# 基于 TF-IDF 的食品风险分析模型的构建与应用

姚振民， 邢家溧*， 承 海， 郑睿行， 毛玲燕， 徐晓蓉， 张书芬， 沈 坚
（宁波市产品食品质量检验研究院(宁波市纤维检验所) 浙江宁波 315048）

**摘要** 食品检测数据作为食品风险分析的重要工具，针对同类食品所做检测项目不同而导致最终的数据矩阵部分缺失，且已有的食品检测数据大部分为未检出等问题，通过引入词频-逆文档频率(term frequency-inverse document frequency，TF-IDF)的权重确定办法，构建一种新型的食品风险分析模型。本文以 2019-2020 年为时间段，收集某市食用农产品的蔬菜样本抽检信息作为分析数据，通过模型计算得到蔬菜中各样品的风险指数。结果显示：2019-2020 年间检测的蔬菜产品中，风险指数高的为韭菜和芹菜，超标指数为毒死蜱，在监管中需加强关注，而其余蔬菜大多呈现低风险情况。本分析模型相较于其它传统分析方法，能给出具体的风险指数，在评价上具有直观性，且当数据样本越大，评价效果越好。同时，本模型基于信息理论来设置权重，消除了主观因素在评价中的影响，在应对多样化食品数据时更具有实用性。模型的建立在大数据的时代背景下，对于深入研究食品安全风险及其评价方法新路径提供一个新思路。

**关键词** 蔬菜产品；风险评价；TF-IDF；指数
**文章编号** 1009-7848(2022)12-0324-08 **DOI**：10.16429/j.1009-7848.2022.12.032

民以食为天，食品是人民生存最基本的物质保障。近年来，国民经济水平的大幅提升，人民生活水平大幅改善，食品质量成为消费者关注的重点。食品安全与人民身体健康息息相关，因此业内外各界都格外关注食品安全管理[1]。然而，“固体饮料冒充配方奶粉”“苏丹红”“红心鸭蛋”等食品安全问题的报道屡见不鲜[2]，这些频繁发生的食品安全事件既严重侵犯了消费者的合法权益，也对国家公信力有一定的影响，因此，食品安全问题引起我国政府与有关部门的高度重视[3]。以综合提升食品安全监管效能为建设目标，科学探索食品安全风险分析模型[4]，积极推进风险监测与预警关口有效前移，实现食品安全源头防控和主动预防的相关研究必将被重点关注[5]。

目前，我国食品安全风险评估和风险监测中存在覆盖面窄，数据共享程度不高，资源投入较少等问题[6]。虽然我国食品安全风险评价已开始，但是与其他国家，尤其是发达国家相比仍存在显著差距，因此在评估中不断借鉴新的经验是必不可少的[7]。目前，国际上的主流方法是通过评级或赋值来对风险发生的可能性和严重程度进行等级排序[8]。基于这些不同的方法，研究人员探索出不少风险分级模型，包括定性、定量和半定量模型 3 类[9]。如美国食品药品监督管理局(FDA)向公众开放的一个基于网络的定量风险评估系统——iRISK，被认为是最适合微生物风险分级的方法[10]。这些风险分级模型的分级效果与风险指标体系构建、初始数据形式、使用的方法等有关，其中定量风险分级模型的分级效果最好，然而，缺点是需要有充足的数据支持[11]。目前美国和欧盟的食品安全风险分级模型主要是针对食品中的微生物和化学污染物[12]，国际上尚无通用的食品安全风险分级模型。

国务院《促进大数据发展行动纲要》(国发[2015]50 号)提出：推动大数据应用，建立并不断完善涵盖基础、数据、技术、平台/工具、管理、安全和应用的大数据标准体系[13]。大数据时代的到来，为食品安全分析模型的建立提供了一种全新的思路。原先单一的食品数据分析方法，例如不合格率、超标率等，渐渐不能满足人们对食品安全信息的要求。基于食品安全大数据特点，开展食品安全

**收稿日期**：2021-12-23
**基金项目**：国家市场监管总局科技计划项目(2019MK080，2020MK117)；浙江省基础公益研究计划项目(LGC20C200013)；宁波市自然科学基金项目(202003N4196，2019A610438，2019A610437)；宁波市泛 3315 创新团队(2018B-18-C)；宁波市高新精英创新团队(甬高科[2018]63 号)
**第一作者**：姚振民，男，学士，工程师
**通信作者**：邢家溧 E-mail：hellojiali77@gmail.com

大数据标准化研究，从而构建食品安全风险分析模型并做出深度分析，乃至进一步预测地区风险才是充分利用大数据特征，响应国家号召、顺应时代潮流的重点方向。

本文以 2019 年 12 月到 2020 年 12 月间某市蔬菜产品各项检测数据为研究对象，通过词频-逆文档频率（term frequency-inverse document frequency，TF-IDF）方法建立指标权重体系，以对检测技术参数及标准限量值的客观认知，通过对结果数据矩阵的处理补全，进一步计算得到最终各个样品的具体风险指数。模型的建立将为我国蔬菜安全分级问题分析与风险预警提供理论依据，且可以推广到各类食品，并为深度挖掘食品安全数据信息，探索构建食品安全监测和风险评估新模式提供一种新的思路。

## 1　材料与方法

### 1.1　数据来源

连续收集 2019 年 12 月至 2020 年 12 月间某市蔬菜类农产品指标检测资料。样品抽检依据国家食品安全抽检计划并遵循“双随机”原则，检测方法按照 2019-2020 国家食品安全检测方法标准进行。检测结果判定依据 GB 2763-2019《食品安全国家标准 食品中农药最大残留限量》。

### 1.2　数据预处理

根据国家食品安全风险评估专家委员会的数据采集需求，单项检测项目未检出的占比低于 60%的，检测值定义为 1/2 检出限（Limit of detection，LOD），高于 60%的定义为 LOD。处理完检测结果为未检出的数据之后，将未检测定义为 0，得到矩阵定义为 $G$。

### 1.3　模型开发

采用 TF-IDF 作为权重确定的数值统计方法。该方法原旨在反映单词对集合或者语料库中文档的重要性，在信息检索、文本挖掘和用户建模的搜索中常用作加权因子，其值与单词在文档中出现的次数呈正比的增加，并被包含该单词的语料库中的文档数量所抵消。其中，TF（term frequency）有多种表达方式，包括原始型、布尔型、对数标度型等，本文中 TF 值以 $tf(t,d)$ 表示，而 IDF（inverse document frequency）是对单词提供多少信息的一种度量，即该信息在所有文档中是常见还是稀有（通过将文档总数除以包含该术语的文档数量，然后取该商的对数来获得），本文中 IDF 值以 $idf(t,D)$ 表示。

本质上来说 TF-IDF 是基于信息论出发的一个统计方法，假定 $D=\{d_1,\cdots,d_N\}$ 为一个文本集合，而 $W=\{w_1,\cdots,w_M\}$ 为 $D$ 中的单词集合，其中 $M$ 和 $N$ 分别表示为单词和文本的总数。分别用 $d_j$ 和 $w_i$ 表示 $D$ 和 $W$ 中的元素，用 $D$ 和 $W$ 表示为 $\{d_1,\cdots,d_N\}$ 和 $\{w_1,\cdots,w_M\}$ 中的随机变量，假定 $D$ 中所有元素取到的概率相等且为 $P(d_j)=1/N$，那么，每个文档计算的信息量为 $-\lg(1/N)$，而随机变量 $D$ 的熵为：

$$H(\Delta)=-\sum_{d_j\in D}P(d_j)\lg P(d_j)=-N\frac{1}{N}\lg\frac{1}{N}=-\lg\frac{1}{N} \tag{1}$$

接下来考虑已知 $w_i(\in W)$ 的情况，令 $N_i$ 为含有 $w_i$ 子集中的文档个数，假定各个文档取到的概率相同，则它们的信息量均为 $-\lg(1/N_i)$，在给定 $w_i$ 的情况下，随机变量 $D$ 的熵为：

$$H(\Delta|w_i)=-\sum_{d_j\in D}P(d_j|w_i)\lg P(d_j|w_i)=-N_i\frac{1}{N_i}\lg\frac{1}{N_i}=-\lg\frac{1}{N_i} \tag{2}$$

假设在选定子集中没有 $w_i$ 的文档的出现概率为零，即这些文件没有任何贡献，上述方程中不可能出现 $N-N_i$。

现在，从整个文本中任取一个单词 $w_i$，将 $d_j$ 中 $w_i$ 的频率表示为 $f_{ij}$，而整个文本中 $w_i$ 的频率表示为 $f_{w_i}$，该文本中的所有单词总数表示为 $F$，则有如下等式成立：$\sum_j\frac{f_{ij}}{F}=f_{w_i}/F$，那么交互信息值可以表示为：

$$\begin{aligned}M(\Delta,\Omega)&=H(\Delta)-H(\Delta|\Omega)=\sum_{w_i}P(w_i)H(\Delta)-H(\Delta|w_i)\\&=\sum_{w_i\in W}\frac{f_{ij}}{F}\left(-\lg\frac{1}{N}+\lg\frac{1}{N_i}\right)=\sum_{w_i\in W}\frac{f_{ij}}{F}\lg\frac{N}{N_i}\\&=\sum_{w_i\in W}\sum_{d_j\in D}\frac{f_{ij}}{F}\lg\frac{N}{N_i}\\&=\sum_{w_i}P(w_i)\cdot idf(w_i)\end{aligned} \tag{3}$$

或根据如下等式：$\sum_j \frac{f_{ij}}{F}=f_{w_i}/F$，表示为：

$$M(\Delta,\Omega)=H(\Delta)-H(\Delta|\Omega)=\sum_{w_i}P(w_i)H(\Delta)-H(\Delta|w_i)$$

$$=\sum_{w_i\in W}\sum_{d_j\in D}\frac{f_{ij}}{F}\lg\frac{N}{N_i} \quad (4)$$

等式(3)和(4)分别为以 $f_{w_i}$ 或 $f_{ij}$ 形式表示TF和除以常数 $F$ 的 IDF 的乘积。IDF 因子表示观察到特定单词后信息量的变化，而 TF 因子表示实际观察到该单词的概率估计。(3)和(4)分别表示两个不同的方面，当 TF 表示为 $f_{w_i}$ 时，TF-IDF 意为单词选择的度量，而当 TF 表示为 $f_{ij}$ 时，TF-IDF 意为单词权重的度量[15]。

对比于本文或者语料库的结构构成，食品数据拥有类似的数据结构构成，同样的数据大量缺失、大量重复。将每个样品视为一句单句，每个检测项目作为构成单句的单词，以方法 LOD 值作为衡量单词出现频率的基准线，从而得到一个类似于文本矩阵的数据矩阵，进一步对每个检测项目进行权重的赋值，最终进行风险值的评估。经尝试，从结果来看，该模型效果很好。

### 1.4 数据处理

以每个样品所做各个指标的检测方法的检出限为基准，假设某样品的某个指标检出为 $m$，所用方法检出限为 $n$，则定义该样品的该个指标为 $m/n$，相当于检出 $m/n$ 次。若未检出，则由 1.2 节所述，根据未检出的占比情况分别定义为 1/2 和 1 次。若未检测则表现为 0 次，最终得到一个频数矩阵，不妨定义为 K。

### 1.5 权重的计算

权重的计算分为两个部分，首先计算 TF 值：

$$tf(t,d)=\frac{f_{t,d}}{\sum_{t'\in d}f_{t',d}} \quad (5)$$

式中：$f_{t,d}$——某一指标 $t$ 在样品 $d$ 中出现的次数，即频数矩阵 $K$ 中第 $d$ 行，第 $t$ 列的值。$\sum_{t'\in d}f_{t',d}$——该样本所有频数之和，在 $K$ 上表现为第 $d$ 行所有数的和。

然后，计算 IDF(inverse document frequency)值：

$$idf(t,D)=\lg\frac{N}{|\{d\in D:t\in d\}|} \quad (6)$$

式中：$N$——样品数，即表现为 $K$ 的行数。上式的分母指 $N$ 个样品中，包含 $t$ 指标的样品的数量，在 $K$ 中表现为 $K$ 中第 $t$ 列不为 0 的行数，如果分母为 0，则变为 $|d\in D:t\in d|+1$。

最终将所得的 $tf(t,d)$ 与 $idf(t,D)$ 相乘，得到权重矩阵 $W_0$，

$$W_{0_{d,t}}=tf(t,d)\times idf(t,D) \quad (7)$$

将其作归一化处理，得到：

$$W_{i,j}=\frac{1}{\sum_{i=1}^{m}W_{0_{i,j}}}\times W_{0_{i,j}} \quad (8)$$

式中 $W_{0_{i,j}}$——$W_0$ 第 $i$ 行，第 $j$ 列的值，最终得到 $W$ 为处理后的权重矩阵。

### 1.6 风险程度的分析

重新处理得到的 $F$ 矩阵，为初步处理的结果矩阵，将其各个元素与对应的食品安全国家标准中的最高限量值 MRL 作比值，即：

$$M_{ij}=\frac{F_{ij}}{\mathrm{MRL}} \quad (9)$$

得到的矩阵 $M$ 为各样本的各项指标值，对于其对应的食品安全国家标准中最高限量值 MRL 的比值矩阵，反映各样本各个指标的检测值对应的风险程度。

### 1.7 风险指数的计算

将所得的权重矩阵与风险矩阵各个点对应相乘即：

$$S_{0_{ij}}=W_{ij}\times M_{ij} \quad (10)$$

将 $S_0$ 进一步处理：

$$S_1=\begin{bmatrix} S_{1_{11}} \\ \vdots \\ S_{1_{n1}} \end{bmatrix}=\begin{bmatrix} S_{0_{11}}+S_{0_{12}}+\cdots+S_{0_{1m-1}}+S_{0_{1m}} \\ \vdots \\ S_{0_{n1}}+S_{0_{n2}}+\cdots+S_{0_{nm-1}}+S_{0_{nm}} \end{bmatrix} \quad (11)$$

再将 $S_1$ 进一步处理：

$$\mathrm{S}=\lg(S_1)=\begin{bmatrix} \lg(S_{1_{11}}) \\ \vdots \\ \lg(S_{1_{n1}}) \end{bmatrix} \quad (12)$$

其 $S$ 中大于 0 的为不合格样品，小于 0 的为合格样品，得到的最终评分越大，其风险程度越高。

## 2 结果与分析

### 2.1 基本情况

蔬菜属植物性农产品，是人体所需膳食纤维、维生素、植物化学物质、矿物质等的主要来源，在

我国居民日常膳食中不可或缺的一种食材。本文中用到的数据包括 2019 年 12 月到 2020 年 12 月间共 1 642 批次（12 872 项次，包括各类检测项次），其中包括豆类蔬菜、根茎类和薯芋类蔬菜、茎类蔬菜、叶菜类蔬菜等 12 类 50 种，检测项目包括阿维菌素、倍硫磷、敌敌畏、啶虫脒、毒死蜱、氟虫腈、腐霉利等 26 类。当前可通过市场手段获得的主要蔬菜及种植业常用农药品类均已覆盖。表 1 为部分批次的部分项次。

**表 1 蔬菜产品原始数据表(mg/kg)**

Table 1　Raw data of vegetable products(mg/kg)

| 序号 | 样品名称 | 项目名称 | 检验依据 | 方法检出限 | 最大允许限(MRL) | 检测结果 | 评价 |
|---|---|---|---|---|---|---|---|
| 1 | 包心菜 | 杀扑磷 | NY/T 761-2008 | 0.03 | 0.05 | 0.04 | 符合 |
| 2 | 西蓝花 | 氯唑磷 | GB/T 20769-2008 | 0.00004 | 0.01 | 0.0001 | 符合 |
| 3 | 茄子 | 腐霉利 | GB 23200.8-2016 | 0.01 | 5 | 0.1 | 符合 |
| 4 | 芹菜 | 氧乐果 | NY/T 761-2008 | 0.02 | 0.02 | 0.2 | 不符合 |
| 5 | 豇豆 | 水胺硫磷 | NY/T 761-2008 | 0.03 | 0.05 | 0.22 | 不符合 |
| 6 | 茄子 | 啶虫脒 | GB/T 20769-2008 | 0.00036 | 1 | 0.6 | 符合 |
| 7 | 黄瓜 | 阿维菌素 | GB 23200.8-2016 | 0.005 | 0.02 | 0.008 | 符合 |
| 8 | 包心菜 | 克百威 | NY/T 761-2008 | 0.01 | 0.02 | 0.01 | 符合 |
| 9 | 包心菜 | 甲基毒死蜱 | NY/T 761-2008 | 0.03 | 0.1 | 0.07 | 符合 |
| 10 | 黄瓜 | 敌敌畏 | NY/T 761-2008 | 0.01 | 0.2 | 0.03 | 符合 |
| 11 | 葱 | 乙酰甲胺磷 | GB 23200.8-2016 | 0.01 | 1 | 0.1 | 符合 |
| 12 | 芹菜 | 毒死蜱 | GB 23200.113-2018 | 0.01 | 0.05 | 0.14 | 不符合 |
| 13 | 菜豆 | 溴氰菊酯 | NY/T 761-2008 | 0.01 | 0.2 | 0.02 | 符合 |
| 14 | 青菜 | 甲胺磷 | GB 23200.8-2016 | 0.01 | 0.05 | 未检出 | 符合 |
| 15 | 鸡毛菜 | 铅(Pb 计) | GB 5009.12-2017 | 0.05 | 0.3 | 0.05 | 符合 |

（余下数据省略）

## 2.2　数据初步处理情况

将原始数据按照上述步骤进行处理，并暂时舍去部分不需要因素（如检验依据等），可得到表 2，即初步处理的结果（部分）

**表 2　蔬菜产品初步处理数据表(mg/kg)**

Table 2　Preliminarily processed data of vegetable products(mg/kg)

| 序号 | 样品名称 | 项目名称 | 检测结果 | 修订结果 | 频数 | 风险比值 |
|---|---|---|---|---|---|---|
| 1 | 包心菜 | 杀扑磷 | 0.03 | 0.03 | 1.333 | 0.8 |
| 2 | 西蓝花 | 氯唑磷 | 0.00004 | 0.00004 | 2.5 | 0.01 |
| 3 | 茄子 | 腐霉利 | 0.01 | 0.01 | 1 | 0.02 |
| 4 | 芹菜 | 氧乐果 | 0.02 | 0.02 | 10 | 10 |
| 5 | 豇豆 | 水胺硫磷 | 0.03 | 0.03 | 7.333 | 4.4 |
| 6 | 茄子 | 啶虫脒 | 0.00036 | 0.00036 | 1 666.667 | 0.6 |
| 7 | 黄瓜 | 阿维菌素 | 0.005 | 0.005 | 1.6 | 0.4 |
| 8 | 包心菜 | 克百威 | 0.01 | 0.01 | 1 | 0.5 |
| 9 | 包心菜 | 甲基毒死蜱 | 0.03 | 0.03 | 2.333 | 0.7 |
| 10 | 黄瓜 | 敌敌畏 | 0.01 | 0.01 | 3 | 0.15 |
| 11 | 葱 | 乙酰甲胺磷 | 0.01 | 0.01 | 10 | 0.1 |
| 12 | 芹菜 | 毒死蜱 | 0.01 | 0.01 | 14 | 2.8 |
| 13 | 菜豆 | 溴氰菊酯 | 0.01 | 0.01 | 2 | 0.1 |
| 14 | 青菜 | 甲胺磷 | 未检出 | 0.01 | 1 | 0.2 |
| 15 | 鸡毛菜 | 铅(以 Pb 计) | 0.05 | 0.05 | 1 | 0.133 |

（余下数据省略）

### 2.3 指标权重构建

权重的建立是整个模型构建的重中之重,权重设置的好、坏影响模型构建的结果以及最终风险评估的结果。方法学上来说,某一检测指标在该样品中的检测值越大,即上述计算得到的频数越大,则在该样品中所占的权重越大,其对该个样本的贡献度就越大[15],而包含该检测项目的样本数量越多,该项目整体权重被抵消的越多。通过计算得到各样品各个检测项目的权重如下表 3 所示。

**表 3 蔬菜产品权重表**

Table 3 Weights of vegetable products

| 序号 | 阿维菌素 | 倍硫磷 | 敌敌畏 | 啶虫脒 | … | 毒死蜱 |
|---|---|---|---|---|---|---|
| 1 | 0.057783132 | 0 | 0 | 0.0157726 | … | 0.017456269 |
| 2 | 0 | 0 | 0 | 0 | … | 0.001733576 |
| 3 | 0.058723353 | 0 | 0 | $4.80877\times10^{-5}$ | … | 0.01774031 |
| 4 | 0.066381235 | 0 | 0.018500756 | 0 | … | 0.040107509 |
| 5 | 0.058706305 | 0 | 0 | $4.80877\times10^{-5}$ | … | 0.01773516 |
| 6 | 0 | 0 | 0.029047634 | 0 | … | 0.006297193 |
| 7 | 0 | 0 | 0 | 0 | … | 0.001733576 |
| 8 | 0.058723353 | 0 | 0 | $4.80877\times10^{-5}$ | … | 0.01774031 |
| 9 | 0 | 0 | 0.032063656 | 0 | … | 0.006951032 |
| 10 | 0 | 0 | 0 | 0 | … | 0.001733576 |
| 11 | 0.058723353 | 0 | 0 | $4.80877\times10^{-5}$ | … | 0.01774031 |
| 12 | 0 | 0 | 0.030309936 | 0 | … | 0 |
| 13 | 0 | 0 | 0 | 0 | … | 0.001733576 |
| 14 | 0 | 0 | 0.05883654 | 0 | … | 0 |
| 15 | 0.058723353 | 0 | 0 | $4.80877\times10^{-5}$ | … | 0.01774031 |
| （余下数据省略） | | | | | | |

注:原序号为样品编号,因涉及保密问题,故用序号代替。

### 2.4 风险值计算

经计算得到各样本的单独风险值见表 4（从高到低排序）。

### 2.5 结果分析

最终得到的风险值越大该样本的风险程度越高,且以 0 作为分界线,当最终的风险值大于 0 时,样本为不合格样本,小于 0 时为合格样本。

2.5.1 单个样本情况分析 从单个样本来看,对于风险值最高的,数值为 1.66333 的批次,样品名为芹菜,甲基异硫磷的检测结果为 0.48 mg/kg,为国家标准最大允许限量值(0.01 mk/kg)的 48 倍,该样品的各指标权重分布情况为:甲基异硫磷的权重占到该样本的 0.95，其余 26 个项目之和为 0.05。风险值第二高的样品名为韭菜,经比对,检测指标中腐霉利的含量为 6.8 mg/kg,为国家标准最大允许限量值(0.2 mg/kg)的 34 倍,该样品的各个指标权重分布情况为:腐霉利的权重占到该样本的 0.94,其余 26 个项目权重之和为 0.05。研

**表 4 蔬菜产品风险值表**

Table 4 Risk values of vegetable products

| 序号 | SUM | 风险值(LOG) |
|---|---|---|
| 1 | 46.06155145 | 1.663338562 |
| 2 | 32.01317839 | 1.505328795 |
| 3 | 18.46944194 | 1.266453773 |
| 4 | 11.97826297 | 1.078393843 |
| 5 | 8.721995443 | 0.940615855 |
| 6 | 8.087647235 | 0.9078222 |
| 7 | 6.662618587 | 0.823644952 |
| 8 | 5.774176853 | 0.761490081 |
| 9 | 5.575727587 | 0.746301547 |
| 10 | 5.280350633 | 0.722662762 |
| 11 | 4.764864091 | 0.678050518 |
| 12 | 3.577680791 | 0.553601589 |
| 13 | 3.504289518 | 0.54459998 |
| 14 | 3.012913633 | 0.478986683 |
| 15 | 2.94130133 | 0.468539519 |

注:原序号为样品编号,因涉及保密问题,故用序号代替。以上数据均用 python 计算得到。

究表明最终风险值大于 1 的都有某项指标显著超过国家标准最大允许限量值。风险值仅为 0.0586 的是青菜的样本,其啶虫脒超标了 1.5 倍,相较于其它不合格产品,相对来说风险性较小。

最终的风险值情况并不是完全取决于某样品是否某个指标超标倍数。例如有一样本为茄子,氧乐果检测结果为 0.36 mg/kg,超过国家标准最大允许限量值(0.02 mg/kg)的 18 倍,比韭菜(毒死蜱超标了 10 倍)风险值高。具体原因有很多种,根本原因是所有含有该检测项目的样本数不同, 导致权重有一定的差异, 以至于最后在计算风险值时有一定的差距。

通过以上分析, 从较为关注的不合格批次来说,单项风险因子超标较多的批次,最终通过模型得到的风险值也较大, 而部分合格批次最终风险值靠近 0 的也有某一项风险因子被检出, 其值甚至靠近定量限。综上,由单个样本的分析得出,该模型对于单项样本的风险度有较好的反馈。

总的来说,通过分析 2019 年 12 月到 2020 年 12 月间蔬菜产品单体数据情况,可以得出芹菜和韭菜是需重点关注的对象, 而毒死蜱和腐霉利作为其中具有较高风险的项目需引起关注。

2.5.2　总体样本情况分析　本文采用数据共 1 642 批次,各类风险因子检测 12 871 项次,其中不合格批次为 27 批,不合格率为 1.64%。各类因子检出 528 项次,检出率为 4.10%,其中不合格项为 27 项次,占检出项的 5.30%。在统计的 26 种风险因子中(表 5),啶虫脒、腐霉利、氯氟氰菊酯和高效氯氟氰菊酯的检出率和不合格率分别为 50.84%和 0.63%,30.46%和 2.65%,8.45%和 0.20%,其中腐霉利不合格率显著高于蔬菜风险因子平均水平($P$<0.01,卡方检验),且不合格批次都为鳞茎类蔬菜(芹菜),需特别关注,其余不合格率均正常分布。

由表 5 可知,同一个风险因子,检出率与不合格率并无相关性, 比如啶虫脒的检出率高达 50.84%,而其不合格率处于正常水平,同样的甲拌磷检出率虽然较低,但是不合格率高居第二。

经过计算不合格批次的风险值均大于 0,通过 TF-IDF 权重确定方法以及模型最终得到的风险值如图 1 所示(从高到低排序)。风险值在 0 以上均为不合格,0 以下均为合格。可以看出绝大部分的蔬菜产品都为低风险甚至无风险产品。

**表 5　风险因子分析表**

Table 5　Risk factor data

| 风险因子 | 检出率/% | 不合格率/% |
|---|---|---|
| 腐霉利 | 30.46 | 2.65 |
| 甲拌磷 | 0.78 | 0.78 |
| 啶虫脒 | 50.84 | 0.63 |
| 氧乐果 | 0.46 | 0.33 |
| 甲基异柳磷 | 0.28 | 0.28 |
| 氯氟氰菊酯和高效氯氟氰菊酯 | 8.45 | 0.20 |
| 水胺硫磷 | 0.11 | 0.11 |
| 毒死蜱 | 0.18 | 0.06 |

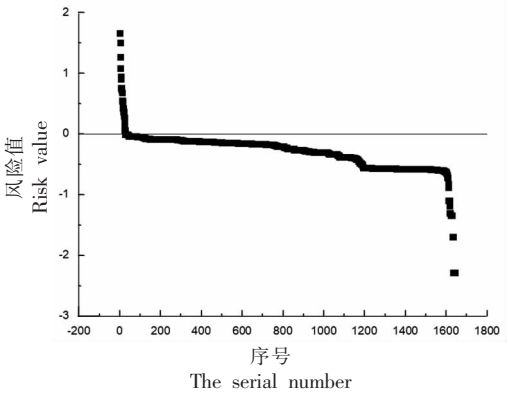


**图 1　最终风险值总体情况**

Fig. 1 The overall performance of final value

考虑到蔬菜具有季节性等特点, 为了更直观反映蔬菜产品风险值与时间的关系, 以时间为横坐标,上文所得风险值作为纵坐标作图,分析蔬菜安全风险变化情况。图 2 为 2019.12 至 2020.11 期间风险值(取每个月平均值)与时间关系图。总体来说, 这段时间的蔬菜风险情况都维持在风险度较低的水平,2019 年 12 月至 2020 年 1 月,2020 年 12 月至 2021 年 1 月的风险趋势类似, 同时也呈现出一定的季节性波动, 夏季相对较低而春秋冬季相对较高,符合一般蔬菜的播种收获情况。此结果表明,基于 TF-IDF 的权重确定模型,不仅单个样本在一定程度上反映具体的风险超标情况,而且对蔬菜整体的季节性有一定的反馈, 反映该模型总体的评价效能,因此,该模型符合对风险概念的认知和应用。

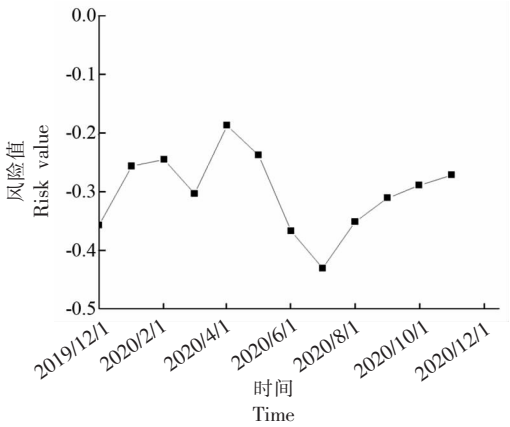


**图 2 蔬菜产品风险值与时间的关系**

Fig.2 Relationship of vegetable products between risk value and time

## 3 结论

鉴于食品检测数据的复杂性，为进一步探索能客观合理分配权重,能更好、更直观体现食品风险程度的方法模型,本文尝试权重赋值方法(TF-IDF) 在食品领域的重新架构应用。收集 2019 年 12 月至 2020 年 12 月间的蔬菜产品的各项检测指标,通过一系列计算分析,将该模型应用于食品风险研判。通过本模型的构建,原先模糊的食品安全风险概念变成具体的指数形式，这对于风险有了更加直观的定义。相较于传统单因素分析方法，该方法引入机器学习的内容，在大数据时代背景下，对于深入研究食品安全风险及其评价方法提供一条新路径。

## 参 考 文 献

[1] 柴懿娜. 浅谈食品安全问题的现状和原因以及应对[J]. 食品安全导刊，2021，15(9)：30-31.
CHAI Y N. Brief discussion on the status quo and causes of food safety problems and countermeasures [J]. China Food Safety Magazine，2021，15 (9)：30-31.

[2] 李晶，张滨. 基于大数据挖掘的食品安全风险智能监测模型[J]. 食品工业，2021，42(4)：384-387.
LI J，ZHANG B. Intelligent monitoring model of food safety risk based on big data mining[J]. The Food Industry，2021，42(4)：384-387.

[3] 陈建千. 关于食品安全的现状及对策探讨[J]. 食品安全导刊，2021，15(9)：36-37.
CHEN J Q. Discussion on the current situation and countermeasures of food safety[J]. China Food Safety Magazine，2021，15(9)：36-37.

[4] 潘焰琼. 风险评估在食品安全监管中的作用[J]. 现代食品，2019，5(15)：112-114.
PAN Y Q. The Role of Risk Assessment in Food Safety Regulation[J]. Modern Food，2019，5(15)：112-114.

[5] 承海，邢家溧，郑睿行，等. 基于熵权-模糊分析法的农产品农药残留安全风险综合评价[J]. 中国食品学报，2021，21(5)：331-339.
CHEN H，XING J L，ZHENG R X，et al. Comprehensive safety evaluation of pesticide residue pollution of agricultural products based on entropy weight-fuzzy mathematics method[J]. Journal of Chinese Institute of Food Science and Technology，2021，21(5)：331-339.

[6] 张旭伟，马勇，张朝飞. 食品安全风险评估与风险监测[J]. 中国食品，2021，50(5)：123-124.
ZHANG X W，MA Y，ZHANG C F. Food safety risk assessment and monitoring [J]. China Food，2021，50(5)：123-124.

[7] 刘莉治. 风险评估在食品安全监管中的作用[J]. 中国食品，2021，50 (5)：120-121.
LIU L Z. The Role of Risk Assessment in Food Safety Regulation [J]. China Food，2021，50 (5)：120-121.

[8] 陈尚，周少君，邓小玲,等. 食品中化学物危害风险分级研究进展[J]. 中国食品卫生杂志，2017，29(3)：374-378.
CHEN S，ZHOU S J，DENG X L，et al. Research progress on hazardous risk ranking of chemical substances in food[J]. Chinese Journal of Food Hygiene，2017，29(3)：374-378.

[9] 朱江辉，宋筱瑜，王晔茹，等. 食品微生物风险分级研究进展[J]. 中国食品卫生杂志，2015，27(3)：322-329.
ZHU J H，SONG X Y，WANG H R，et al. Progress of risk ranking for food microbiological hazards[J]. Chinese Journal of Food Hygiene，2015，27(3)：322-329.

[10] CHEN Y，DENNIS S B，HARTNETT E，et al. FDA-iRISK-a comparative risk assessment system for evaluating and ranking food-hazard pairs：case studies on microbial hazards[J]. Journal of Food Protection，2013，76(3)：376-85.

[11] EFSA. Scientific opinion on the development of a risk ranking toolbox for the EFSA BIOHAZ Panel[J]. EFSA Journal, 2015, 13(1): 3939.
[12] 王芳, 孙晓红, 陶光灿. 中国食品安全风险分级研究进展[J]. 食品科学, 2021, 42(21): 271-277.
WANG F, SUN X H, TAO G C. Progress in Risk Ranking for Food Safety in China[J]. Food Science, 2021, 42(21):271-277.
[13] 张飞, 蒋云, 杨斌, 等. 食品安全大数据标准化存在问题及对策[J]. 食品工业, 2021, 42(4): 329-334.
ZHANG F, JIANG Y, YANG B, et al. The problems and countermeasures of big data food safety standardization[J]. The Food Industry, 2021, 42(4): 329-334.
[14] AKIKO A. An information-theoretic perspective of tf-idf measures[J]. Information Processing and Management, 2003, 39(1): 45-65.
[15] ROBERTSON S. Understanding inverse document frequency: on theoretical arguments for IDF[J]. Emerald Insight, 2004, 60(5): 503-520.

## The Building and Applying of Food Risk Analysis Model Based on TF-IDF

Yao Zhenmin[1], Xing Jiali[*], Cheng Hai, Zheng Ruihang, Mao Lingyan, Xu Xiaorong, Zhang Shufen, Shen Jian
(*Ningbo Academy of Product and Food Quality Inspection (Ningbo Fibre Inspection Institute), Ningbo 315048, Zhejiang*)

**Abstract** Food testing data is an important tool for food risk analysis. The final data matrix is missing due to different testing items for similar foods, and most of the existing food testing data is undetected. Through the introduction of TF-IDF (The term frequency-inverse document frequency) weight determination method has constructed a new type of food risk analysis model. This paper uses the sampling information of vegetable samples of edible agricultural products in a city from 2019 to 2020 as the research data, and calculates the risk index of each sample in the vegetable through the model. The results show that among the vegetable products tested from 2019 to 2020, the high-risk index is leeks and celery, and the over-standard index is chlorpyrifos, which needs to be paid more attention in supervision, while most of the remaining vegetables are low-risk. Compared with other traditional analysis methods, this analysis model can give a specific risk index, is intuitive in evaluation, and shows better evaluation performance in big data analysis. At the same time, this model sets weights in an objective and universal mode, which eliminates the influence of subjective factors in the evaluation and further enhances the practicability in diversified data analysis. The model is established in the context of the era of big data, and provides a new way of thinking for further in-depth research and exploration of new paths for food safety risks and evaluation methods.

**Keywords** vegetable products; risk assessment; TF-IDF; index