Establishment of Risk Prediction Model for Soil and Groundwater Pollution of Gas Station with Machine Learning Techniques

I-Cheng Chang, Shen-De Chen, and Tai-Yi Yu*

***Abstract*—With the rapid development of network technology and the digital economy, the wave of the era of artificial intelligence has swept the world. Facing the era of big data and artificial intelligence, data-oriented technologies are undoubtedly served as the practical research trend. Therefore, the precise analysis provided by big data and artificial intelligence can provide effective and accurate knowledge and decision-making references for all sectors. In order to effectively and appropriately evaluate the potential risk to soil and groundwater for gas station industry, this study focuses on the potential risk factors affecting soil and groundwater pollution. In the past, our team has evaluated the risk factors affecting the remediation cost of soil and groundwater pollution for possible potential pollution sources such as gas stations, this study proceeds with the existing industrial database for in-depth discussion, uses machine learning technology to evaluate the key factors of pollution risk for soil and groundwater, and compares the differences, applicability and relative importance of the three machine learning techniques (such as neural networks, random forests and support vector machine). The performance indicators reveal that the random forest algorithm is better than support vector machine and artificial neural network. The relative importance of parameters of different machine learning models is not consistent, and the first five dominant parameters are location, number of gas monitoring wells, age of gas station, numbers of gasoline oil nozzle, and number of fuel dispenser for random forest model.**

***Index Terms*—Neural network, support vector machine, random forest, gas station, soil and groundwater pollution**

I. Introduction

The environmental medium of soil and groundwater, carrying all pollutants in air, water and soil, have complex characteristics such as long-term accumulation, dynamic change, difficult to find and cross medium pollution. According to the pollution characteristics of soil and groundwater, pollution events may lead to the risk of physical injury and financial loss of third parties, the risk of physical injury, financial loss and liability of enterprises, and major environmental pollution events will cause huge financial risks and even threaten the credit ranking of the government. Environmental risk often belongs to the fat tail risk with low loss frequency and huge loss. In order to effectively and appropriately evaluate the potential risks of the operators to soil and groundwater, this study cooperates with the existing industrial database of Taiwan gas stations and uses machine learning technology to predict and evaluate the crucial factors affecting the pollution risks of soil and groundwater. Risk identification and risk analysis are the dominant works to evaluate the risks and hazards caused by soil and groundwater from a specific industry for a prework of a risk management plan. There is uncertainty and variability in the process of risk assessment. Uncertainty arises from a lack of understanding of the risk assess model, and scientific methods can be applied to reduce uncertainty with more or better data, but it cannot be completely eliminated. Variability comes from the differences in the exposure behavior and degree of the pollution source to the receptor, and its influencing factors may include the characteristics of the pollution source itself, geographical location, soil, geology, groundwater, pipeline characteristics, and coating materials. This study performs the following procedures (1) collect the related database of the gas station industry; (2) collect related risk factors of soil and groundwater pollution at gas station site; (3) apply machine learning technologies to predict potential risks of gas stations that would cause pollution of soil and groundwater; (4) extract the dominant factors of the basic information and monitoring results for gas stations, and (5) evaluate the performance of different machine learning approaches.

The selected AI technology is the key issue to develop an accurate and effective risk model of soil and groundwater, and the model must have appropriate statistical variables, measurable variables and relevant statistical parameters. In recent years, machine learning technologies have been used to establish appropriate pollution models, such as neural network [1], support vector machine (SVM) [2], and random forest algorithm [3].

In recent years, the application of neural networks, Ehteshami *et al.* [4] applied neural networks to predict the nitrate pollution of groundwater, and the input parameters considered hydrogeology, soil nitrogen content, soil organic matter and soil carbon content, etc. The differences between the backpropagation (BP) and radial basis functions were compared, and found that the model differences of the two functions were not significant, and they could well predict the nitrate pollution in groundwater. However, the radial basis model showed marginally better performance compared to back-propagation by 30 %. Bieganowski *et al.* [5] studied polluted soils of different geology to understand the decline in oil concentration and changes in soil moisture, analyzed commercial gasoline and diesel to distinguish various

Manuscript received June 4, 2022; revised July 1, 2022; accepted August 3, 2022.

I-Cheng Chang is with the Department of Environmental Engineering, National Ilan University, Yilan, Taiwan. Email: icchang@niu.edu.tw (I.C.C.)

Shen-De Chen is with the Apollo Technology Co., Ltd, Taipei, Taiwan. Email: shende@apollotech.com.tw (S.D.C.)

Tai-Yi Yu is with the Department of Risk Management, Ming Chuan University, Taipei, Taiwan.

*Correspondence: yti@mail.mcu.edu.tw (T.Y.Y.)

hydrocarbons, and used principal component analysis (PCA) and artificial neural network (ANN) to interpret soil volatility fingerprints recorded by electronic noses. Hou *et al*. [6] compared BP neural network optimized by genetic algorithm to predict cadmium concentration in rice grains. Based on Pearson correlation analysis and geodesy, variables such as total soil cadmium concentration, clay content, nickel concentration, cation exchange capacity, organic matter and pH were selected as input factors for the prediction model. Based on the cadmium concentration in food predicted by the model, human exposure and health risks can be quickly assessed, and timely measures can be taken to reduce the transfer of cadmium from the soil to the food chain and reduce the exposure risk of organisms. Liu *et al*. [7] combined BP neural network with particle swarm optimization, which led to an integrated PSO-BPNN method, to estimate three heavy metals (cadmium, mercury and arsenic) in soil content.

Based on the minimization principle of structural risk, support vector machine (SVM) can avoid the over fitting problem [8], and has the ability of minimization on structural risk, and can avoid the dilemma of ANN models falling into local minimum [9]. Sakizadeh *et al*. [10] collected 229 soil samples, analyzed 12 kinds of heavy metals (Ag, Co, Pb, Tl, Be, Ni, Cd, Ba, Cu, V, Zn and Cr) to predict soil pollution index (SPI) with SVM and ANN algorithms. Jia *et al*. [11] used three different machine learning methods, including SVM, naive Bayesian (NB) and ANN, to predict and classify potential polluting enterprises in China's Yangtze River Delta, and classified geographical regions and industries based on the geographical statistical data of more than 260,000 enterprises.

Compared to SVM and ANN models, random forest (RF) technique may be a user-friendly technique [12], RF manner may provide better results and avoid overfitting [13]; there are only two model parameters (the number of variables in a random subset of each node, and the number of trees in the forest) to be decided. Rodriguez-Galiano *et al*. [3] explored the performance of RF in the prediction and simulation of nitrate pollution, taking agricultural areas as the verification object, and setting the trigger value of 50 mg/L of nitrate concentration in groundwater, based on comprehensive GIS database, including hydrogeological attributes, driving forces, remote sensing variables and physicochemical variables, a total of 24 parameters, which are used as input parameters to predict a nitrate pollution; the RF assessment results are also consistent with the logistic regression method. The prediction results show that RF could supply certain degree of accuracy and rank relative importance of different variables, and RF approach is suitable to predict and interpret complex sources and high dimensional data, and RF manner is easy to select the dominant variables that affect prediction model [14–17].

II. Materials and Methods

Based on the objectives and requirements of this study, the temporospatial scope and data sources is currently defined on the basis of the complete attributes of gas stations and the records of improvement sites, which were collected from Taiwan Environmental Protection Administration in the past 15 years. The information of gas stations has accumulated more than 3,200 records of site information with more than 40 fields, and 150 records of improvement records. This research proceeds data analysis and comparison of artificial intelligence and predicts risk potential of soil and groundwater pollution for gas station. After data cleaning process (such as data missing/ white space/ blank data, non-correlated variable processing, etc.), this research performs the following four data cleaning operations.

1) Measurable variables: Based on the existing data of gas stations in Taiwan, after brainstorming within the three experts, this study captures the relevant data fields of subsequent data into the modeling operation. The data fields include basic information of the gas station (including setting date, location, oil used, business type, business status, announcement status, announcement date, releasing announcement date), monitoring results (soil monitoring results, groundwater monitoring results, pollution potential in last year, monitoring method of storage tank, monitoring method of pipeline, number of monitoring wells for soil gas), pipeline protection material, pipeline type, storage tank type, protection material of storage tank, leakage prevention facilities, and leakage records.
2) Derived variables: In this study, variables such as the date of the previous disclosure, potential risk, storage tank protection, pipeline protection, overflow protection device, monitoring method of storage tank, pipeline monitoring method, and leakage protection facility were sequentially included. According to the content of the corresponding fields in the original data set, and the principles of this study, they are respectively created as gas station age, pollution potential levels, double wall of storage tank, protection measure of pipeline system, overflow protection device, monitoring measure of oil storage tank, monitoring measure of pipeline system, anti-leakage device for tankers. Several variables are converted into flag variables. The possible options for pipeline material were glass fiber, galvanized steel pipe, single-layer flexible hose, double-layer flexible hose, seamless steel pipe and others. Storage tank material included protective steel, single layer glass fiber and double layer glass fiber. The type of storage tank protection contained cathodic protection, coating, epoxy resin, glass fiber coating, asphalt coating, secondary barrier layer, polyethylene PE, PU, anti-corrosion belt and others.
3) Discrete coding: In view of the current domestic gas station data, including the station age, number of gasoline nozzle, number of fuel dispenser, number of storage tanks, number of groundwater monitoring wells, number of monitoring wells of soil gas, improvement cost for in-situ remediation and other numerical variables. In order to match the needs of machine learning analysis, this study introduces the K-means algorithm to discretize the conversion variables and encode the non-missing /non-blank values in the aforementioned variable range.
4) Data record screening/data re-cleaning: After cleaning the possible missing values (missing/white space/blank) in

the conversion variable field, establish and determine the role of risk factor variables. The 150 records of improvement records were multiplied 15 times and added to the database.

Machine learning is the branch of artificial intelligence that focuses on training computers to learn from data and improve based on experience, rather than running jobs according to explicit code. In machine learning, algorithms are trained to find patterns and correlations in large data sets, and based on that analysis, make the best decisions and predictions. Machine learning applications continue to improve with users, accessing more data and increasing accuracy.

Artificial neural networks (ANN): an artificial neural network (ANN) is composed of a node layer, including an input layer, one or more hidden layers, and an output layer. Each node or artificial neuron will be connected to another, and has a weighted sum threshold. If the output of any individual node is higher than the specified threshold, the node is immediately started and the data is transmitted to the next layer of the network. Otherwise, no data will be transmitted to the next layer. It captures the nonlinear behavior between dependent variables and independent variables, so it is widely used as prediction, classification and optimization methods in various fields. Most neural networks are feedforward, which means that they flow in only one direction, from input to output. However, the model can be trained through back propagation. Back propagation allows the model to calculate and attribute the errors associated with each neuron, so that model can appropriately adjust and fit the parameters of the model.

Support vector machine (SVM) is the earliest proposed new machine learning technique, the development of SVM manner is based on minimization of structured risk to minimize the upper limit of generalization error. The SVM can achieve good generalization results on both classification and regression, because the convergence principle gives it a greater ability to regress the relationship between input and output values, and obtain satisfactory performance on new input data. Least-squares support vector machines (LS-SVMs) have now emerged as an attractive semi-supervised statistical learning technique for rapidly solving multivariate calibration problems. Compared to traditional SVMs with direct quadratic programming, least squares linear systems can help SVMs solve regression and classification problems.

The RF can be regarded as integrated learning based on decision tree algorithm. Multiple decision trees are generated by bagging method, and combined with the prediction results of multiple decision trees, the category with the largest number of votes is selected from many decision trees by voting. In the regression model, the result of random forest output will be the average of many decision trees. Compared with decision tree algorithm, RF has stronger generalization ability, has ability to handle more input variables, and can evaluate the importance of each variable. For datasets with uneven classification, RF can reduce the error and is less prone to over fitting.

In this study, a confusion matrix was applied to evaluate the performance of a machine learning model. Three performance indicators were cited as (1) Accuracy: It defines how often the model predicts the correct output. It can be calculated as the ratio of the number of correct predictions made by the classifier to all number of predictions made by the classifiers. (2) Precision: It is defined as the number of correct outputs provided by the model or out of all positive classes that have predicted correctly by the model, how many of them were actually true. (3) Recall: It is denoted as the out of total positive classes, how our model predicted correctly.

III. Results and Discussion

This study adopts three machine learning methods, such as ANN, SVM and RF, to predict the risk potential of gas station site, after preprocessing the field data, all data were incorporated into as input data of machine learning approaches. To provide the same performance basis for comparison, the confusion matrix and the accuracy, precision, and recall rates were employed. Since the proportion of positive and negative data in this case is obviously unevenly distributed (the positive data in this study represents the improvement of the gas station site, usually other fields also pay attention to a few cases such as disease, defects, credit failure, etc.). The numbers of samples for classification algorithm and dominant parameters need to be adjusted accordingly, and the classifier would capture and identify the feature differences between positives and negatives. Facing with such challenges, the ROC indicator, adjustment of the proportion of data, adjustment of the classification rules would be common approaches. In this study, the proportion of data was adjusted for the comparison of three machine learning manners.

TABLE I: The Confusion Matrix of with Three Machine Learning Techniques

SVM	Real		Performance Indicators (%)	
Prediction	1	0	Accuracy	88.49
1	1,316	294	Precision	81.74
0	163	2,196	Recall	88.98
RF	Real			
Prediction	1	0	Accuracy	92.92
1	1,565	242	Precision	86.61
0	34	2,057	Recall	97.87
ANN	Real			
Prediction	1	0	Accuracy	64.55
1	224	1,386	Precision	13.91
0	21	2,338	Recall	91.43

The results of machine learning analysis (Table I) are summarized as follows: (1) The accuracies of the SVM, RF and ANN models are 88.49, 92.92 and 64.55%, respectively. The accuracies of SVM and RF models were greater than 85%. (2) The precisions of the SVM, RF and ANN models are 81.74, 86.61 and 13.91%, respectively. The precision of RF model was greater than 85%. (3) The recall rates of the SVM, RF and ANN models are 88.98, 97.87 and 91.43%, respectively. The recall rates of these three models were greater than 85%. (4) The analytical results of three performance indicators, demonstrated that RF model was the most suitable one. Only precision of RF model was suitable for a qualified model. (5) The machine learning model established in this study can indeed effectively predict the risk potential of alternative gas station. It is still necessary to collect relevant improvement site data to assist in the establishment of risk potential data for soil and groundwater

pollution as a basis for decision-making for supervisors and consultant industry.

Performance indicators are important factors to measure whether the model performance is effective or not. Therefore, the indicators selected in this study are important indicators, which must meet certain performance level. For the pollution cases of soil and groundwater from gas stations in this study, the importance of false negative is much higher than that of false positive. The false negative ratio represents the correct ratio of the polluted site to the non-polluted site, and the false positive represents the correct ratio of the non-polluted site to the polluted site. Therefore, the performance of ANN on the false negative ratio of improved samples is far lower than that of SVM and RF models. Compared to SVM, the false negative ratio of RF is relatively better, that is a need to expand certain number of improved sites to increase the performance level of RF model.

According to the above gas station data described in the method section, all relevant and available data are incorporated into database to proceed machine learning analysis, and analytical results of the revised data for different machine learning methods reveal following: (1) The relative importance of parameters in this study (Table II), the first five important factors are risk potential in last year, location, pipeline protection materials, numbers of gasoline oil nozzle, and number of fuel dispenser for the SVM model; location, the number of soil gas monitoring wells, gas station age, numbers of gasoline oil nozzle, and number of fuel dispenser for the RF model; and location, the number of soil gas monitoring wells, age of gas station, numbers of gasoline oil nozzle and number of fuel dispenser for the ANN model. (2) The relative importance of the parameters of different machine learning models are not consistent, and these results reflect the difference in methodology of different techniques. (3) The suitability and availability of machine learning approach depend on the type of model, the number of samples, and the ratio of positive and negative values. Maybe the machine learning model already has a certain degree of performance, there is a need to meet the performance standard to provide good explanatory power and high-quality decisions. (4) On the basis of performance indicators for three machine learning approaches, the RF is the best one of three models. The first five important factors of all parameters are location, the number of soil gas monitoring wells, gas station age, numbers of gasoline oil nozzle, and number of fuel dispenser for this study. By the way, location is the first, second, and first important factor for the SVM, RF, and ANN models. In the SVM model, risk potential in last year, pipeline protection type have the relative high importance; No. of monitoring wells for soil gas, gas station age, No. of gasoline oil nozzle for the RF model; gas station age, No. of monitoring wells for soil gas, and No. of gasoline oil nozzle for the ANN model. (5) Considering the numbers of important factors, there are three (risk potential in last year, location, pipeline protection type), five (location, No. of gasoline oil nozzle, No. of monitoring wells for soil gas, No. of fuel dispenser, and gas station age), and four factors (location, No. of monitoring wells for soil gas, No. of gasoline oil nozzle, and gas station age) for SVM, RF and ANN models, respectively.

TABLE II: THE RELATIVE IMPORTANCE OF TOP TEN VARIABLES FOR DISTINCT MODELS

Items	SVM	RF	ANN
Risk potential in last year	0.32	0.08	0.04
Location (county or city)	0.24	0.32	0.25
Pipeline protection type	0.14		
No. of gasoline oil nozzle	0.05	0.14	0.10
No. of monitoring wells for soil gas	0.04	0.18	0.14
No. of fuel dispenser	0.03	0.10	0.06
Pipeline protection material	0.03		0.05
leakage prevention device	0.03	0.02	0.01
No. of storage tank	0.03	0.07	0.02
storage tank material	0.03	0.03	
gas station age		0.16	0.24
No. of groundwater wells		0.02	0.01

IV. CONCLUSIONS

Based on the systematic thinking of the data-oriented approach, big data and artificial intelligence, this study applied three machine learning algorithms (ANN, SVM and RF manners) to construct a pollution risk model of soil and groundwater pollution for gas stations in Taiwan, and provides the relative importance analysis of model parameters. The results led to following conclusions: (1) Integration of diverse data of existing environmental protection authorities at all levels to form a complete and correct environmental database, which should provide important and detailed reference for decision-making. Technologies such as big data and artificial intelligence have become mature technologies, which can provide decision-making institutions with fine decision-making reference. (2) Environment is the receptor of pollution sources, the data of raw materials, imported and exported of raw materials, process data, emitted data from industrial sectors, waster data, environmental causality events may be included in the environmental database, and this integration will be great benefits to the environmental monitoring, environmental management of industrial sectors. (3) Based on the risk potential data of gas station data in Taiwan, this study establishes and compares performance indicators with three machine learning approaches, and finds that the random forest algorithm is better than support vector machine and artificial neural network. The performance of artificial neural network is the worst of the three algorithms in this study. (4) The relative importance of top ten parameters of different machine learning models is not consistent. From the perspective of random forest model, the first five dominant parameters are location, number of gas monitoring wells, age of gas station, numbers of gasoline oil nozzle, and number of fuel dispenser. (4) Utilization of machine learning methods to evaluate the pollution potential of gas stations to soil and groundwater pollution, this study integrates various machine learning methods to evaluate the importance of various risk factors, and found that location (county /city) and gas station age are the two most important risk factors. This study suggested that the future risk assessment of pollution for soil and groundwater could be combined with information such as industrial characteristics, land use, and facility age, as references for the delineation of potential areas of soil and groundwater pollution.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, S.-D. Chen, T.-Y Yu; methodology, T.-Y Yu and I.-C. Chang; software, T.-Y. Yu; validation, S.-D. Chen, I.-C. Chang; data curation, T.-Y Yu and I.-C. Chang; writing, S.-D. Chen, T.-Y Yu; review and editing, I.-C. Chang; supervision, S.-D. Chen; All authors have read and agreed to the published version of the manuscript.

Acknowledgment

The authors express their gratitude to the Ministry of Science and Technology, Taiwan (MOST 109-2511-H-130-003-MY2) for funding this study.

References

[1] S. M. Cabaneros, J. K. Calautit, and B. R. Hughes, "A review of artificial neural network models for ambient air pollution prediction," *Environmental Modelling & Software*, vol. 119, pp. 285–304, September 2019.

[2] Z. Huang, H. Chen, C. J. Hsu, W. H. Chen, and S. Wu, "Credit rating analysis with support vector machines and neural networks: A market comparative study," *Decision Support Systems*, vol. 37, no. 4, pp. 543–558, September 2004.

[3] V. Rodriguez-Galiano, M. P. Mendes, M. J. Garcia-Soldado, M. Chica-Olmo, and L. I. Ribeiro, "Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain)," *Science of the Total Environment*, vol. 476, pp. 189–206, April 2014.

[4] M. Ehteshami, N. D. Farahani, and S. Tavassoli, "Simulation of nitrate contamination in groundwater using artificial neural networks," *Modeling Earth Systems and Environment*, vol. 2, no. 1, pp. 1–10, February 2016.

[5] A. Bieganowski, G. Józefaciuk, L. Bandura, Ł. Guz, G. Łagód, and W. Franus, "Evaluation of hydrocarbon soil pollution using e-nose," *Sensors*, vol. 18, no. 8, p. 2463, July 2018.

[6] Y. X. Hou, H. F. Zhao, Z. Zhang, and K. N. Wu, "A novel method for predicting cadmium concentration in rice grain using genetic algorithm and back-propagation neural network based on soil properties," *Environmental Science and Pollution Research*, vol. 25, no. 35, pp. 35682–35692, October 2018.

[7] P. Liu, Z. Liu, Y. Hu, Z. Shi, Y. Pan, L. Wang, and G. Wang, "Integrating a hybrid back propagation neural network and particle swarm optimization for estimating soil heavy metal contents using hyperspectral data," *Sustainability*, vol. 11, no. 2, p. 419, January 2019.

[8] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, September 1995.

[9] L. Yu, S. Y. Wang, K. K. Lai, and L. Zhou, *Bio-Inspired Credit Risk Analysis: Computational Intelligence with Support Vector Machines*, Springer, 2008.

[10] M. Sakizadeh, R. Mirzaei, and H. Ghorbani, "Support vector machine and artificial neural network to model soil pollution: A case study in Semnan Province, Iran," *Neural Computing and Applications*, vol. 28, no. 11, pp. 3229–3238, November 2017.

[11] X. Jia, B. Hu, B. P. Marchant, L. Zhou, Z. Shi, and Y. Zhu, "A methodological framework for identifying potential sources of soil heavy metal pollution based on machine learning: A case study in the Yangtze Delta, China," *Environmental Pollution*, vol. 250, pp. 601–609, July 2019.

[12] T. Han, D. Jiang, Q. Zhao, L. Wang, and K. Yin, "Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery," *Transactions of the Institute of Measurement and Control*, vol. 40, no. 8, pp. 2681–2693, May 2018.

[13] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, October 2001.

[14] V. F. Rodriguez-Galiano, M. Chica-Olmo, F. Abarca-Hernandez, P. M. Atkinson, and C. Jeganathan, "Random Forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture," *Remote Sensing of Environment*, vol. 121, pp. 93–107, June 2012.

[15] V. F. Rodriguez-Galiano and M. Chica-Olmo, "Land cover change analysis of a Mediterranean area in Spain using different sources of data: Multi-seasonal Landsat images, land surface temperature, digital terrain models and texture," *Applied Geography*, vol. 35, no. 1–2, pp. 208–218, November 2012.

[16] X. Jia, D. O'Connor, Z. Shi, and D. Hou, "VIRS based detection in combination with machine learning for mapping soil pollution," *Environmental Pollution*, vol. 268, 115845, January 2021.

[17] K. Tan, H. Wang, L. Chen, Q. Du, P. Du, and C. Pan, "Estimation of the spatial distribution of heavy metal in agricultural soils using airborne hyperspectral imaging and random forest," *Journal of Hazardous Materials*, vol. 382, 120987, January 2020.