

基于近红外光谱技术预测玉米伏马菌素

Forecasting method of Fumonisin in corn using near infrared spectra technique

张林 党楠 杨琳 李亚文 袁训锋

ZHANG Lin DANG Nan YANG Lin LI Ya-wen YUAN Xun-feng

(商洛学院电子信息与电气工程学院, 陕西 商洛 726000)

(School of Electronic Information and Electrical Engineering, Shang Luo University, Shangluo, Shaanxi 726000, China)

摘要:为了更好地利用近红外光谱分析技术对玉米伏马菌素含量进行预测,减小因玉米产地间的差异对玉米近红外光谱预测模型的影响,以不同产地的玉米作为研究对象,利用 $x-y$ 共生距的方法将试验样本划分为校正集与验证集,采用经典的偏最小二乘法分别建立不同产地和混合产地的玉米伏马菌素预测模型,并采用验证集样本分别对模型的预测精度进行验证。为了减小建模及预测过程的运算量,采用连续投影算法(SPA)和竞争性自适应加权算法(CARS)对不同产地玉米的近红外光谱的特征波长进行筛选,筛选出 22 个特征波长变量作为输入,大大降低了建模及预测过程的运算量,同时预测准确度也有所改善,其预测相关系数达到 0.954,为快速、无损地实现对玉米伏马菌素的检测提供了可靠的理论依据。

关键词:玉米;伏马菌素;近红外光谱;预测

Abstract: In order to forecast the content of fumonisin in corn using the infrared spectrum analysis technology, and reduce the differences caused by their yield region, the influence of experiment using 4 different origin of domestic corn were investigated. The method of using $x-y$ co-occurrence distance could be divided into calibration sample and validation sets, using the classical and different regions based on the partial least squares, and then the prediction model of fumonisin maize hybrid origin, and USES the validation set samples to validate the prediction precision, respectively. In order to reduce the computational complexity of modeling and forecasting process, experiments using continuous projection algorithm (SPA) and competitive adaptive weighting algorithm (CARS) the characteristics of the infrared spectra of different origin corn wavelength filter, and 22 characteristics were filtered out. Then these 22 wavelengths were input as variables, and this greatly reduced the computational complexity of

modeling and forecasting process, as well as improved the prediction accuracy, with the correlation coefficient at 0.954.

Keywords: Corn; Fumonisin; NIR; Predict

伏马菌素(Fumonisin, FBs)是被国际卫生组织列为几种严重致癌真菌毒素之一^[1],是在合适的温度与湿度条件下由串珠镰刀菌代谢产生的二级代谢产物,其主要存在于玉米等谷物中。20 世纪 80 年代 Gelder blom 等^[2]在玉米制品中发现并成功分离出了 B1 型伏马菌素,紧接着 Laurent 等^[3]对伏马菌素进行了深入的研究,发现伏马菌素其实包含有两种结构极其相近的有毒有害物质,分别将其命名为伏马菌素 B1 和伏马菌素 FB2,随着研究的深入,如今研究者^[3]已经发现多种类型的伏马菌素,其主要可以分为 A、B、C 和 P 4 种类型。在玉米中最常见的是 FB1 型伏马菌素,在目前所发现的所有种类的伏马菌素当中,B1 型伏马菌素对人体的危害最大,毒性最强。

早在 1993 年国际癌症研究院就把伏马菌素列为强致癌物质目录,将伏马菌素认定为可以危害动物和人体健康的物质之一^[4]。目前,国际卫生组织认为伏马菌素是几种具有重大研究意义的真菌毒素之一^[4]。传统的伏马菌素测定方法主要是通过分析化学的方法进行测定与研究^[5],包括色谱测定法和湿化学分析两种方法,这两种方法能够准确测定伏马菌素的含量,但测量过程繁琐,无法实现实时、无损的测定^[6]。

本研究拟在分析玉米近红外光谱特征的基础上,采用近红外光谱分析技术间接对玉米中伏马菌素含量进行定量预测研究,同时采用偏最小二乘法(Partial least squares regression, PLS)建立玉米伏马菌素含量预测模型,并对模型的精度与稳定性进行验证。

1 试验方法

1.1 试验材料

试验中的玉米样本选自陕西、山西、河南和黑龙江,从中

基金项目:商洛学院科学研究项目(编号:16SKY-FWDF005);商洛市科学技术研究发展计划项目(编号:SK2016-52)

作者简介:张林(1986—),男,商洛学院讲师,硕士。

E-mail:759013520@qq.com

收稿日期:2016—10—10

挑选 360 个玉米样本进行试验,其中 80 个样本产自陕西,90 个产自山西,100 个产自河南,90 个产自黑龙江。将玉米样本破碎、筛选并依次编号,标记光谱采集区域。试验进行前先将玉米样本取出在 20 ℃ 的室温下放置 12 h,使玉米样本与室内温度保持一致,尽量减少试验中温度对光谱采集的影响;

近红外光谱仪:Antaris II 型,美国 Nicolet 公司。

1.2 伏马菌素测定与光谱采集

试验过程中采用 Antaris 傅里叶变换近红外光谱仪对玉米的近红外光谱进行采集,选择对波长 800~2 400 nm 范围的光谱进行采集,扫描次数 32 次,分辨率 3.0 cm⁻¹。玉米样本的光谱采集完成后采用数字阿贝折光仪对玉米中的伏马菌素含量进行测定。

1.3 样本的划分

试验中采用 $x-y$ 共生距的划分方法,同时兼顾样本的 x 变量(近红外光谱数据)和 y 变量(玉米伏马菌素含量)的欧氏距离。为了保证试验样本在 $x-y$ 的空间的权重相同,将 x,y 的距离进行标准化:

$$d_{x-y}(p, q) = \frac{d_x(p, q)}{\max_{p, q \in [1, n]} d_x(p, q)} + \frac{d_y(p, q)}{\max_{p, q \in [1, n]} d_y(p, q)},$$

$$p, q \in [1, n], \quad (1)$$

式中:

n —— 试验样本总数;

$d_{x/y}(p, q)$ —— 试验中任意的两个样本之间的欧氏距离^[7];

在试验过程中分别对 4 个产地的玉米样本进行分组,分别将其分为校正集和预测集。

1.4 特征波长的选取

由于近红外光谱中不但包含有需要的效信息,且包含着大量的冗余信息^[8],在试验的过程中采用连续投影算法(Successive projections algorithm, SPA)和竞争自适应加权算法(Competitive Adaptive Reweighted Sampling, CARS)的方法选取玉米伏马菌素有效近红外光谱的特征波长。

1.5 玉米伏马菌素近红外光谱检测模型的建立

偏最小二乘法(PLS)因其具有良好的抗干扰性能和稳定性,已在近红外光谱分析领域得到了广泛的应用^[9],也是经典的近红外光谱建模方法之一。试验过程中首先采用偏最小二乘法分别建立单一产地的玉米伏马菌素近红外光谱检测模型,然后再建立 2 个产地混合、3 个产地混合和 4 个产地混合的玉米伏马菌素近红外光谱预测模型。最后,采用验证集玉米样本分别对模型的预测精度和稳定性进行验证。

2 结果与讨论

2.1 玉米样品的近红外光谱分析以及样本的划分

为了减少光谱噪声的影响及运算量,试验过程中选取 4 000~12 000 cm⁻¹ 共计 3 200 个波长点进行研究。图 1 为玉米样本的原始漫反射近红外光谱,仅从光谱图很难区分不同产地玉米样本光谱间的差异,对其进行主成分分析可以发

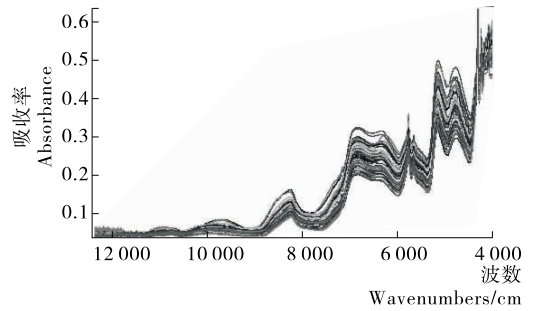


图 1 玉米原始光谱图

Figure 1 Corn raw spectra

现前 5 个主成分(Principal Component, PC)包含了光谱数据的 99% 以上的信息。采用非参数检验(Kruskal Wallis)^[10]的方法对玉米样本的前 5 个主成分间的差异进行检验,所有主成分对应的 $P < 0.05$,表明不同产地的玉米近红外光谱间存在着比较明显的差异。

在将试验所用玉米样本进行分组之前,首先采用马氏距离、杠杆值和学生氏残差检验法对试验中样本进行检验,剔除其中的异常样本,然后将样本分为校正集和预测集两组。不同产地玉米伏马菌素含量测定值分布和分组情况见表 1。

2.2 单一产地的玉米伏马菌素近红外预测结果

采用偏最小二乘法和校正集样本分别建立不同产地玉米的伏马菌素含量预测模型,然后分别采用事先准备好的验证集样本对模型的预测精度进行验证,结果见表 2。采用单一产地玉米样本建立的玉米伏马菌素近红外预测模型在对相应的玉米进行预测时模型具有较好的预测精度(达到 0.9 以上),但该模型用于其他产地的玉米伏马菌素含量进行预测时,预测精度大幅度下降。

2.3 混合产地的玉米伏马菌素近红外光谱模型预测结果

为了克服单一产地建立的玉米伏马菌素近红外光谱模型在对其他产地玉米伏马菌素含量预测时预测结果明显变坏的问题,试验中将 4 个产地的校正集玉米样本进行混合,建立混合产地的玉米伏马菌素近红外光谱预测模型,并采用 4 个产地的玉米对模型的预测精度进行检验,结果见表 3。通过对建模结果的分析,可以发现随着选取建立模型的玉米产地种类混合数量的增加,模型的预测精度在不断提高,将 4 个产地的玉米混合后作为校正集建立的玉米伏马菌素预测模型的精度最佳。通过以上的研究可知,当校正集样本中包含的玉米种类增加,玉米样本的近红外光谱信息也增加,这样建立的预测模型具有更好的适应性,能够减少因玉米产地的差异而导致对玉米伏马菌素预测的影响。

2.4 混合产地模型的优化

为了进一步减小玉米伏马菌素预测过程运算的复杂性同时提高预测的精度,在 4 000~12 000 cm⁻¹ 光谱范围内,采用竞争自适应加权算法(CARS)对 4 个产地的玉米混合后的校正集样本的伏马菌素进行有效特征光谱的筛选。由于通过竞争自适应加权算法得到的最优采样次数会有细微的差异,因此将进行 60 次试验,选取其得到的交叉验证均方根

表 1 玉米样本伏马菌素含量测试结果统计数据

Table 1 Content of fumonisin corn sample test results of statistical data

产地	样本总量	异常样本数	组别	样本数	最大值	最小值	平均值
陕西	80	2	校正集	62	18.57	5.84	9.19
			预测集	16	18.39	6.04	9.02
山西	90	0	校正集	72	15.39	5.79	9.11
			预测集	18	14.67	6.71	8.92
河南	100	5	校正集	76	17.58	6.38	9.58
			预测集	19	16.91	6.47	9.28
黑龙江	90	1	校正集	70	14.38	5.31	8.57
			预测集	18	13.87	6.46	8.89

表 2 单一产地的玉米伏马菌素近红外光谱建模及检测结果[†]

Table 2 Single origin of fumonisin corn near infrared spectrum modeling and testing results

校正集	预测集							
	A		B		C		D	
	RMSEP	R_p	RMSEP	R_p	RMSEP	R_p	RMSEP	R_p
A	0.815	0.906	0.978	0.779	1.481	0.814	1.263	0.661
B	1.117	0.768	0.656	0.911	1.196	0.743	1.325	0.715
C	1.084	0.801	1.120	0.783	0.973	0.901	1.223	0.611
D	1.184	0.708	1.201	0.803	0.993	0.765	1.014	0.907

[†] A 为陕西,B 为山西,C 为河南,D 为黑龙江; R_p 为预测集相关系数;RMSEP 为预测均方差。

表 3 混合产地的玉米伏马菌素近红外光谱模型预测结果[†]

Table 3 Hybrid origin of fumonisin corn near infrared spectrum model predicted results

校正集	预测集									
	A		B		C		D		A+B+C+D	
	RMSEP	R_p	RMSEP	R_p	RMSEP	R_p	RMSEP	R_p	RMSEP	R_p
A+B	0.721	0.906	0.682	0.909	0.699	0.906	0.769	0.908	0.934	0.919
A+C	0.684	0.911	0.946	0.910	0.994	0.911	0.971	0.911	0.789	0.921
A+D	0.674	0.909	0.915	0.907	0.913	0.909	0.947	0.917	0.944	0.917
B+C	0.891	0.910	0.669	0.909	0.877	0.911	0.736	0.912	0.917	0.923
B+D	0.931	0.914	0.897	0.906	0.947	0.914	0.874	0.905	0.981	0.921
C+D	0.879	0.906	0.968	0.915	0.911	0.901	0.881	0.912	0.783	0.925
A+B+C	0.661	0.918	0.681	0.912	0.696	0.926	0.704	0.914	0.644	0.933
A+B+D	0.666	0.921	0.691	0.917	0.691	0.928	0.659	0.924	0.617	0.931
A+C+D	0.698	0.930	0.684	0.926	0.687	0.927	0.699	0.938	0.611	0.937
B+C+D	0.686	0.926	0.669	0.914	0.688	0.927	0.676	0.936	0.641	0.931
A+B+C+D	0.636	0.929	0.641	0.931	0.621	0.937	0.637	0.928	0.607	0.949

[†] A 为陕西,B 为山西,C 为河南,D 为黑龙江; R_p 为预测集相关系数;RMSEP 为预测均方差。

误差(RMSECV)最小的一次,本次竞争自适应加权算法对玉米伏马菌素特征波长的选取过程见图 2。当采样次数为 45 次时系统的交叉验证均方根误差最小,此时对应的光谱特征变量个数为 109。将选定的 109 个特征波长作为输入变量,建立玉米伏马菌素近红外光谱预测模型,其结果见表 4。与全波段建模相比不但减少了预测过程的运算量,且模型的预测精度也有所提升。

采用竞争自适应加权算法剔除了玉米近红外光谱中的大量无效信息,但剩余的 109 个特征波长的光谱中仍然包含着一些具有共线性的冗余信息,接下来采用连续投影算法(SPA)在 109 个有效光谱点中进一步优选,剔除冗余项,得到 22 个玉米近红外光谱特征波长变量,并用这 22 个特征波

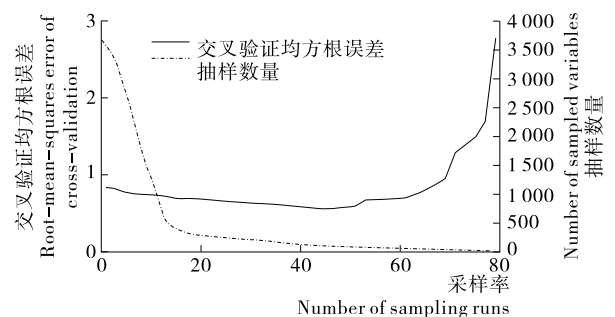


图 2 竞争自适应加权算法对玉米近红外光谱特征变量的筛选
Figure 2 Competition adaptive weighting algorithm for corn screening of near infrared spectral characteristics of the variable

表4 玉米伏马菌素模型优化结果

Table 4 Fumonisin corn model optimization results

模型	变量数	预测集									
		A		B		C		D		A+B+C+D	
		RMSEP	R_p	RMSEP	R_p	RMSEP	R_p	RMSEP	R_p	RMSEP	R_p
PLS	3200	0.636	0.929	0.641	0.931	0.621	0.937	0.637	0.928	0.607	0.949
PLS+CARS	109	0.671	0.932	0.628	0.929	0.611	0.942	0.598	0.932	0.601	0.952
PLS+CARS+SPA	127	0.664	0.936	0.631	0.922	0.621	0.916	0.608	0.927	0.597	0.946

† A为陕西,B为山西,C为河南,D为黑龙江; R_p 为预测集相关系数;RMSEP为预测均方差。

长建立玉米的伏马菌素含量预测模型。与全波段建模相比采用SPA简化后的模型更加简单,其对4个产地玉米的预测集样本伏马菌素检测相关系数 $R_p = 0.954$,预测均方根误差 $RMSEP = 0.417$ 。应用该模型对4个不同产地的玉米伏马菌素预测结果见图3。通过去除光谱中的冗余信息,不但减少了建模与预测过程中的运算量,同时进一步提高了模型的预测精度。

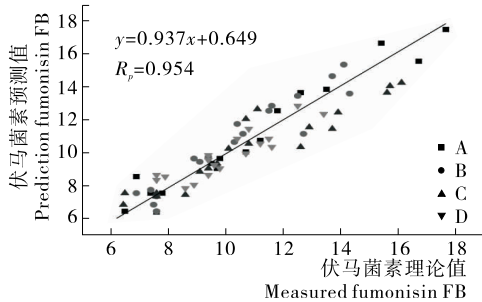


图3 采用CARS+SPA优化模型的预测集样本预测结果与真实值散点图

Figure 3 Used CARS + SPA optimization model set sample prediction results and the real value of a scatter diagram

3 结论

本试验以4个产地的玉米作为研究对象,采用近红外光谱分析技术对玉米伏马菌素检测方法进行了研究。研究发现采用混合产地的玉米样本建立的模型相比采用单一产地玉米样本建立的模型具有更好的适应性。为了减少建模及预测过程的运算量,试验中采用CARS+SPA从光谱的3200个自变量中筛选出有效的22个特征光谱作为系统的输入变量,大大减少了系统的运算量和模型的复杂程度,同时预测结果也有所改善。试验结果表明在对含有不同产地的玉米建立预测模型时,玉米的近红外光谱中含有大量的线性变量,有效剔除这些共线性变量能够提高玉米伏马菌素含量的预测精度,减小由于玉米产地差异对玉米伏马菌素近红外光谱检测的影响。本研究方法能够较好地克服由于不同产地、不同品种导致近红外光谱对玉米伏马菌素含量预测精度低、系统稳定性差的问题,采用本试验方法建立的模型具有更高的适应性。

参考文献

- [1] CAMPS Dachoupan Sirisomboon, REDDY Putthang, PETTR Sirisomboon. Application of near infrared spectroscopy to detect aflatoxigenic fungal contamination in rice [J]. Food Control, 2013, 33(4): 207-214.
- [2] SMITA T, MISHRA L. A rapid FT-NIR method for estimation of aflatoxin B₁ in red chili powder [J]. Food Control, 2009, 20(11): 840-846.
- [3] XIA Chu. Research progress in the technology for rapid detection of mycotoxins in grain and its products [J]. Science and Technology of Cereals, Oils and Foods, 2013, 21(3): 76-81.
- [4] JAMES K. Deoxynivalenol: mechanisms of action, human exposure, and toxicological relevance [J]. Arch Toxicol, 2010, 84: 663-679.
- [5] 袁景, 郭小玉, 杨天溪, 等. 基于光谱技术的食品中常见真菌毒素的快速检测研究进展 [J]. 上海师范大学学报: 自然科学版, 2015, 44(5): 571-579.
- [6] 刘秀英, 申健, 常庆瑞, 等. 基于可见/近红外光谱的牡丹叶片花青素含量预测 [J]. 农业机械学报, 2015, 46(9): 319-327.
- [7] SIMS D A, GAMON J A. Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and development stages [J]. Remote Sensing of Environment, 2002, 81(2/3): 337-354.
- [8] 邹小波, 陈正伟, 石吉勇, 等. 基于近红外高光谱图像的黄瓜叶片色素含量快速检测 [J]. 农业机械学报, 2012, 43(5): 152-156.
- [9] 刘秀英, 王力, 宋荣杰, 等. 黄绵土风干过程中土壤含水率的光谱预测 [J]. 农业机械学报, 2015, 46(4): 266-272.
- [10] 李栓民, 郭银巧, 王克如, 等. 小麦籽粒蛋白质光谱特征变量筛选方法研究 [J]. 中国农业科学, 2015, 48(12): 2317-2326.