Applied Mathematics, Modeling and Computer Simulation
C.-H. Chen et al. (Eds.)

doi:10.3233/ATDE221100

Risk Classification Study of Carbofuran in Vegetables Based on K-Means++ Algorithm

Tongqiang JIANG [a,b], Qi YANG[a,b], Tianqi LIU[a,b], Wei DONG[a,b,1], Yongchun JIAO[a,b], Qingchuan ZHANG[a,b]
[a] *School of E-Business and Logistics, Beijing Technology and Business University, Beijing, China*
[b] *National Engineering Research Centre for Agri-Product Quality Traceability, Beijing Technology and Business University, Beijing, China*

Abstract. This study used dietary exposure assessment method and K-means++ clustering algorithm to construct a carbofuran risk grading model to assess the risk of carbofuran in six vegetable categories in 20 provinces in China using carbofuran sampling and testing data and consumption data in 2020. The clustering algorithm classified the risk level of carbofuran in vegetables into 3 levels: low, medium, and high. The number of low risk combinations accounted for 92.5%, and the high-risk combinations were bulb vegetables in Hebei and leafy vegetables in Shaanxi. This study uses objective data to build a risk classification model and a data-driven approach to risk classification, enhancing the objectivity and validity of the experimental results.

Keywords. Risk assessment, K-means++, risk classification, carbofuran, vegetables

1. Introduction

Carbofuran is a common pesticide used for the control of pests in crops. Carbofuran is harmful to humans because it can be absorbed by the body, and the distribution and metabolism in the body are rapid. Carbofuran poisoning includes symptoms such as headache, nausea, and convulsions.

Food risk classification is an emerging form of risk assessment that integrates factors such as contaminant concentration levels, consumers' dietary exposures, and the degree of public health hazards, and then ranks the risks in a hierarchy that allows for more scientific and rapid identification of risk levels and facilitates better resource allocation by decision makers [1].

In the existing studies, food risk classification methods all suffer from the problem of quantifying risk values and determining risk levels too subjectively. The clustering algorithm is data-driven and classifies samples according to their similarity, which can enhance the objectivity of experimental results [2]. K-means algorithm, as a classical clustering method based on distance division, has the characteristics of simplicity,

1 Corresponding Author, Beijing Technology and Business University, Beijing, China; E-mail: dongwei2019@btbu.edu.cn.

speed and wide applicability. K-means++ algorithm is an optimized improvement of it, which can gives preference to points that are farther away when selecting the initial points, and the clustering results are less influenced by the initial points[3]. Therefore, the K-means++ clustering algorithm was used for conducting a risk classification study of carbofuran in vegetables.

In summary, a food safety risk classification model based on sampling data was developed in this study. Data were obtained from the national sampling data and vegetable consumption data of each province in vegetables in 2020. The dietary exposure assessment method was used to calculate the dietary intake of carbofuran through different types of vegetables for residents of 20 provinces in China respectively, and finally the K-means++ clustering algorithm was used to determine the risk level.

2. Data

2.1 Data Sources

There were 12,940 samples of contaminant level data, all from the 2020 National Food Safety Sampling Database.We obtained data on the consumption of China's residents by region from the Fifth China Total Diet Study.The study subdivided vegetables into six categories: leafy vegetables, root potatoes, legumes, melons, bulbs and eggplants.Toxicological data were obtained from the Joint Meeting on Pesticide Residues.

2.2 Data Pre-processing

In this study, the data were pre-processed on the grounds of the principles proposed by WHO GEMS, replacing the non-detected values for all contaminants with a detection limit of 1/2.

3. Food Risk Assessment Model

3.1 Evaluation Indicators

In this study, Hazard Quotient (HQ), Harm Index (HI) and Nemerow integrated pollution index (NIPI) were selected as model indicators according to the method of contaminant risk assessment. The mean, median, 95th percentile and maximum values of carbofuran sampling data were chosen as food contamination characteristics to calculate the exposure to carbofuran at different contamination standards.

3.1.1 Chronic Risk Assessment Indicators

This system uses the risk quotient (HQ) to evaluate the health risk of carbofuran in vegetables [4]. An indicator value greater than 1 indicates that there is some risk, and the greater the HQ value, the greater the risk; an indicator value less than 1 indicates

that there is no risk, and the smaller the HQ value, the lower the risk level.The model is expressed as:

$$HQ_{m,n} = \frac{EDI_{m,n}^{50}}{ADI} \tag{1}$$

$$EDI_{m,n}^{50} = \frac{FC_{m,n} \times X_{m,n}^{50}}{W} \tag{2}$$

In equations (1) and (2), the $HQ_{m,n}$ is the chronic dietary intake risk of carbofuran in vegetable *n* in province *m*. $EDI_{m,n}^{50}$ is the estimated daily intake of carbofuran via vegetable *n* in province *m* at a medium exposure. $FC_{m,n}$ is the average daily consumption of vegetable *n* in province *m* (kg/d). $X_{m,n}^{50}$ is the median of the detected values of carbofuran in vegetable *n* in province *m* (mg/kg); W is the average body weight of the population, with a value of 60 kg.

3.1.2 Acute Risk Assessment Indicators

The hazard index (HI) was used to assess the acute toxicity risk of carbofuran ingestion in food [5], with HI > 1 indicating the presence of some risk, the larger the HI value, the greater the risk present; HI < 1 indicating no risk, and the smaller the HI value, the lower the degree of risk. The model is expressed as:

$$HI_{m,n} = \frac{EDI_{m,n}^{95}}{ARfD} \tag{3}$$

$$EDI_{m,n}^{95} = \frac{FC_{m,n} \times X_{m,n}^{95}}{W} \tag{4}$$

In equations (3) and (4), the $HI_{m,n}$ is the acute risk of carbofuran in vegetable *n* in province *m*. $EDI_{m,n}^{95}$ is the estimated daily intake of carbofuran via vegetable *n* in province *m* at high exposure. $FC_{m,n}$ is the average daily consumption of vegetable *n* in province *m* (kg/d). $X_{m,n}^{95}$ is the 95th percentile of carbofuran sampled in vegetable *n* in province *m* (mg/kg). W is the average body weight of the population, with a value of 60 kg.

3.1.3 Nemerow Integrated Pollution Index

The Nemerow integrated pollution index (NIPI) calculates the degree of carbofuran contamination in the sampled samples, reflecting the food contamination characteristics [6].The model is expressed as:

$$NIPI_{m.n} = \sqrt{\frac{P_{\max(m.n)}^2 + P_{ave(m.n)}^2}{2}} \tag{5}$$

In equation (5), $P_{max(m,n)}$ is the maximum value of pollution index in vegetable *n* in province *m*. $P_{ave(m,n)}$ is the average value of the pollution index in vegetable *n* in region *m*.

$$P_{m,n} = \frac{X_{m,n}}{S_n} \tag{6}$$

In equation (6), the $P_{m,n}$ is the contamination index in vegetable n in province m.$X_{m,n}$ is the test value of carbofuran in vegetable n in province m.S_n is the evaluation standard of carbofuran in vegetable n.

3.2 Clustering Classification

The above three indicators were constructed to quantify the risk factors of contaminants by integrating the likelihood of exceedance, exposure and hazard of food contaminants, and to establish a food safety risk assessment model. K-means clustering algorithm is the most common clustering algorithm, which can well identify the clusters of convex data, and its results are not affected by the order of the input data. However, in K-means clustering, the initial clustering centroids are chosen randomly, which may cause the problem that the clustering results differ greatly from the actual distribution of the data. Since the K-means++ clustering algorithm gives preference to points that are farther away when selecting the initial points, and the clustering results are less influenced by the initial points, this study uses the K-means++ algorithm and compares them by contour coefficients to select the better clustering method.

The main process of K-means ++ clustering algorithm is as follows.

Input: number of data samples of n, number of clusters k.

Output: the set of k clusters and their class numbers.

Steps.

(1) Select any k objects from the n data objects as initial clustering centers.

(2) Calculate the distance between the n samples and the nearest cluster center, set as D(X).

(3) Selecting new data as clustering centers, where the selection probability is proportional to the size of D(X).

(4) Iterate steps (2) to (3) until all k clustering centers have been selected, then the iteration ends.

(5) So far all the k initial clustering centers have been selected, and then traditional K-means clustering is run.

How to determine the number of grading levels scientifically and rationally is one of the main problems faced in this study. This study expects to determine the number of risk classes of food contaminants using clustering algorithm, which corresponds to the problem of the number of clusters in the algorithm.The silhouette coefficient is a common evaluation method for good or bad clustering [7], and is evaluated from two perspectives: the degree of cohesion and the degree of separation, which takes values between [-1,1], and the closer the value is to 1, the higher the similarity of samples within clusters and the lower the similarity of samples between clusters, and the better the clustering performance. The data-driven approach to risk grading in this study can enhance the objectivity of risk grading. The expression of the silhouette coefficient is as follows.

$$S = \frac{b-a}{max(a,b)} \tag{7}$$

In equation (8), S is the silhouette coefficient, a is the average of its distance to every other sample in the same category, and b is the average distance to all samples in the cluster adjacent to it.

4. Experimental Results

After obtaining the three index values of HQ, HI, and NIPI based on the above, the obtained index values need to be normalized in order to avoid the influence of different orders of magnitude of the three index values on the clustering effect. Then, the number of dietary intake risk classification of carbofuran in vegetables was determined by silhouette coefficients, and the silhouette coefficients of different categories are shown in figure 1.

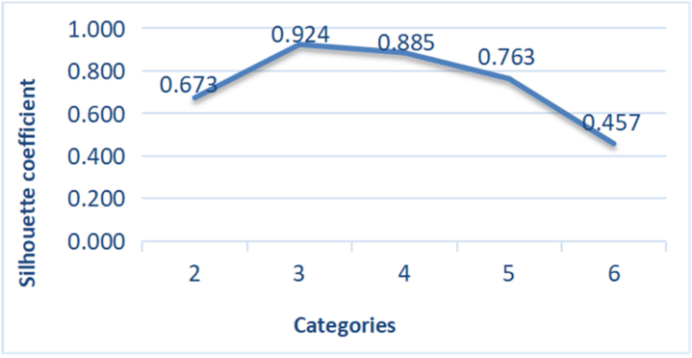

Figure 1. Silhouette coefficients for clustering number of classes 2-6

As shown in figure 1, the value of the corresponding silhouette coefficient was the largest (0.924) when the number of clustering categories was 3.Therefore, the risk levels of carbofuran in vegetables were classified into three levels: low, medium and high. The corresponding index values of the clustering centers of the three levels are shown in Table 1.

Table 1. Clustering centers of the three risk levels

Category	HQ	HI	NIPI	Risk Level
1	0.042	0.051	0.706	Low
2	0.008	0.019	9.419	Medium
3	0.085	0.150	19.986	High

The results of the risk level of carbofuran in vegetables obtained by the K-means++ algorithm are illustrated in Figure 2 and Table 2, which are composed of different provinces and different kinds of vegetables. The combinations of medium and high risk levels were ranked from high to low, as shown in table 3.

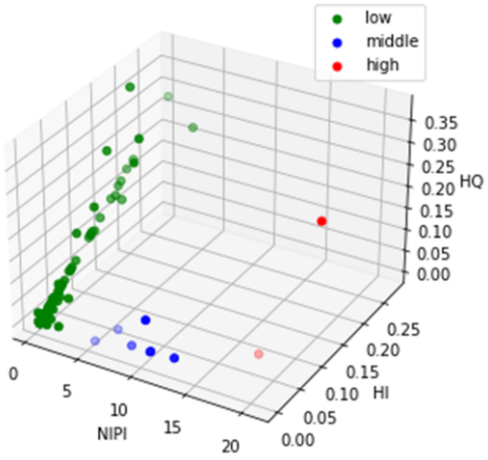

Figure 2. K-means++ clustering results

It can be seen from the table 2 and table 3 that the combinations of province-vegetable categories with risk level 1 accounted for 92.5% of the total, and the high-risk combinations were bulb vegetables in Hebei and leafy vegetables in Shaanxi. The vegetable categories with relatively high risk are legumes and bulb vegetables.

Compared with traditional dietary exposure assessment methods, this study used objective data to build a risk ranking model and adopted a data-driven approach to confirm the number of risk classes through a clustering algorithm, increasing the objectivity, scientific validity and validity of the experimental results. This model ranked the risk combinations of 120 provincial and municipal-vegetable varieties across China, and the results were not significantly different when compared with the results of traditional risk assessment. The Fifth China Population Diet Study showed that the risk of carbofuran was mainly from vegetables, and the risk assessment of carbofuran considering food consumption factors of the Chinese population showed that the risk of ketamine was higher in Hebei, Fujian and Guangdong, again verifying the validity of the experimental results of this study.

However, it is important to note in particular that the grading is based on the intercomparison of the province and vegetable category portfolios, not on absolute risk.Therefore, the combinations of medium and high risks mainly indicate that they should receive priority attention from the population and regulatory authorities.

Table 2. Results of vegetable risk grading by province

Province	Legumes	Root and potato vegetables	Melon vegetables	Bulb vegetables	Eggplant and fruit vegetables	Leafy Vegetables
Beijing	1	1	1	1	1	1
Fujian	2	1	1	2	1	1
Guangdong	2	1	1	1	1	1
Guangxi	2	1	1	1	1	1
Hebei	1	1	1	3	1	1
Henan	1	1	1	2	1	1

Heilongjiang	1	1	1	1	1	1
Hubei	1	1	1	1	1	1
Hunan	1	1	1	1	1	1
Jilin	1	1	1	1	1	1
Jiangsu	1	1	1	1	1	1
Jiangxi	1	1	1	1	1	1
Liaoning	2	1	1	1	1	1
Inner Mongolia	1	1	1	1	1	1
Ningxia	1	1	1	1	1	1
Qinghai	1	1	1	1	1	1
Shaanxi	1	1	1	1	1	3
Shanghai	1	1	1	1	1	1
Sichuan	1	1	1	1	1	1
Zhejiang	2	1	1	1	1	1

Table 3. Combination of province-vegetable categories with medium and high risk levels

Province	Vegetable Category	Risk Level
Hebei	Bulb vegetables	High
Shaanxi	Leafy Vegetables	High
Guangdong	Legumes	Medium
Guangxi	Legumes	Medium
Henan	Bulb vegetables	Medium
Zhejiang	Legumes	Medium
Fujian	Legumes	Medium
Fujian	Bulb vegetables	Medium
Liaoning	Legumes	Medium

5. Conclusion

In this study, we established a risk classification model for carbofuran based on the exposure assessment method, and achieved the risk classification of 20 provinces and different vegetable categories across China by K-means++ clustering algorithm. Compared with the traditional dietary exposure assessment method, the risk grading model enables us to consider the combined effects of multiple indicators in a comprehensive and objective manner. The study conducted a comparative study through the silhouette coefficients and finally selected the K-means++ algorithm in order to determine the number of risk levels and achieve the risk classification of the combination of provinces and vegetable categories. The results show that the model in this study can be effective for the risk classification of carbofuran, but more validation tests are needed to see if this model is applicable to other hazard categories.

Acknowledgments

This research was supported by the National Key Technology R&D Program of China (No.2019YFC1606401); the Beijing Natural Science Foundation (No.4202014); the Natural Science Foundation of China (No.61873027); the Humanity and Social Science Youth Foundation of Ministry of Education of China (No.20YJCZH229); and the Social Science Research Common Program of Beijing Municipal Commission of Education (No.SM202010011013).

References

[1] 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.

[2] Zhang Y, Zhou Y, School S. Review of clustering algorithms[J]. Journal of Computer Applications, 2019.

[3] Arthur D, Vassilvitskii S. k-means++: the advantages of careful seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, SIAM, pp. 1027-1035. 2007.

[4] Thompson H M. The use of the Hazard Quotient approach to assess the potential risk to honeybees (Apis mellifera) posed by pesticide residues detected in bee-relevant matrices is not appropriate[J]. Pest Management Science, 2021.

[5] Lei M, Zeng M, Wang L H, et al. Arsenic, lead, and cadmium pollution in rice from Hunan markets and contaminated areas and their health risk assessment[J]. Huanjing Kexue Xuebao / Acta Scientiae Circumstantiae,2010, 30(11):2314-2320.

[6] Wang L H, Li M M, Zhang Y, et al. Pollution characteristics and health risk assessment of heavy metals in soil of a vegetable base in North China[J]. Acta Geoscientica Sinica, 2014, 35(2):191-196.

[7] Peter R J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis[J]. Journal of Computational & Applied Mathematics, 1987, 20.