安全科学与工程学报
Journal of Safety Science and Engineering
第 3 卷◆第 4 期◆版本 1.0◆2025 年
文章类型：论文丨刊号（ISSN）：2972-4716(P) / 2972-4724(O)

基于数据分布矢量化的实验室检测质量风险研判模型的研究

钱仪嘉　韩晶　韩祎陟　梁召
中国海关科学技术研究中心 呼和浩特海关技术中心
DOI:10.32629/jsse.v3i4.17868

[摘　要] 本文旨在通过矢量化分布对实验室机构检验检测数据进行分析,识别数据质量风险、检测质量风险以及食品检测中的潜在问题。通过采集大量食品领域实验室的检验数据,进行数据预处理后,建立基于数据分布矢量化风险研判模型。经过对数据的清洗、统计和风险评估,分析出风险研判结果,帮助机构监管部门对机构进行数字化、靶向性监管,同时帮助检测机构提升内部检测质量,用数字化方式降低质量控制的成本,前置化发现风险点,从提升检测质量的角度促进食品安全工作的良性发展。
[关键词] 数据质量；检测质量；风险评估；食品安全
中图分类号：TS201.6　文献标识码：A

Laboratory Testing Quality Risk Assessment Model Based on Vectorization of Data Distribution

Yijia Qian　Jing Han　Huizhi Han　Zhao Liang
Science and Technology Research Center of China Customs　Technical Center of Hohhot Customs
[Abstract] This paper aims to analyze the inspection and testing data of laboratory institutions through vectorized distribution, so as to identify data quality risks, testing quality risks and potential problems in food testing. By collecting a large amount of inspection data from laboratories in the food field, and after data preprocessing, a risk assessment model based on vectorized data distribution is established. Through data cleaning, statistics and risk assessment, the risk assessment results are analyzed, which can help the institutional supervision departments carry out digital and targeted supervision over institutions. Meanwhile, it can assist testing institutions in improving their internal testing quality, reduce the cost of quality control through digital means, identify risk points in advance, and promote the benign development of food safety work from the perspective of improving testing quality.
[Key words] data quality; testing quality; risk assessment; food safety

引言

随着全球化进程的加速和经济的持续发展,食品供应链日益复杂,食品安全问题成为社会各界广泛关注的焦点。食品安全风险的存在不仅直接关系到消费者的身体健康,还影响着市场的稳定以及国际贸易的顺利进行。实验室检测机构对食品中的各类指标进行检测和分析,为食品安全监管提供数据支持和技术依据。如果检测数据存在质量问题,很可能会导致对食品安全状况的误判,进而影响监管决策的科学性和合理性。

因此,对食品检测数据进行全面、深入的质量评估和风险分析显得尤为重要。通过对检测数据的矢量化分布进行研究,可以发现数据分布中的异常特征和潜在问题,有效识别数据质量风险和检测质量风险。这不仅有助于提升检测机构的工作质量,确保检测结果的准确性和可靠性,更能为食品安全监管提供更有力的支持和保障,有利于保障国际贸易的顺利进行和食品行业的健康发展。

1 数据采集

1.1样本来源

本研究原始数据来源于某实验室lims系统录入的食品抽检与检验数据,取2023年至2025年这三年的数据,其中包括样品与检测项目的详细信息,具体批次与项次数见表1。

表1 原始数据按年份批次统计表

年份	批次	项次
2023	12077	55004
2024	3354	30520
2025	637	3829

1.2数据汇集

原始数据涉及6家不同的检验机构,不同检验机构对于样品名称、检验项目、检验方法等都有不同的命名。为了统一汇集数据,根据国家市场监督管理总局公告(2020年第8号),发布新修订的《食品生产许可分类目录》,该目录采用“食品、食品添加剂类别-类别编号-类别名称-品种明细”的分类层次,共将食品分为32大类287细类。

由此构建了食品的四级分类：食品大类、食品亚类、食品次亚类、食品细类。本研究根据样品名称进行了食品四级分类的匹配工作,并针对检验项目、检验方法、检验结果单位以及检验结果等进行了一系列严谨的数据清洗操作,最终统计结果如下：

在2023-2025年间共检测了16069份抽样单,涉及278个检验项目,共计89354项次,平均每个抽样单检验8个项目,所有样品的抽样单位与生产企业遍及全国。在抽检的样品中,不合格批次为88批次,占总批次的0.5%,总体合格率为99.5%。

样品总计6456种,涵盖314种不同的食品四级分类,广泛分布于32个食品大类之中。其中,食用农产品无论是在批次占比还是项次占比上均居于首位,批次占比达48%,项次占比更是高达65%；位于第二位的是调味品,其批次占比为8.9%,项次占比为6.3%。

本研究以食品次亚类为对象,按照食品次亚类的批次统计信息见表2。

表2 原始数据按食品分类统计表

食品次亚类	批次	批次占比
茄果类蔬菜	1035	6.40%
畜肉	929	5.70%
根茎类和薯芋类蔬菜	803	4.90%
叶菜类蔬菜	729	4.50%
糕点	692	4.30%

通过对数据的汇集,可以清晰地了解到食品抽检的总体情况以及各类食品的质量状况,为后续的食品安全监管和研究提供有力的数据支持。

2 数据预处理

2.1异常值处理

统计的所有数据中,约91%的数据以“<XX”的形式呈现,缺失具体的检测值结果。针对此类检验结果,查询对应的检验方法,对检出限进行核实。将这些异常记录分为两类处理：(1)值小于该检验方法的检出限,定义为“未检出”；(2)值未小于检出限,定义为实测值。

数据记录异常会直接影响数据分析的结果,本研究的定义方式能在一定程度上避免数据记录异常对数据分析结果产生的负面影响,为后续的数据分析与决策提供坚实的数据基础。

2.2重复值处理

从食品抽检规范方面,抽样单编号是抽样单位内部对所采集样品的编号,按《国家食品安全抽样检验抽样单编号规则》编制,一个样品所对应的编号具有唯一性。然而,本研究收集的数据中出现了同一抽样单下检验项目重复的情况,尤其是在细菌类项目上。针对此类问题,本研究对数据进行了严格筛查与清洗,即采用唯一值清洗法,通过比对抽样单编号、检验项目名称等关键信息,识别并剔除了重复记录,确保每一项检验结果的唯一性。

2.3缺失值处理

从数据追溯方面,原始数据未提供所有样品对应的条形码信息,导致无法将抽检结果精确匹配至具体产品批次。这一缺失限制了数据对比分析的精度,削弱了研究结论的可追溯性与实证性。同时,部分样品的生产企业名称存在缺失情况。经深入调研与核查后得知,部分缺失生产企业名称的样品与相似样品属于同一生产企业,由于记录不完整,导致名称未能完整呈现,出现生产企业名称缺失的现象。

为缓解这一问题,本研究依据样品已有的生产企业名称进行产品追溯,通过对生产企业信息的匹配,尽可能缩小抽检结果与具体产品之间的关联范围,提高数据对比分析的可靠性。

3 模型原理

3.1风险原理

为分析实验室的检测质量,从具体的检测值入手,针对特定的食品、检验项目、检验方法,检测值应该较为稳定。原因如下：

(1)食品的同质性：特定食品在原料、生产工艺、储存条件等相对稳定的情况下,其各项质量指标应具有一定的相似性。因此,对于同一食品的多次检测,其检测值应在一定范围内波动,不会出现大幅度的离散。(2)检验方法的精确性：经过规范化的检验方法具有较高的精确度和重复性。在相同的实验条件下,对同一样品进行多次检测,其结果应较为接近,且符合该检验方法的精密度要求。(3)质量控制体系：实验室的质量控制体系包括标准操作程序、仪器校准、人员培训、内部质量控制等环节。这些措施旨在确保检测过程的一致性和稳定性,从而保证检测值的可靠性。

由以上所述,检测值的分布应该具有一定的稳定性。但是,不同的检验机构在检验的过程中可能存在多种因素导致误差,如偶然误差、系统误差、样品问题、地区差异、食品种类的多样性以及抽检的随机性等原因。

基于以上内容,本研究针对实验室检测质量风险判定的原理为：在大数据条件下,根据趋于稳定的数据分布,识别异常分布,再根据异常分布具体的检测值,验证多种维度的因素,排查异常原因。

3.2模型原理

3.2.1基准

从概率论的角度,当样本容量足够大时,某一类数据的分布会趋于稳定。因为食品的命名、样品、细类多种多样,所以需要

搜集大量食品抽检检验数据作为历史大数据。量化分布有多种方式,在本研究中,各个分类下分布的形态是未知且复杂的,需要基于海量的历史数据特点展开深入分析与精准计算,所以本研究使用最常见的分位数来构造分布。

由于抽检数据存在“未检出”结果,未检出并不完全表示检验结果为“0”,而是受仪器精度或数据记录习惯的影响。所以需要单独建立一个区间为“未检出”。因此,在相同的分类条件下,将数据划分为13个区间的占比分布情况。其中,第一个区间被定义为“未检出”。剩余的12个分布区间则由检出值的最小值、10个分位数以及最大值共同构成。通过对这13个区间的历史数据占比进行统计,建立了一个对比“基准”。

基准如同一把精准的标尺,为评估当前相同条件下的检测值数据提供了可靠的参照依据,不仅便于后续的计算与分析,还能够更直观地呈现出数据在不同区间内的分布特征。

3.2.2矢量化度量分布

为了能在相同的条件下处理和分析原始数据,我们将其依据既定基准划分为对应区间,并计算出每个区间的占比。通过这种方式,将离散且复杂的数据矢量化。具体来说,将原始数据转换为一个n维的矢量,每个维度代表一个区间的占比。通过比较原始数据与基准形成的两个矢量之间的差异,可以量化原始数据与历史大数据之间的差异。这种差异的量化不仅能够识别异常值,还能评估数据的分布特征是否与历史数据一致。

3.2.3距离度量分布差异

为了更加准确地对比原始数据与基准形成的两个矢量分布之间的差异,需要构建指标来量化。本研究的分布是由分位数以及“未检出”构成的13个区间,并且分布未知,采用距离的度量方式最为合理。

本研究基于矢量分布,期望通过度量两个分布之间的差异来识别异常,所以需要评估数据的相似性与差异性,因此采用切比雪夫距离与Hellinger距离度量方法,这两种方法可以同时考虑单个维度的最大差异与整体的细微差异,提供了可靠的量化依据,有助于精准识别偏离基准的数据特征。

切比雪夫距离只关注最大值,对个别异常值不敏感。用于需要考虑多维空间中最大差异的场景,公式为:

$$D(x,y)=\max_{i=1}|x_i-y_i|$$

x_i与y_i分别表示待研究数据和基准数据在第i个区间的占比。

Hellinger距离因其优越的数学特性和统计学意义,用于衡量两个概率分布相似性的距离度量。它的值范围在0到1之间,其中0表示两个分布完全相同,1表示两个分布完全不同。适用于统计学、数据科学和机器学习等领域。公式为:

$$H(P,Q)=\frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{n}(\sqrt{P_i}-\sqrt{Q_i})^2}$$

P_i与Q_i分别表示待研究数据和基准数据在第i个区间的占比。

本研究中将全部数据集的80%作为训练数据,20%作为测试数据,以确保模型在广泛的样本上进行训练和验证。

3.2.4机器学习算法在异常检测中的应用

对两个矢量分布之间的差异进行距离量化后,仅通过判定距离多大会成为异常是具有挑战性的。因为距离度量本身只能提供数据点之间的相对差异,而无法直接反映出哪些差异可以被视为异常。这种局限性促使我们引入机器学习算法,以更全面地分析数据特性并有效地识别异常。

机器学习,尤其是无监督学习算法,能够在没有预先标记的训练数据的情况下,自动识别数据中的模式和结构。本研究主要使用异常检测算法,识别数据集中显著偏离正常模式的样本,如孤立森林(iForest,isolationforest)、局部离群因子(LOF)和主成分分析(PCA)等算法,三种算法在多个角度的对比见表3。

表3 不同异常检测算法对比表

算法	训练时间	数据规模适应性	异常值比例适应性	参数敏感性	数据分布假设
孤立森林	快速	大规模	低比例	低	无特定假设
局部离群因子(LOF)	较慢	中等	中等比例	高	局部密度差异
主成分分析(PCA)	中等	中等	低比例	中等	线性关系

综上所述,孤立森林是一种时间复杂度低、适用于大规模数据集、对数据分布无假设,且无需大量正常样本进行训练的算法,能够有效识别少量异常值。在本研究中,数据量较大且异常值比例较低,孤立森林算法因其高效的计算性能和对少量异常值的敏感性,成为更为合适的选择。

本研究将异常分数排名前5%的样本筛选出来,作为潜在异常值进行人工标注。这一比例的选择旨在平衡敏感性和特异性,确保识别出足够多的潜在异常值,同时避免过多的误报。同时人工标注融合了人工经验和专业知识,以提高标注的准确性。

3.3检测质量风险研判模型

基于以上模型原理构建的异常检测模型,可以识别出数据分布中的异常,而异常的表现多种多样,并且需要结合实际判断是否存在真实风险,所以考虑将异常定义为不同的风险。

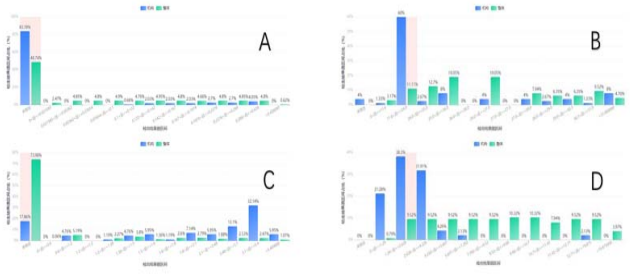

图1 四种可能风险类型的数据分布图

(A)未检出率过高;(B)检出值高度集中;(C)检出值整体偏大;(D)检出值整体偏小

根据距离算法度量结果可以将风险等级定义为高、中、低，而根据数据分布形态以及实验室检测值的实际情况，可以将风险类型定义为：未检出率过高、检出值高度集中、检出值整体偏大或整体偏小等，这四种可能风险类型的数据分布见图1。

此外，由于基准的建立规则受到食品分类的限制，在本模型设定的同一个食品次亚类下，包含多种样品，而不同样品下同一个项目的检测值可能存在较大的差异，因此，不能仅依靠模型判定结果来定义风险，还需要考虑样品的来源、具体产品差异等因素。

为了解决这一问题，本研究进行了抽检独特性分析，在被抽样单位企业所在地市、生产企业所在区县、食品细类三个维度进行验证，在相同分类下，验证这三个维度是否存在异常。如果排除异常，则最终确认风险。并且在历史数据中进行产品独特性分析，找到相同条形码的产品，进行检测值的对比，提供更有利的证据。此外，本研究还结合人工经验与专家判断，对模型的判定结果进行审核。通过人工审核的方式，确保每个潜在风险点都经过全面评估，从而提高风险评估的准确性和可靠性。

通过上述优化措施，本研究建立了一个更为全面和准确的实验室检测质量风险研判模型，为食品安全监管提供了更为可靠的数据支持。

4 模型结果与数据验证

将检测质量风险研判模型运用于原始数据（以下简称风险数据），结合数据本身与产品本身的特性，可以得到具体的风险项见表4。

表4　风险详情

食品次亚类	检验项目	检验方法	检验结果单位	风险类型	风险等级
畜肉	挥发性盐基氮	GB 5009.228-2016	mg/100g	检出值高度集中	高

畜肉中的挥发性盐基氮，采用GB 5009.228-2016半微量定氮法检测，结果单位为mg/100g。基准数据量为4430批次，存在12家其他检验机构，风险数据76批次，均由风险机构A检测。

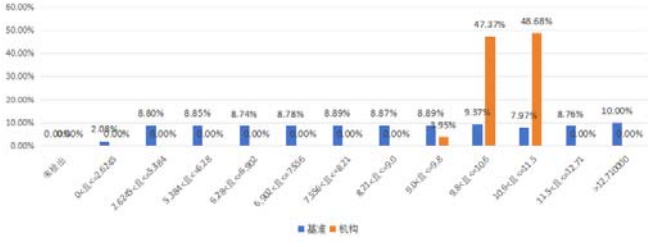

图2 畜肉-挥发性盐基氮-检测值分布

(1)风险研判。通过矢量化分布、基准对比、模型预测三步，提取出该分类下的数据分布存在异常。(2)风险特征。风险机构的数据分布较基准分布差异较大（见图2），有96%的检测值集中在(9.8mg/100g, 11.5mg/100g]区间，呈现“检出值高度集中”。(3)风险验证。在验证被抽样单位地市、生产企业区县、食品细类三个维度下，其他机构与基准分布并没有呈现明显的数据差异，证实该机构风险。

在具体产品对比中，畜肉下有2个食品细类（猪肉和牛肉），提取具体的产品、相同生产企业下的其他机构的检测值进行对比。牛肉之间的差异不大，猪肉存在3个生产企业的同产品对比。具体对比结果见表5。

表5 三个生产企业同产品对比表

生产企业名称	风险机构检测值均值	其他机构检测值均值	差异	风险机构样本量	其他机构样本量	其他机构数
公司A	10.65	8.66	22.98%	28	41	12
公司B	10.81	9.36	15.49%	10	4	2
公司C	10.51	7.94	32.37%	8	34	12
总计	10.65	8.39	26.94%	46	79	12

综上，猪肉中的挥发性盐基氮，在三家同厂商的均值对比下，风险机构的检测值高度集中在较大的区间内。某机构的检测均值(10.65mg/100g)显著高于其他机构(8.39mg/100g)，平均偏差达26.94%。例如，公司A的产品在某机构检测中均值为10.65mg/100g，而其他机构为8.66mg/100g，差异显著(22.98%)。

经过风险研判模型的分析，确证了该机构在畜肉中的挥发性盐基氮的检测值存在一定的风险，可能存在检测数据真实性的问题，也可能是由于设备精度、人员操作规范性、环境等各类因素导致数据出现不准确。

5 结论

本研究提出了一种创新的实验室检测质量风险研判模型，该模型结合了矢量化、距离度量、孤立森林与决策树算法、业务验证。

首先，本研究根据原始数据的特点，对数据进行预处理。根据在大数据下，相同分类的数据分布趋于稳定的原理，搜集大量历史数据构建基准。通过数据分布矢量化，利用切比雪夫距离与Hellinger 距离度量机构分布与基准分布的差异，最终结合孤立森林与决策树算法提出异常检测模型。

其次，为了实现对异常的有效评估，本研究不仅根据实际情况对风险等级、风险类型进行了划分，还基于样品的多样性进行了业务验证，形成完整的检测质量风险研判模型。该模型从异常值检测、抽检独特性、产品特殊性三个角度验证风险的有效性，并且提取具体的产品进行检测值对比验证风险的真实性，提高了准确性与全面性。

实验结果表明，本研究的风险研判模型能够准确识别异常的风险项，在业务验证中提供有利的证据。不仅能提高异常检测的准确性和可靠性，还为食品安全监管和质量控制提供了有力支持。通过综合考虑样品来源、产品差异等因素，模型能够更全面地评估风险，确保检测结果的真实性和有效性。这项研究为实验室检测质量的风险管理提供了新的思路和方法，具有重要的实际应用价值。

[基金项目]

海关总署科研项目(2024HK239)。

[参考文献]

[1]潘盛波,高林,李毅.贵阳市食品检验检测体系建设现状与思考[J].中国初级卫生保健,2016,30(11):50-51.

[2]LIUFT,TINGKM,ZHOUZH.Isolation forest[C]//The IEEE International Conference on Data Mining.IEEE,2008:413-422.

[3]BREUNIG M M,KRIEGEL H P,NG R T.LOF:identifying density-based local outliers[C]//ACM SIGMOD International Conference on Management of Data.ACM,2000:93-104.

[4]蔡玲嘉,李雄,王泽涌,等.基于改进孤立森林算法的审计数据异常主动预警研究[J].信息技术与信息化,2025,(03):131-134.

[5]余翔,陈国洪,李霆,等.基于孤立森林算法的用电数据异常检测研究[J].信息技术,2018,42(12):88-92.

[6]刘超晔,应月,黄孟丽.食品检测实验室检验过程风险评估和控制[J].现代食品,2017,(09):59-62.

[7]李倩,韩斌,汪旭祥.基于模糊孤立森林算法的多维数据异常检测方法[J].计算机与数字工程,2020,48(04):862-866.

作者简介：

钱仪嘉(2001--),女,汉族,北京人,本科,助理工程师,主要从事信息化和网络安全工作。