Transactions on Environment, Energy and Earth Sciences
ISSN: 2960-1924, eISSN: 2960-2300 | Volume 3
GRET 2024

Research on Enterprise Emission Risk Assessment Model Based on XGBoost-AHP

Tong Wu*

School of Information Technology & Management, University of International Business and Economics, Beijing, China, 100029

* Corresponding Author Email: wtong2990@gmail.com

Abstract. This study aims to address the immature development of environmental pollution-related insurance in China by introducing machine learning technology to improve the traditional risk assessment model in order to scientifically and accurately quantify the pollution caused by enterprises in the production process. In this paper, public data such as annual reports and social responsibility reports of listed companies are collected, and the data on corporate emissions are identified by KMeans and GMM clustering, and then the probability of corporate illegal emissions behavior is predicted by using XGBoost classifier. At the same time, considering the environmental pollution risk of the location of the enterprise, Random Forest was used to predict the comprehensive risk and transfer probability. Finally, the enterprise discharge risk is evaluated by hierarchical analysis method. This study not only helps the enterprise's own risk management, but also provides a pricing basis for the insurance company and promotes the development of environmental pollution-related insurance in China, which has important theoretical and practical significance.

Keywords: Environmental Pollution Insurance; XGBoost; Corporate Emissions; Risk Management.

1. Introduction

With the rapid advance of industrialization and urbanization, the environmental risks posed by corporate wastewater discharges are becoming more and more prominent, which poses a serious challenge to the protection of the environment, the preservation of the ecological balance and the safeguarding of public health. In order to accurately assess these risks, this paper urgently needs to rely on advanced technological tools. In recent years, the rapid development of machine learning technology provides a powerful tool to solve this problem. In the field of wastewater discharge risk assessment, several studies have accumulated rich experiences and insights for this paper. Govender and Sivakumar's study successfully classified air pollution data by using KMeans clustering method, which not only revealed the hidden patterns in the data, but also provided a brand-new perspective on wastewater data processing [1]. This means that KMeans clustering can also be applied to the analysis of wastewater emission data, so as to identify the sources and types of pollution more accurately. Meanwhile, the study of Alahamade et al. demonstrated the excellent performance of random forest model in air quality prediction, and the accurate prediction model they constructed successfully predicted the trend of air quality [2]. This result provides solid theoretical support and practical guidance for this paper to predict the environmental impacts of wastewater discharges using the random forest model. In addition, Xuchun Zhang et al. used the XGBoost algorithm to effectively predict the sewage treatment alarm, which fully proved the superiority of the algorithm in dealing with complex environmental data [3]. It is also worth mentioning that Junxia Wang et al. also successfully identified abnormal emissions during wastewater treatment through clustering algorithms, a finding that is of great guiding significance for this paper to improve the monitoring of wastewater quality and treatment efficiency [4]. Based on the insights from the above studies, this paper will integrate advanced machine learning techniques such as KMeans clustering, Random Forest, and XGBoost [5], and take listed companies in Jiangsu Province as an example to explore in depth the risk assessment methods of corporate wastewater discharges. By integrating these techniques, this paper expects to construct a more accurate risk assessment model, so as to provide more scientific support for environmental protection and sustainable development.

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

2. Integrated learning-based classification and prediction of corporate sewage behavior

2.1. Data sources

In this study, the data on the amount of emissions from listed companies were obtained from the dataset of the paper 'Financing Platform Debt and Environmental Pollution Governance' (Mao Jie, Guo Yuqing, Jing Cao et al.); the information on population, greening, and water quality of each prefecture-level city in Jiangsu Province was obtained from the publicly disclosed statistical data of the Jiangsu Provincial Bureau of Statistics, the Jiangsu Provincial Department of Water Conservancy, and the Jiangsu Provincial Bureau of Forestry; the data related to the penalties for company emissions and the cities in Jiangsu Province with respect to air, water quality, rubbish, bicarbonate. The smart environmental protection scores are taken from IPE Public Environmental Research Centre, and are obtained by searching and downloading the data through automatic computer simulation of manual operation. The data used in this paper are pre-processed through normalization and standardization, and the processed data are shown in Table 1:

Table 1. Display of some data processing results

Stock Code	Year	Chemical Oxygen Demand	Ammoniacal Nitrogen Emissions	Total Nitrogen	Total Phosphorus	Logarithm of Integrated Pollution Equivalent of Water Body
35	2007	-0.27	-2.14	-2.09	-1.41	-2.39
35	2008	-0.73	-2.04	-2.12	-1.86	-2.69
...	...	...	...	...	...	...
300936	2022	0.12	2.15	0.51	0.35	0.97
301003	2022	1.17	2.16	0.91	2.22	1.42

2.2. Data categorization

The data in the datasets used in this paper are derived from the annual reports of listed companies, social responsibility reports or public information on the company's official website, and their authenticity has to be verified. In order to gain a deeper understanding of the patterns and characteristics of corporate discharging behaviors and accurately assess the discharging risks faced by enterprises, this paper expects to predict the probability of corporate water pollution violations in order to correct the model accordingly. To this end, this paper first adopts two clustering methods, KMeans and GMM, to identify and categorize the discharge data of listed companies.

According to the characteristics of the original data, normal data and abnormal data can be distinguished by clustering first. The initial clustering is firstly carried out using KMeans algorithm, and then the secondary clustering is carried out using GMM algorithm. Since the normal data accounts for the largest proportion, the data with the largest proportion after clustering is classified as normal type, and the rest of the data is classified as abnormal class [6-8].

2.2.1. First KMeans clustering

In this paper, we use the KMeans algorithm, set the number of clusters to 2, divide the data into two classes, and add the clustered labels as a new column to the original data in order to influence the secondary clustering as new features.

The evaluation index of the first clustering is shown in Table 2.

Table 2. KMeans clustering results

Evaluation indicators	metric	Performance and Quality
Contour Coefficient	0.43	Clustering works well
Calinski-Harabasz Index	2180.48	Clustering works well
Davies-Bouldin Index	0.89	Clustering works well

2.2.2. Second GMM clustering

The model established here also sets the number of clusters as 2, and the evaluation indexes are shown in Table 3, which shows that the profile coefficient and Calinski-Harabasz index of the model after the second clustering increase, and the Davies-Bouldin index decreases, so the fitting effect after the second clustering is further improved than that of KMeans clustering only.

Table 3. GMM clustering results

Evaluation indicators	metric	Performance and Quality
Contour Coefficient	0.46	Clustering works well
Calinski-Harabasz Index	2374.59	Clustering works well
Davies-Bouldin Index	0.85	Clustering works well

2.2.3. Data clustering results

The clustering results were visualized by plotting the scatter plot of the final clustered data to visualize the distribution between different categories and the final cluster distribution. Figure 1-4 shows the distribution of four groups of characteristics, namely, COD and ammonia emissions, log of ammonia emissions and water body integrated pollution equivalent, total nitrogen and total phosphorus, and log of COD and water body integrated pollution equivalent, using scatter plots, and coloring the data points according to the final clustering results, with each color also representing a cluster of clusters.

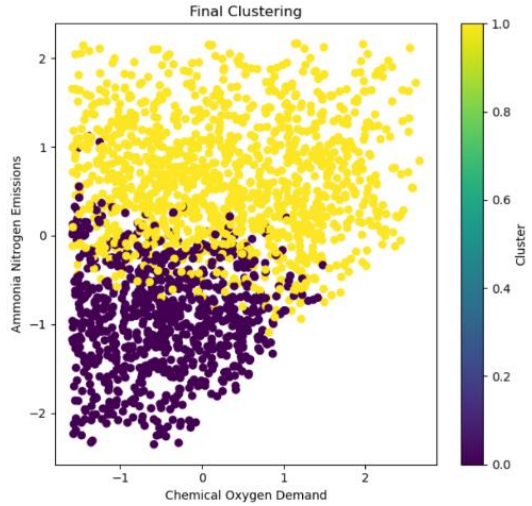

Figure 1. Clustering of COD and ammonia emissions

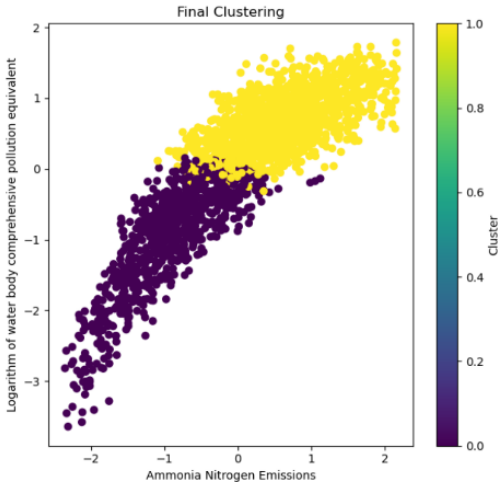

Figure 2. Logarithmic clustering of ammoniacal nitrogen emissions and combined pollution equivalent of water bodies

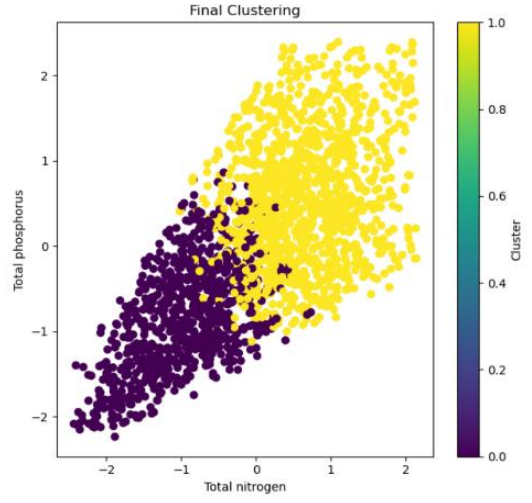

Figure 3. Total nitrogen and total phosphorus clustering

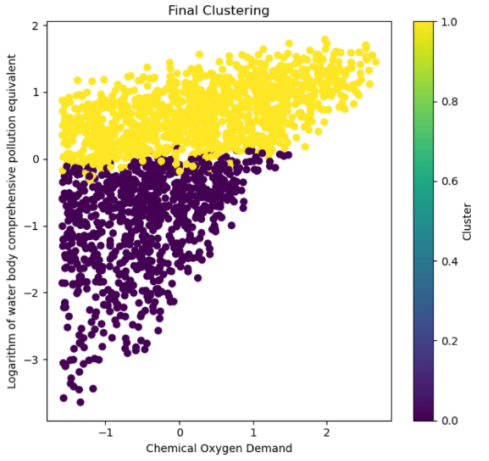

Figure 4. Logarithmic clustering of chemical oxygen demand and combined water pollution equivalents

According to the Measures for Determination and Handling of Environmental Monitoring Data Falsification and Falsification Acts issued by the Ministry of Ecology and Environment as well as the Regulations on the Management of Emission Permits, the abnormalities of enterprise emission data are mostly caused by illegal behaviors of enterprises such as tampering and falsification of self-monitoring data, as well as smuggling and discharge, etc., and the probability of the appearance of abnormal data can be interpreted as the probability of enterprise illegal behaviors. Among them, the proportion of normal data is more than one, and the probability of not violating the law is large, and 0 and 1 after clustering can be interpreted as the emergence of enterprises with and without violations of the law, respectively.

2.3. Prediction of probability of illegal discharge by enterprises based on XGBoost model

Based on the above clustering, this paper uses the clustered data to build and train XGBoost classifiers to identify and predict the likelihood of whether a company may have violated the law by stealing emissions, tampering with data or falsifying data.

2.3.1. Model construction and results

In this paper, XGBoost model is used to fit the data to create a classifier, and the data that has been clustered is used as a dataset, which is divided into a training set and a test set in a ratio of 7:3. The training set data was called to train the model and predictions were made on the test set, and the predicted results of the probability of each enterprise belonging to each class were obtained. Calculating the average probability of the overall two categories, the predicted probability of the enterprise's illegal discharge is about 0.37, which is in line with the expectation. The prediction results for some enterprises are shown in Figure 5.

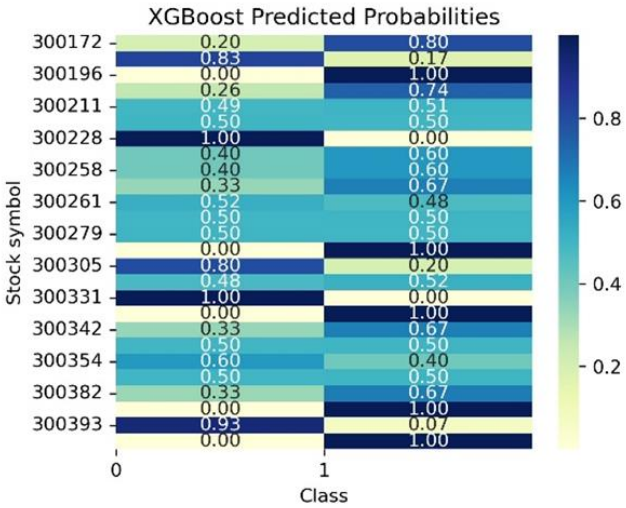

Figure 5. XGBoost partial prediction results

2.3.2. Performance comparison

In this paper, we compare the performance of XGBoost model with GBDT model and Random Forest model on the test set, the indexes include Mean Absolute Error (MAE), Root Mean Square Error (RMSE), in order to evaluate their performance in predicting the normal probability of corporate sewage data, and the results are shown in Table 4, which shows that XGBoost performs the best among the three models.

Table 4. Comparison of the performance of XGBoost and related models

Model name	MAE	RMSE
XGBoost model	0.010	0.099
GBDT model	0.012	0.107
Random forest model	0.013	0.115

2.4. Integrated risk assessment of sewage discharge

The risk of wastewater discharge by enterprises is not only affected by their own behavior, but also by the environment in which they are located. For this reason, this paper collects population, greening, air, water quality, bicarbonate, sewage and rubbish treatment, intelligent environmental protection and other environment-related data from the publicly disclosed statistics of Jiangsu Provincial Bureau of Statistics, Jiangsu Provincial Department of Water Resources, and Jiangsu Provincial Bureau of Forestry in 13 prefectural-level cities in Jiangsu Province, and selects the data related to water pollution discharge to be combined with the existing data on enterprise pollution discharge and penalty records to establish a random forest model to comprehensively assess the emissions risk faced by enterprises.

Due to the many factors considered in this paper, in order to prevent overfitting, correlation analysis is used to eliminate indicators with high correlation. The correlation coefficients between the variables are shown in Figure 6.

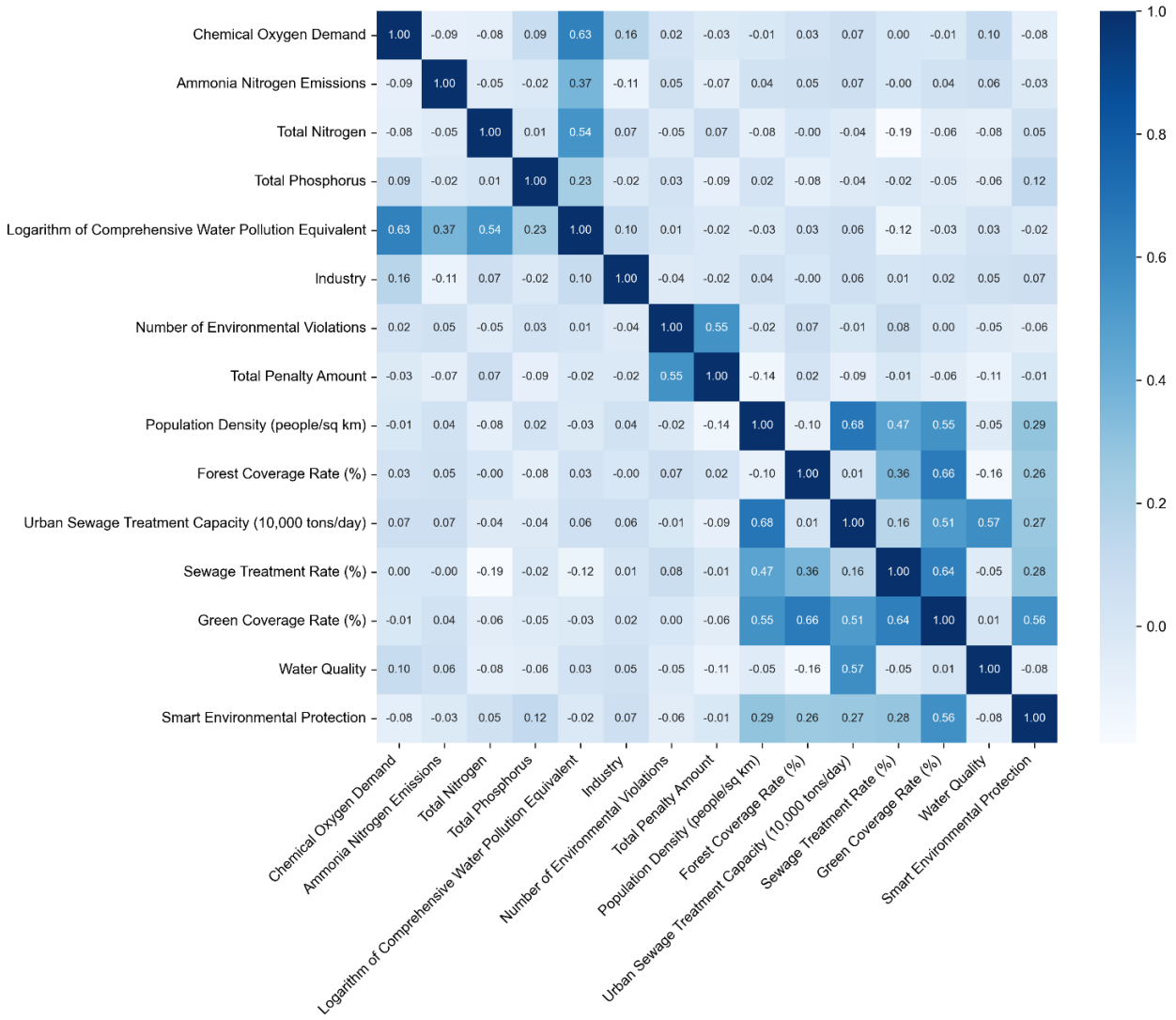

Figure 6. Correlation coefficients of indicators related to enterprise discharge variables

Considering that indicators with too high correlation will affect the clustering effect, this paper calculates the correlation coefficients between the indicators, and retains the most representative data in six categories: the logarithm of the integrated pollution equivalent of the water body, the industry to which it belongs, the number of environmental supervision records, the cumulative amount of penalties, the quality of the water, and the percentage of green space coverage. Nine data were missing for the water quality score alone, so the mean value was used to fill in the missing values, and the industry column was coded to standardize the data in each column.

2.4.1. Random forest parameterization and model evaluation

In this paper, the training set test set is divided according to 7:3, and the initial parameters are used to model and tune the parameters, and the best parameters and model evaluation indexes are obtained as shown in Table 5 and Table 6.

Table 5. Optimal parameter selection for Random Forest

Parameter name	Optimal parameter
Maximum depth	5
Maximum number of features	4
Number of decision trees	100
Randomness seed	42

Table 6. Random Forest Evaluation Indicators

Evaluation indicators	Scale
Silhouette Coefficient	0.86
Accuracy_BS	0.97
Macro-average AUC	0.99

The profile coefficients, accuracy and AUC macro averages of the best parametric model on the test set show that the model performs better. Adding the indicators one by one in order of importance, the accuracy of the model was higher than 0.97. In particular, the accuracy dropped from 0.985 to 0.972 when the 7th indicator was included, but considering the comprehensive nature of risk quantification, it was chosen to keep all indicators and not to delete them [9]. The final classification of companies into five categories and the corresponding eigenvalues are shown in Table 7.

Table 7. Corresponding centroids for each type of indicator

Type number	Logarithm of combined water body pollution equivalent	Affiliated Industries	Number of environmental regulatory records	Cumulative penalties	Water quality	Green area coverage
1	0.14	1.22	1.27	1.49	66.16	0.08
2	0.14	1.00	20.00	1785.00	61.50	0.02
3	0.14	1.00	7.33	135.10	66.93	0.10
4	0.14	1.00	11.50	291.78	61.70	0.13
5	0.14	1.33	4.67	58.17	67.06	0.09

The probability of each firm falling into each category is shown in Table 8.

Table 8. Probability that a business falls into one of the five categories

	Class 1	Class 2	Class 3	Class 4	Class 5
0	0.61286	1.95246e-223	0.03232	6.78837e-07	0.35482
1	0.61230	1.97418e-223	0.03242	6.82646e-07	0.35529
...	...	...	...	...	...
221	0.61297	1.95736e-223	0.03233	6.80480e-07	0.35470
222	0.61297	1.95736e-223	0.03233	6.80480e-07	0.35470

3. Results

This paper assigns the following judgement matrices to the six indicators, namely the logarithm of integrated pollution equivalent of the water body, the industry to which it belongs, the number of environmental regulatory records, the cumulative amount of penalties, the quality of the water, and the percentage of green space coverage, based on the opinions of experts in the relevant fields.

$$x = \begin{vmatrix} 1 & 3 & 5 & 7 & 9 & 2 \\ \frac{1}{3} & 1 & 3 & 5 & 7 & \frac{1}{2} \\ \frac{1}{5} & \frac{1}{3} & 1 & 3 & 5 & \frac{1}{3} \\ \frac{1}{7} & \frac{1}{5} & \frac{1}{3} & 1 & 3 & \frac{1}{4} \\ \frac{1}{9} & \frac{1}{7} & \frac{1}{5} & \frac{1}{3} & 1 & \frac{1}{5} \\ \frac{1}{2} & 2 & 3 & 4 & 5 & 1 \end{vmatrix} \quad (1)$$

Afterwards, the weights corresponding to each indicator were calculated according to the TOPSIS method [10]. The corresponding weights for the probability of belonging to each category are then calculated by combining the coordinates corresponding to the center point of each category in Table 7 as follows.

$$[0.0587 \;\; 0.145 \;\; 0.213 \;\; 0.371 \;\; 0.212] \quad (2)$$

Combining the above conclusions, this paper scores the comprehensive risk faced by each corporate discharge, and the formula for calculating the specific score is as follows:

$$score_i = \sum_{j=1}^{5} (p_{ij} \times n_{ij}) \quad (3)$$

Where, p_{ij} and n_{ij} denote the probability and weight of the ith firm belonging to the jth type, respectively, $j = 1,2,3,4,5$.

Each firm specific score will be uploaded with the data and visualized using bar charts as in Figure 7.

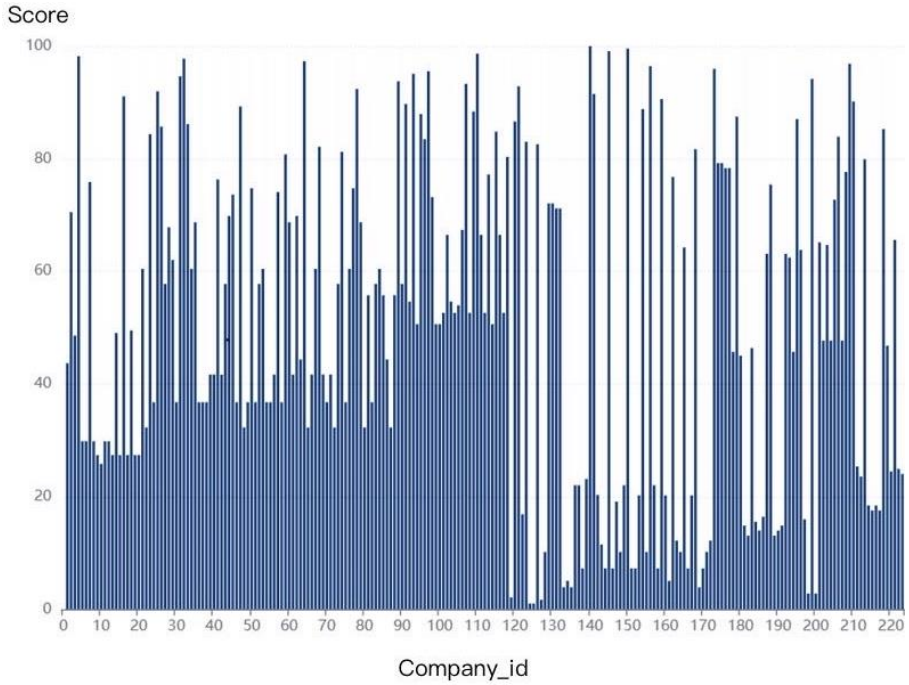

Figure 7. Combined enterprise discharge risk score

The overall description of the enterprise risk score is shown in Table 7, and its maximum, mean, median, standard deviation, kurtosis, skewness and coefficient of variation are as expected. Therefore it is reasonable and realistic to use this score to assess the comprehensive risk of corporate emissions.

Table 9. Overall description of risk scores

Maximum	Minimum	Average	Standard deviation
100	1	50.388	28.83

(continued)

Median	Kurtosis	Skewness	Coefficient of variation
50.612	-1.2	0.001	0.572

Ultimately, this paper can be interpreted from the enterprise emissions penalties and the environmental conditions of the five categories of enterprises, the interpretation of the results are shown in Table 10, while the final score of the enterprise is a combination of the relevant indicators to produce an objective score.

Table 10. Classification of enterprises and related explanations

Type number	Emission penalties	Environmental conditions
1	Very little	Good
2	Critical	Vulnerable and in urgent need of improvement
3	Rather serious	Good
4	Rather serious	Vulnerable
5	relatively small	Good

4. Conclusions

This study innovatively combines the insurance perspective with environmental pollution assessment, and proposes a new approach to corporate emissions assessment that takes into account both economy and environmental protection. The paper focuses on data authenticity, develops innovative assessment tools, and enhances the accuracy and flexibility of problem solving through multi-model fusion. The study integrates multiple data sources to construct a new type of corporate environmental risk profile, and also considers geographical and external environmental factors to provide strong support for regulation. Meanwhile, practical policy recommendations and regulatory tools provide important guidance for governments and insurance organizations.

Looking ahead, future research will continue to explore deep learning networks to improve the accuracy of model predictions. It is expected that the research will be extended to a wider area to enhance the generalizability of the model. In addition, future research plans to further improve the risk assessment indicator system through in-depth data analysis and expert consultation, so as to enhance the comprehensive assessment capability of the model.

References

[1] Govender P, Sivakumar V. Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019)[J]. Atmospheric pollution research, 2020, 11(1): 40-56.

[2] Alahamade W, Lake I, Reeves C E, et al. A multi-variate time series clustering approach based on intermediate fusion: A case study in air pollution data imputation[J]. Neurocomputing, 2022, 490: 229-245.

[3] Zhang X. Chun. Monitoring and early warning of wastewater treatment plant discharge based on CatBoost model[D]. Lanzhou: Lanzhou University, 2021.

[4] Wang Junxia, Liu Tonghao, Zhang Shoubin, et al. Research on supervision and inspection technology of self-monitoring of emission units[J]. China Environmental Monitoring,2019,35(2):23-28.

[5] Shrestha S M, Shakya A. A customer churn prediction model using XGBoost for the telecommunication industry in Nepal[J]. Procedia Computer Science, 2022, 215: 652-661.

[6] Wei Yan, Lai Jingxian, Zhou Qilong, et al. Exploration of identification rules and processing methods of abnormal data in automatic monitoring of pollution sources[J]. Environmental Monitoring Management and Technology,2022,34(2):56- 59.

[7] CHEN Chong, HE Wei, ZHONG Tianfu, et al. Identification of emission data anomalies in catalytic cracking unit based on isolated forest method[J]. Journal of Xi'an Petroleum University (Natural Science Edition), 2021,36(4):119-126.

[8] Wang Weijiu, Xu Minya, Xu Boshi et al. Construction and application of sentencing prediction model for illegal business offence based on XGBoost algorithm[J]. Intelligence Exploration,2022(9):20-28.

[9] Huang Lei, Huang Yujia, Liu Penghui, et al. Research on the methodological system of regional integrated environmental risk assessment[J]. China Environmental Science,2020,40(12):5468-5474.

[10] JIANG Dejuan, YU Haozhe, LI Lijuan. Dynamic evaluation of water resources carrying capacity in Shandong Province based on comprehensive empowerment and TOPSIS model[J]. Resource Science,2024,46(03):538-548.