doi:10.3969/j.issn.2095-6002.2014.05.014 文章编号:2095-6002(2014)05-0074-06
引用格式:刘翠玲,胡玉君,吴胜男,等. 近红外光谱奇异样本剔除方法研究. 食品科学技术学报,2014,32(5):74-79.
LIU Cuiling, HU Yujun, WU Shengnan, et al. Outlier sample eliminating methods for building calibration model of near infrared spectroscopy analysis. Journal of Food Science and Technology, 2014,32(5):74-79.

近红外光谱奇异样本剔除方法研究

刘翠玲, 胡玉君, 吴胜男, 孙晓荣, 窦森磊, 苗雨晴, 窦 颖
(北京工商大学 计算机与信息工程学院, 北京 100048)

摘 要: 采用近红外光谱分析技术建立面粉校正模型,对面粉中灰分含量进行定量分析,并对异常样本进行剔除. 试验中采用马氏距离法和蒙特卡洛采样法分别对异常样本进行了剔除,结果表明:用马氏距离法剔除异常样本,当权重系数为1.5,剔除样本数为3时,得到最好结果,相关系数(R^2)为92.67,交互验证均方差RMSECV为0.048 5;MCCV法剔除异常样本,剔除样本数为3,得到最好结果,相关系数(R^2)为94.64,交互验证均方差RMSECV为0.041 1. 故马氏距离法剔除异常样本能在一定程度上提高校正模型的精度和预测精度,但MCCV法剔除异常样本后模型精度和预测精度优于马氏距离法.

关键词: 近红外光谱; 异常样本; 马氏距离法; MCCV; 灰分
中图分类号: TS211.7; TS207.3; TP391.9 **文献标志码:** A

人们的日常生活离不开面粉,面粉的品质问题随着生活水平的提高而得到普遍关注,食品监管部门对面粉品质的控制尤为重要. 目前对面粉的评价指标主要有水分、灰分以及面筋等[1-4]. 传统的面粉品质检测方法(物理化学法)存在多种缺陷,不仅耗费时间,而且容易对面粉造成二次污染[5],而被广泛应用于农作物品种检测和分析的近红外光谱分析技术能够在不破坏样品的前提下对样品进行准确、迅速的检测,在一定程度上克服了传统检测方法的缺陷.

近红外光谱分析技术是一种物理测试技术,主要通过建立近红外光谱分析模型对未知的面粉样品进行预测,分析模型的准确程度能够直接影响对未知样品的预测精度[6-8]. 在建立面粉的近红外光谱分析模型时要求面粉的近红外光谱图和化学值之间存在一定的相关性,异常样品的存在能够降低谱图与化学值之间的相关性,降低模型的预测精度,因此需要对异常样本进行判别和处理. 王建义[9]等人对产生异常样品的原因进行了详细的介绍,本文主要探讨马氏距离法以及蒙特卡洛交叉验证法对剔除异常样本后的数据建立近红外光谱分析模型,通过测定模型的准确度对两种方法进行比较,从而提高近红外光谱面粉品质检测模型的精确性和可靠性.

1 试验材料、仪器与方法

1.1 样品的准备

试验所用面粉样本,是从合作单位古船面粉厂取得的不同日期、不同生产线生产的不同种类的面粉,共计60个.

1.2 样品化学值的测量

试验采用国标法850 ℃乙酸镁法,准确测量面粉样本的灰分含量,所测值作为建模时的化学值.

1.3 样品近红外光谱的采集

本次试验使用傅里叶变换近红外光谱仪 VER-

收稿日期:2014-03-06
基金项目:北京市科技创新平台资助项目(pxm_2012_014213_000023);北京市教委科技发展重点资助项目(KZ201310011012);北京市优秀人才基金资助项目(2012D005003000007).
作者简介:刘翠玲,女,教授,博士,主要从事检测技术及智能信息处理方面的研究.

TEX 70，将上述面粉样品放置在漫反射样品台的样品杯中，进行近红外光谱采集. 大样品杯旋转采样，环境温度 23 ~ 25 ℃，扫描次数 64 次，波数范围 12 000 ~ 4 000 cm^{-1}，分辨率 8 cm^{-1}. 对 60 个面粉样本进行近红外漫反射扫描后的光谱图如图 1.

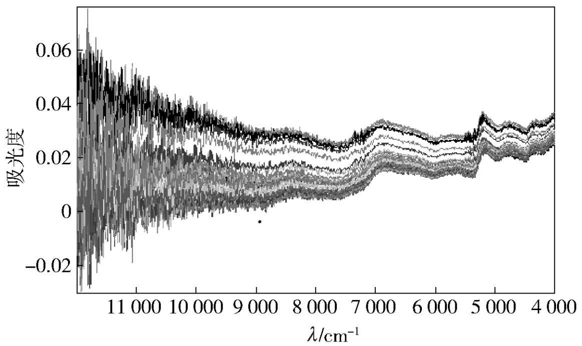

图 1　面粉样本的近红外漫反射光谱图

Fig. 1　NIR diffuse reflectance spectra of flour samples

1.4　马氏距离与蒙特卡洛交叉验证算法

1.4.1　马氏距离算法

将面粉的光谱图转换成数据矩阵后将成为一个 $n \times k$ 的矩阵 $\boldsymbol{A}$.

计算 n 个样品的平均光谱：

$$\overline{A} = \sum_{i=1}^{n} A_{ij}/n, \tag{1}$$

式(1)中，A_{ij}为样品光谱矩阵元素；n 为样品的个数；j 为波长序号；$\overline{A}$ 样品光谱的平均值.

将光谱数据减去平均值做光谱数据中心化处理：

$$\boldsymbol{A}_u = \boldsymbol{A} - \overline{\boldsymbol{A}}, \tag{2}$$

式(2)中，$\boldsymbol{A}_u$ 代表中心化处理后的光谱矩阵，$\boldsymbol{A}$ 代表原光谱矩阵，$\overline{\boldsymbol{A}}$ 代表光谱的平均值阵.

然后计算出原标准光谱数据集的协方差阵：

$$\boldsymbol{M} = \boldsymbol{A}^{\mathrm{T}}{}_u \boldsymbol{A}_u/(n-1), \tag{3}$$

式(3)中，$\boldsymbol{M}$ 代表标准光谱数据集的协方差阵，$\boldsymbol{A}_u^{\mathrm{T}}$ 代表中心化处理后的光谱矩阵的转置，$\boldsymbol{A}_u$ 代表中心化后的光谱阵，n 代表样品数.

根据校正集样品数据和平均光谱数据计算两者之间的马氏距离：

$$D^2 = (A_i - \overline{A})\boldsymbol{M}^{-1}(A_i - \boldsymbol{A})^{\mathrm{T}}, \tag{4}$$

式(4)中，A_i 代表校正集样品数据，$\overline{A}$ 代表平均光谱数据. $\boldsymbol{M}^{-1}$代表标准光谱数据集的协方差阵的逆矩阵.

为了检验 n 个样品中是否存在异常样本，首先要设置一个阈值，这个阈值是根据计算出的 n 个马氏距离设置的. 计算阈值范围如下：

$$D_t = \overline{D} + e \times \sigma_D, \tag{5}$$

式(5)中，$\overline{D}$ 代表马氏距离的平均值；σ_D 代表马氏距离的标准差；e 代表调整闭值范围的参数.

当样品 i 与样品的平均光谱十分相近时，即存在 $D_i \leqslant D_t$，则称之为平均样品的邻近样品. 陈斌[10-13]等人详细介绍了如何通过设置不同的阈值范围参数 e，调节样品的临近样品个数，并采用 PLS 建模进行回归预测，根据预测结果选取最佳 e 值.

1.4.2　蒙特卡洛交叉验证算法

蒙特卡洛交叉验证算法(Monte Carlo cross validation，MCCV)又称为统计模拟方法，能够用于解决复杂统计模型和矩阵高维问题[14-15]. 蒙特卡洛交叉验证算法的核心是对样本的抽取，如何从给定的目标函数分布中进行高效抽样成为关键所在. 蒙特卡洛随机取样(Monte Carlo sampling，MCS)法提出选取一定的校正集(占样品量的 80%)建立偏最小二乘模型，剩余的 20% 作预测集对模型进行验证，经过多次循环后能够得到一组预测残差，通过预测残差计算出预测残差的均值(MEAN)与方差(STD)，从而判断异常样本.

通过校正集相关系数(R^2)、交叉验证均方差 RMSECV、预测均方差 RMSEP 对模型进行评价，从而验证剔除异常样本是否有利于模型精度的提高.

2　结果与讨论

2.1　含异常样品的面粉近红外光谱分析

将 60 个样本应用于近红外定量分析，通过 Kennard-Stone(KS)方法，确定校正集 50 个样本，剩余 10 个样本用于模型验证. 通过 OPUS 6.5 软件的分析和优化，选择最优处理算法，寻找面粉的吸收光谱较丰富的波段. 分析表明，面粉对光谱信息贡献量最大的谱区范围是 4 848.4 ~ 4 246.7 cm^{-1}，维数为 6，利用 PLS 方法进行建模，可得相关系数(R^2)为 85.69，交互验证均方差 RMSECV 为 0.067 2，50 个面粉样本近红外光谱图交叉验证后灰分的近红外计算值与化学分析值如图 2.

部分异常样品的存在使模型的相关系数比较低，模型预测结果缺乏可信度，所以需要把异常样本剔除.

2.2　马氏距离法剔除异常样品

对 50 个校正集样本的近红外光谱进行马氏距离计算，可得到马氏距离分布图，如图 3.

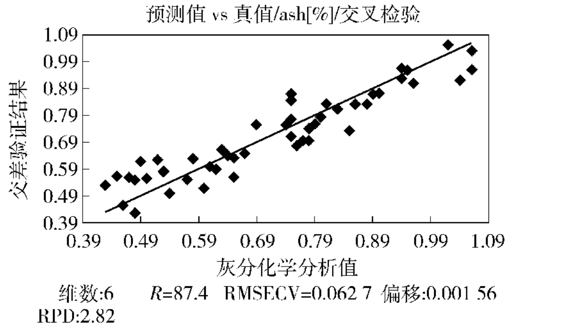

图 2 近红外光谱交叉验证计算值与化学分析值

Fig. 2 Near-infrared spectroscopy cross-validation calculated values and chemical analysis values

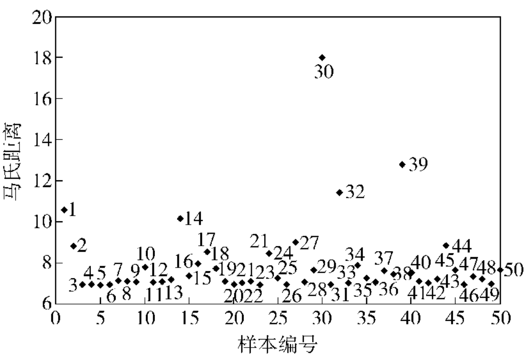

图 3 校正集的马氏距离分布图

Fig. 3 Calibration set Mahalanobis distance distribution

从图 3 中可以看出,一些样品如 30,39 等的马氏距离过大而成为异常样本. 设定 6 个不同的权重系数对异常样本进行判断和分析,可将权重设置为 e(3,2.5,1.5,1.2,1.0, 0.5) ,分别剔除异常样本为:30(e=3);30,39(e=2.5);30,32,39(e=1.5);1,30,32,39(e=1.2);1,14,30,32,39(e=1.0);1,14,27,30,32,39(e=0.5). 剔除异常样品后,对光谱信息贡献量最大的谱区范围 4 848.4 ~4 246.7 cm^{-1}. 采取偏最小二乘方法建模,所得结果如表 1,

表 1 不同阈值剔除后 PLS 校正模型交互校验结果

Tab. 1 Interact verification results of PLS calibration model after removal of different thresholds

权重系数	剔除个数	主成分数	R^2	RMSECV
∞	0	6	85.69	0.067 2
3	1	6	88.47	0.060 5
2.5	2	7	91.78	0.051 6
1.5	3	8	92.67	0.048 5
1.2	4	8	92.48	0.049 5
1.0	5	8	92.35	0.049 1
0.5	6	8	91.60	0.052

马氏距离法剔除异常样品后交叉验证计算值与化学分析值如图 4.

由表 1 可知,当权重系数为 1.5,主成分数为 8,剔除异常样本数为 3 时,得到最好结果,相关系数(R^2)为 92.67,交互验证均方差(RMSECV)为 0.048 5.

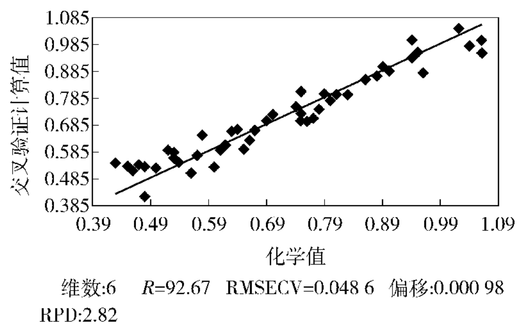

图 4 马氏距离法剔除异常样品后交叉验证计算值与化学值

Fig. 4 Cross-validation calculated and chemical values of Mahalanobis distance method excluding anomalous samples

2.3 蒙特卡洛交叉验证算法剔除异常样本

在 50 个校正集样本中,用蒙特卡洛随机取样法选取校正集和预测集,然后建立偏最小二乘模型,循环 2 000 次后得到各样本的预测残差值,并计算出均值与方差的 MEAN-STD 图,如图 5,为了确定异常样本,绘制误差的火柴梗图,如图 6.

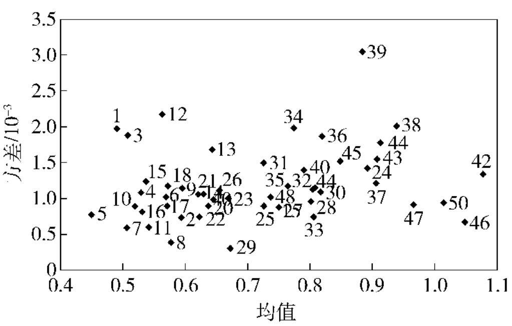

图 5 均值方差分布

Fig. 5 Mean-variance distribution

从图 5 可知,某些样本明显偏离主体样本,如 39,12 这些样本可视为奇异样本,应该剔除,由 MEAN-STD 图和火柴梗图确定出需要剔除异常样本. 奇异样本剔除前后 PLS 校正模型的 RMSECV 的变化情况见表 2. MCCV 剔除异常样品后交叉验证计算值与化学分析值如图 7.

由表 2 可知,剔除异常样品个数为 3,得到最好结果,相关系数(R^2)为 94.64,交互验证均方差 RMSECV 为 0.041 1.

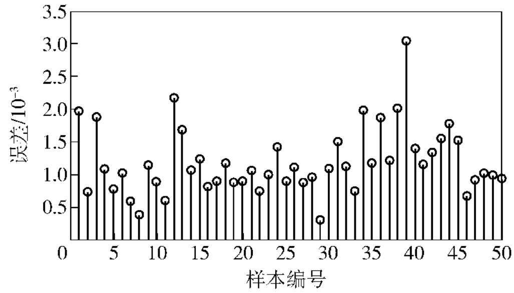

图 6　预测误差的火柴梗图

Fig. 6　Stick Figure of prediction error

表 2　剔除异常样本前后 PLS 校正模型交互校验结果

Tab. 2　PLS calibration model cross validation results after excluding outliers

剔除样本编号	主成分数	R^2	RMSECV
12	6	85. 69	0. 067 2
39	8	93. 42	0. 045 7
12,39	8	93. 97	0. 043 4
1,12,39	8	94. 64	0. 041 1
1,12,34,39	8	94. 44	0. 042 3
1,3,12,34,39	8	93. 90	0. 044 3

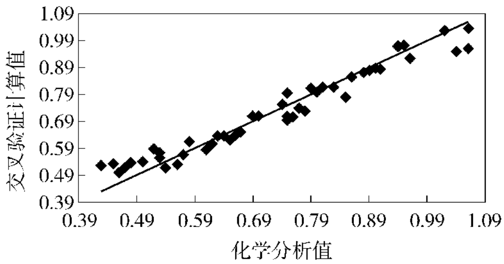

图 7　MCCV 法剔除异常样品后交叉验证计算值与化学分析值

Fig. 7　Cross-validation calculated and chemical analysis values of MCCV law after excluding abnormal samples

2.4　预测模型的精度比较

为了验证剔除异常样本的准确性,对预测集的 10 个样本进行预测,预测结果如表 3. 真实值与预测值之间的相关图如图 8 ~ 图 10.

表 3　剔除异常样本后校正模型的预测结果

Tab. 3　Calibration model predictions after excluding outliers

剔除方法	剔除个数	主成分数	R^2	RMSECV	RMSEP
不剔除	0	6	85. 69	0. 067 2	0. 020 5
马氏距离法	3	8	92. 67	0. 048 5	0. 015 1
MCCV 法	3	8	94. 64	0. 041 1	0. 012 7

由表 3 可知,用马氏距离法和 MCCV 法剔除异常样本后校正模型的精度和预测精度确实有所提高,MCCV 法剔除异常样本模型精度和预测精度提高的相对明显.

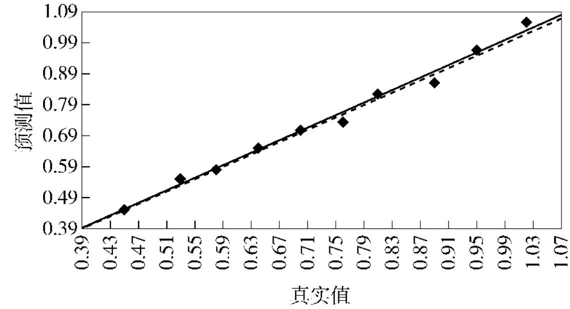

图 8　未剔除样本的模型真实值与预测值相关图

Fig. 8　Real and predicted values correlation chart without excluding sample model

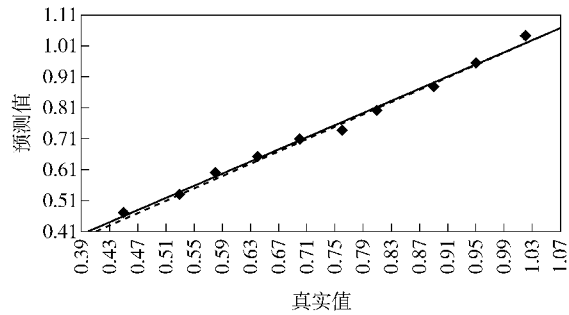

图 9　马氏距离法剔除样本模型真实值与预测值相关图

Fig. 9　Real and predicted values correlation chart with Mahalanobis distance method excluding sample

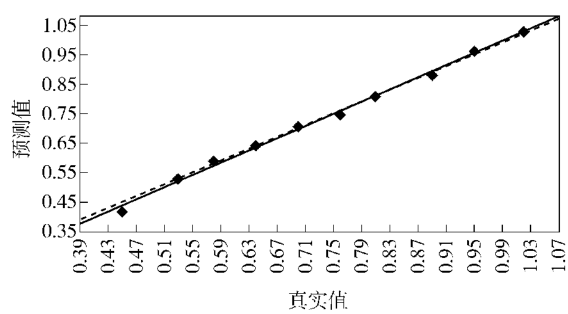

图 10　MCCV 法剔除样本模型真实值与预测值相关图

Fig. 10　Real and predicted values correlation chart with MCCV law Excluding sample

3 结论与展望

本次试验用马氏距离法和蒙特卡洛采样法分别对异常样本进行了剔除，用马氏距离法剔除异常样本，当权重系数为 1.5，剔除样本数为 3 时，得到较好结果，相关系数（R^2）为 92.67，交互验证均方差 RMSECV 为 0.048 5. MCCV 法剔除异常样本，剔除异常样本数为 3 时，得到较好结果，相关系数（R^2）为 94.64，交互验证均方差 RMSECV 为 0.041 1. 结果表明：马氏距离法剔除异常样本确实能提高校正模型的精度和预测精度，但 MCCV 法剔除异常样本模型精度和预测精度提高的相对更明显.

在本次试验中发现，虽然 2 种异常样本剔除方法都使模型精度得到提高并且剔除异常样本的个数相同，但是剔除的样本并不同，可能存在以下问题：1）在没有样本的化学值的情况下，仍然可以采用马氏距离法剔除异常样本，MCCV 法不仅需要光谱数据而且需要样本的化学值，可能存在由于人为误差导致化学值测量不准确，从而导致 2 种方法剔除不同的样本. 2）2 种方法的原理不同，马氏距离法是通过光谱数据验证样本间的距离，MCCV 方法是通过对光谱数据进行多次 PLS 建模验证得到结果，所以这两种方法所得到的结果不同. 目前对剔除异常样本进行了初步的研究，所做的都是验证工作，下一步的工作目标是找到问题存在的原因，并且寻找更好的异常样本剔除方法，从而提高预测模型的准确性和稳定性.

参考文献：

[1] 陆婉珍，袁洪福，徐广通. 现代近红外光谱分析技术［M］. 北京：中国石油化工出版社，2000：37 – 45.

[2] 倪永年. 化学计量学在分析化学中的应用［M］. 北京：科学出版社，2004：304 – 310.

[3] 刘建学. 实用近红外光谱分析技术［M］. 北京：科学出版社，2008：168 – 186.

[4] 邹小波，赵杰文. 农产品无损检测技术与数据分析方法［M］. 北京：中国轻工业出版社，2008：197 – 220.

[5] 闫李慧，王金水，金华丽，等. 基于近红外光谱技术的通用面粉水分无损检测模型的建立［J］. 现代食品科技，2011，27（2）：235.

[6] Karande A D, Heng P W S, Liew C V. In-line quantification of micronized drug and excipients in tablets by near infrared (NIR) spectroscopy: real time monitoring of tabletting process［J］. International Journal of Pharmaceutics, 2010，396：63 – 74.

[7] CHEN Quansheng, PEI Jiang, ZHAO Jiewen. Measurement of total flavones content in snow lotus (saussurea involucrate) using near infrared spectroscopy combined with interval PLS and genetic algorithm［J］. Spectrochimica Acta Part A, Molecular and Biomolecular Spectroscopy, 2010，76：50 – 55.

[8] QU Nan, ZHU Mingchao, MI Hong, et al. Nondestructive determination of compound amoxicillin powder by NIR spectroscopy with the aid of chemometrics［J］. Spectrochimica Acta Part A, Molecular and Biomolecular Spectroscopy, 2008，70：1146 – 1151.

[9] 王建义，雷蒙. 近红外光谱煤质分析模型中异常样品的剔除方法［J］. 工矿自动化，2011，11（11）：75 – 76.

[10] 陈斌，邹贤勇，朱文静. PCA 结合马氏距离法剔除近红外异常样品［J］. 江苏大学学报，2008，29（4）：277 – 279.

[11] 王毅. 近红外光谱分析技术在食用植物油品质检测中的应用［D］. 镇江：江苏大学，2010.

[12] SHAO Yongni, HE Yong. Measurement of soluble solids and pH of Yogurt using visible/near infrared spectroscopy and chemometrics［J］. Food Bioprocess Techno, 2009（2）：229 – 233.

[13] Edword J. Graphical modelling and the mahalanobis distance［J］. Journal of Applied Statistics, 2005, 32（9）：959 – 967.

[14] 李水芳，单杨，范伟，等. 基于 MCCV 奇异样本筛选和 CARS 变量选择法对蜂蜜 pH 值和酸度的近红外光谱检测［J］. 食品科学，2011，32（8）：182 – 184.

[15] LIU Yande, Ying Yibin, JIANG Haiyan. Rapid determination of maturity in apple using outlier detection and calibration model optimization［J］. Transactions of the ASAB E, 2006, 49（1）：91 – 95.

Outlier Sample Eliminating Methods for Building Calibration Model of Near Infrared Spectroscopy Analysis

LIU Cuiling, HU Yujun, WU Shengnan, SUN Xiaorong, DOU Senlei, MIAO Yuqing, DOU Ying

(*School of Computer Science and Information Engineering*, *Beijing Technology and Business University*, *Beijing* 100048, *China*)

Abstract: The accuracy of the prediction model is affected by the near-infrared spectrum of flour and flour ash contents was quantitative analyzed. While the presence of outlier data seriously interfere with the reliability of the model, therefore, it is essential to identify and deal with the outlier samples to improve the predictive ability. Mahalanobis distance and the Monte Carlo cross validation (MCCV) methods were used to remove the outlier samples. When the weight coefficient was 1.5, excluding sample number was 3 with the former method it could get the best results, and the related coefficient (R^2) was 92.67, cross-validation mean square error (RMSECV) was 0.048 5. While with the latter method the correlation coefficient (R^2) was 94.64, cross-validation mean square error (RMSECV) was 0.041 1. Therefore, Mahalanobis distance method can improve the calibration model and prediction accuracy to a certain extent, while the calibration model and prediction accuracy of MCCV without outliers samples was better than that of the Mahalanobis distance method.

Key words: near infrared spectroscopy; outlier samples; Mahalanobis distance; MCCV; flour ash

(责任编辑:檀彩莲)

(上接第 62 页)

Application of HACCP System in Fried White Feather Chicken Nuggets Production

CHEN Qiumei[1], CHEN Yanying[1], CHEN Ming[2], WANG Shaoyun[1,*]

(1. *College of Biological Science and Engineering*, *Fuzhou University*, *Fuzhou* 350002, *China*; 2. *Fujian Shengnong Co. Ltd.*, *Guangze* 350104, *China*)

Abstract: In the production process of fried White Feather Chicken nuggets, biological, physical, and chemical hazard analysis methods based on the HACCP system were applied. The potential hazards existing in pre-processing, frozen treated, and storage process were analyzed and the critical control points and critical limits were revealed. The HACCP plan was made based on the raw materials reception, additives launching, and metal detection to ensure the product quality and safety of fried White Feather Chicken nuggets.

Key words: HACCP system; fried; White Feather Chicken; nuggets; frozen prepared foods

(责任编辑:李　宁)