA-PLS（b）和 MC-BOA-PLS（c）的预测结果

5 结论

本文提出了一种新的双集成建模方法 MC-BOA-PLS，该方法使用近红外光谱对土壤样本中的有机质含量进行定量分析。确定了 BOA 迭代次数、蝴蝶数量和 PLS 子模型迭代次数。采用最优参数，开发了 MC-BOA-PLS 模型，并成功预测了土壤中的有机质含量。此外，比较了 MC-BOA-PLS 和 PLS 的预测性能。结果表明，MC-BOA-PLS 方法表现出最高的 R 和最小的 PMSEP，从而产生了最佳的预测结果。因此，所提出的近红外光谱方法是测定土壤样本中有机质含量的有效和准确的工具。

参考文献：

[1] Y. L. Ba, J. B. Liu, J. C. Han, X. C. Zhang. Application of Vis-NIR spectroscopy for determination the content of organic matter in saline-alkali soils. Spectrochimica Acta Part A-Molecular and Biomolecular Spectroscopy. 229 (2019) 117863.

[2] S. G. Xie, F. J. Ding, S. G. Chen, X. Wang, Y. H. Li, K. Ma. Prediction of soil organic matter content based on characteristic band selection method. Spectrochimica Acta Part A-Molecular and Biomolecular Spectroscopy 273 (2022) 120949.

[3] J. B. Liu, Z. Y. Dong, J. S. Xia, H. Y. Wang, T. T. Meng, R. Q. Zhang, J. C. Han, N. Wang, J. C. Xie. Estimation of soil organic matter content based on CARS algorithm coupled with random forest. Spectrochimica Acta Part A-Molecular and Biomolecular Spectroscopy, 258 (2021) 119823.

[4] S. G. Xie, Y. H. Li, X. Wang, Z. X. Liu, K. L. Ma, L. W. Ding. Research on estimation models

of the spectral characteristics of soil organic matter based on the soil particle size. Spectrochimica Acta Part A-Molecular and Biomolecular Spectroscopy. 260 (2021) 119963.

[5] J. Chan, A. F. Plante, C. Peltre, T. Baumgartl, P. Erskine. Quantitative differentiation of coal, char and soil organic matter in an Australian coal minesoil. Thermochimica Acta 650 (2017) 44-55.

[6] G. A. Ehlers, F. Clark, T. Sean, K. E. Scherr, A. P. Loibner, L. J. Janik. Influence of the nature of soil organic matter on the sorption behaviour of pentadecane as determined by PLS analysis of mid-infrared DRIFT and solid-state ^{13}C NMR spectra. Environmental Pollution 158 (2010) 285-291.

[7] N. Dupuy, F. Douay. Infrared and chemometrics study of the interaction between heavy metals and organic matter in soils. Spectrochimica Acta Part A-Molecular and Biomolecular Spectroscopy 57 (2001) 1037-1047.

[8] H. Z. Chen, L. L. Xu, J. Gu, F. X. Meng, H. L. Qiao. A quasi-qualitative strategy for FT-NIR discriminant prediction: Case study on rapid detection of soil organic matter. Chemometrics and Intelligent Laboratory Systems 224 (2022) 104547.

[9] A. Nasonova, G. J. Levy, O. Rinot, G. Eshel, M. Borisover. Organic matter in aqueous soil extracts: Prediction of compositional attributes from bulk soil mid-IR spectra using partial least square regressions. Geoderma 411 (2022) 115678.

[10] M. St Luce, N. Ziadi, B. J. Zebarth, C. A. Grant, G. F. Tremblay, E. G. Gregorich. Rapid determination of soil organic matter quality indicators using visible near infrared reflectance spectroscopy. Geoderma 232 (2014) 449-458.

[11] X. Z. Song, Y. Huang, H. Yan, Y. M. Xiong, S. G. Min. A novel algorithm for spectral interval combination optimization. Analytica Chimica Acta 948 (2016) 19-29.

[12] Y. W. Lin, N. Xiao, L. L. Wang, C. Q. Li, Q. S. Xu. Ordered homogeneity pursuit lasso for group variable selection with applications to spectroscopic data. Chemometrics and Intelligent Laboratory Systems 168 (2017) 62-71.

[13] W. Y. Yang, Y. R. Xiong, H. H. Wang, T. Wu, Y. P. Du. Interval interaction moving window partial least squares for wavelength interval selection in near infrared spectroscopy. Chemometrics and Intelligent Laboratory Systems 241 (2023) 104976.

[14] S. F. Xie, B. R. Xiang, L. Y. Yu and H. S. Deng. Tailoring noise frequency spectrum to

improve NIR determinations. Talanta 80 (2009) 895-902.

[15] X. F. Sun, H. L. Li, Y. Yi, H. M. Hua, Y. Guan and C. Chen. Rapid detection and quantification of adulteration in Chinese hawthorn fruits powder by near-infrared spectroscopy combined with chemometrics. Spectrochimica Acta Part A-Molecular and Biomolecular Spectroscopy 250 (2021) 119346.

[16] M. Sjostrom, S. Wold, W. Lindberg, J. A. Persson and H. Martens, A multivariate calibration problem in analytical chemistry solved by partial least-squares models in latent variables. Analytica Chimica Acta 150 (1983) 61-70.

[17] J. P. Cruz-Tirado, J. M. Amigo and D. F. Barbin. Determination of protein content in single black fly soldier (Hermetia illucens L.) larvae by near infrared hyperspectral imaging (NIR-HSI) and chemometrics. Food Control 143 (2022) 109266.

[18] X. G. Shao, X. H. Bian, J. J. Liu, M. Zhang and W. S. Cai. Multivariate calibration methods in near-infrared spectroscopic analysis, Analttical Methods 2 (2010) 1662-1666.

[19] H. Zhang, X. Y. Hu, L. M. Liu, J. F. Wei and X. H. Bian. Near infrared spectroscopy combined with chemometrics for quantitative analysis of corn oil in edible blend oil. Spectrochimica Acta Part A-Molecular and Biomolecular Spectroscopy 270 (2022) 120841.

[20] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity, The Bulletin of Mathematical Biology 5 (1943) 115-133.

[21] A. Badura, J. Krysinski, A. Nowaczyk and A. Bucinski. Application of artificial neural networks to the prediction of antifungal activity of imidazole derivatives against Candida albicans. Chemometrics and Intelligent Laboratory Systems 222 (2022) 104501.

[22] R. G. Brereton and G. R. Lloyd. Support vector machines for classification and regression. Analyst 135 (2010) 230-267.

[23] W. Q. Yang, F. S. Li, Y. C. Zhao, X. Lu, S. Y. Yang and P. F. Zhu. Quantitative analysis of heavy metals in soil by X-ray fluorescence with PCA-ANOVA and support vector regression. Analytical Methods 14 (2022) 3944-3952.

[24] G. B. Huang, Q. Y. Zhu and C. K. Siew. Extreme learning machine: theory and applications. Neurocomputing 70 (2006) 489-501.

[25] C. Tan, H. Chen and Z. Lin. Brand classification of detergent powder using near-infrared spectroscopy and extreme learning machines. Microchemical Journal 160 (2021) 105691.

[26] K.Y. Wang, X.H. Bian, X.Y. Tan, H.T. Wang, Y.K. Li. A new ensemble modeling method for multivariate calibration of near infrared spectroscopy. Analytical Methods 13 (2021) 1374-1380.

[27] X. H. Bian, P. Y. Diwu, Y. R. Liu, P. Liu, Q. Li, X. Y. Tan. Ensemble calibration for the spectral quantitative analysis of complex samples. Journal of Chemometrics 32 (2018) 2940.

[28] H. Li, P. C. Wu, J. S. Dai, X. B. Zou. A Monte Carlo resampling based multiple feature-spaces ensemble (MFE) strategy for consistency-enhanced spectral variable selection. Analytica Chimica Acta 1279 (2023) 341782.

[29] L. X. Zhang, Z. Yuan, P. W. Li, X. F. Wang, J. Mao, Q. Zhang, C. D. Hu. Targeted multivariate adulteration detection based on fatty acid profiles and Monte Carlo one-class partial least squares. Chemometrics and Intelligent Laboratory Systems 169 (2017) 94-99.

[30] R. Rinnan, A. Rinnan. Application of near infrared reflectance (NIR) and fluorescence spectroscopy to analysis of microbiological and chemical properties of arctic soil. Soil Biology & Biochemistry 39 (2007) 1664-1673.

通讯作者简介：

卞希慧，女，1983 年生，天津工业大学化学工程与技术学院教授，主要进行化学计量学算法研究及其在中药、食品、环境等方面的应用研究。

E-mail：bianxihui@163.com

第一作者简介

李莹霞，女，1998 年生，硕士研究生，研究方向为基于变量选择的化学计量学算法研究。

E-mail：17303453629@163.com