

Article

Big-Data Analysis and Machine Learning Based on Oil Pollution Remediation Cases from CERCLA Database

Hangyu Li ¹, Ze Zhou ¹, Tao Long ², Yao Wei ³, Jianchun Xu ¹, Shuyang Liu ¹  and Xiaopu Wang ^{1,*}¹ School of Petroleum Engineering, China University of Petroleum (East China), Qingdao 266000, China² State Environmental Protection Key Laboratory of Soil Environmental Management and Pollution Control, Nanjing Institute of Environmental Sciences, Ministry of Ecology and Environment, Nanjing 210042, China³ School of Computer Science & Engineering, South China University of Technology, Guangzhou 510006, China

* Correspondence: wxp@upc.edu.cn

Abstract: The U.S. Environmental Protection Agency's (EPA) Superfund—the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) database—has collected and built an open-source database based on nearly 2000 US soil remediation cases since 1980, providing detailed information and references for researchers worldwide to carry out remediation work. However, the cases were relatively independent to each other, so the whole database lacks systematicness and instructiveness to some extent. In this study, the basic features of all 144 soil remediation projects in four major oil-producing states (California, Texas, Oklahoma and Alaska) were extracted from the CERCLA database and the correlations among the pollutant species, pollutant site characteristics and selection of remediation methods were analyzed using traditional and machine learning techniques. The Decision Tree Classifier was selected as the machine learning model. The results showed that the growth of new contaminated sites has slowed down in recent years; physical remediation was the most commonly used method, and the probability of its application is more than 80%. The presence of benzene, toluene, ethylbenzene and xylene (BTEX) substances and the geographical location of the site were the two most influential factors in the choice of remediation method for a specific site; the maximum weights of these two features reaches 0.304 and 0.288.

Keywords: CERCLA; oil-contaminated soil; soil remediation; machine learning

Citation: Li, H.; Zhou, Z.; Long, T.; Wei, Y.; Xu, J.; Liu, S.; Wang, X. Big-Data Analysis and Machine Learning Based on Oil Pollution Remediation Cases from CERCLA Database. *Energies* **2022**, *15*, 5698. <https://doi.org/10.3390/en15155698>

Academic Editor: Gabriele Di Giacomo

Received: 30 June 2022

Accepted: 2 August 2022

Published: 5 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In the 1970s, the U.S. economy and work centres moved from city to suburb, north to south and east to west. After relocating, many enterprises left behind many 'brown land parcels', including industrial land, filling stations, abandoned warehouses and abandoned residential buildings that may have contained lead or asbestos. These sites were contaminated by industrial wastes to varying degrees, with high levels of hazardous substances in the soil and water, posing a severe threat to human health and the environment without legislative support and effective treatments.

Toxic waste dumps such as 'Love Canal' and 'Valley of the Drums' attracted attention from all over the U.S. when the public became aware of the risks of contaminated sites to human health and the environment. In response, the U.S. Congress established the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) in 1980, informally known as the Superfund. The Act provides broad federal authority to tax the chemical and petroleum industries and directly address releases or threats of releases of hazardous substances that may endanger public health or the environment. For example, the U.S. Environmental Protection Agency (EPA) can use the Superfund to prepay a site's treatment costs. The EPA can then recover all costs from one responsible party and from this responsible party to another. In other words, 'treatment first, accountability later' can effectively improve the efficiency of contaminated site treatment. Specific information on

the types of contaminants and remediation operations at each contaminated site in the U.S. is also provided, but each case study is independent and not systematic or instructive [1–4].

The EPA's database CERCLA has been selected as the data source relating to hazardous wastes being dumped, exposed, or mismanaged. There are thousands of contaminated sites in the U.S. associated with manufacturing plants, processing plants, landfills and mining sites and oil-contaminated sites are the main concern in this text because the current evidence suggests potential health impacts due to exposure to oil pollutant, such as cancer, liver damage, immunodeficiency and neurological symptoms. Adverse impacts to soil, air and water quality in oil-contaminated areas were also identified [5,6].

Machine learning is an effective tool for finding relationships within the system. In the past 20 years, with the emergence of more open-source algorithms and the publication of soil data, machine learning has been applied in pollution remediation. For example, machine learning calibration models interpret soil spectral data and predict contamination levels [7]. In addition, machine learning models can predict the bioavailability of 16 polycyclic aromatic hydrocarbons (PAHs) in compost-contaminated soils [8]. Other developments are a back propagation-approximate nearest neighbour (BP-ANN) model for rapid prediction of PAH concentrations in soils [9] and an artificial neural network for prediction of PAH levels in Caspian Sea sediments [10]. There are numerous machine learning model predictions on flora and fauna communities in contaminated soils [11–13]: a machine learning model was used to predict the HM immobilization efficiency of biochar in biochar-amended soils to study soil heavy metal immobilization and to predict the remediation results [14,15]; the adsorption of six heavy metals on 44 biochar were modelled using artificial neural networks and random forests, and the adsorption efficiency was accurately predicted, concluding that the characteristics of the biochar had the greatest influence on the adsorption efficiency [16]. Three modelling approaches were used to model various levels of oil-contaminated soils, identify data before and after oil spills, and effectively identify oil-contaminated and uncontaminated sites [17]. However, these are works for a given pollutant and a specific site. Therefore, the variations of oil-contaminated sites and pollutant categories need to be studied over time and space in order to gain a clear understanding of how the various factors affecting the remedial actions, to look for potential links between independent cases and to improve the systematicity and guidance of the CERCLA data to some extent. Specifically, pollutant information and restoration status in each state were systematically and theoretically analysed, organized and compared. Furthermore, a decision tree model was constructed to predict the remediation methods to be adopted under different pollution conditions. Finally, the weights that each feature contributes to the choice of remediation method were determined. In this paper, information on the type of pollutants and remediation operations at various remediation sites in Texas and other areas with high relevance to the petroleum industry was obtained. Using a machine learning method, pollutant information and remediation status in each state were theoretically analysed, summarized and compared.

2. Methods

2.1. Data Collection

The U.S. EPA's website (<https://www.epa.gov/superfund/search-superfund-sites-where-you-live>, accessed on 15 May 2021) documents various contaminated sites in the U.S. and their remediation status. The data can be divided into two types: data from tables on the database site (as used in the present study) and detailed data such as strata, pollution and restoration in PDFs. The PDFs data format and content of each contaminated site vary considerably owing to large time scales, missing data and other problems. Therefore, an automated data acquisition tool was used to obtain pollutant types and remediation practices for various locations in four states (California, Texas, Alaska and Oklahoma) from the CERCLA's website.

2.2. Model Selection and Methods

After acquiring the data, the existing information was processed for machine learning. First, a Decision Tree Classifier model was selected. Data such as contaminant names and clean-up technologies are discrete and hard to standardize. Additionally, the amount of data was not sufficient and the Decision Tree Classifier requires less data for training. It is insensitive to missing values, has a low requirement for data normalization and is able to handle numerical and categorical data, while other models can usually only be used to analyse datasets specifically for a particular type of variable. As an example-based inductive learning method, it was possible to classify decision trees from unordered training samples into tree models in real time, and people are capable of understanding the meaning of a decision tree after it has been explained. The code language was Python 3.9, the compiler was PyCharm and the Decision Tree Classifier algorithm was derived from scikit-learn (<https://scikit-learn.org/stable/index.html>, accessed on 10 May 2021).

The data were divided into time; location; benzene, toluene, ethylbenzene and xylene (BTEX); persistent organic pollutants (PoPs); PAHs; metals; highly toxic substances (Cyanide, dichlorodiphenyltrichloroethane (DDT) and other chemicals that are extremely harmful to the human body are defined as ‘highly toxic substances’ here); and clean-up technologies. The first six types of data were used as characteristic variables in the model, which attempts to predict clean-up technologies. Time was the node on which the site underwent remediation operation the first time and was standardized to −1980 so that the values were not too large and the resulting weight was not too high. The reason for not performing the normalization process was that time was the only digital variable. The time difference after the normalization process was too small to reduce the time weight artificially. Location characteristics were numbered according to the state where the site is located. In the processing of this article, California was recorded as 0 and Texas as 1. These characteristics are present in BTEX, PoPs, PAHs, metals and highly toxic substances. If such substances appear in the pollutant list of a site, the corresponding characteristic value of a given substance is 1, otherwise it is 0. The clean-up technologies are divided into three categories: physical repair corresponding to 1, chemical remediation corresponding to 2 and biological repair corresponding to 3. For sites that adopted multiple repair methods, the corresponding label was the repair method with the highest value, because two or more remediation methods were used in almost every case, but the number of specific repair techniques used in these three remediations varied from case to case. Therefore, it is meaningful to identify the main remediation as the main one; this is also to simplify the data for modelling purposes. For example, if a site simultaneously had physical and chemical repairs, and the proportion of the physical repair was greater, the corresponding label was 1.

3. Results and Discussion

The composition and content of pollutants could change over time. Hence, relative remediation operations should be adopted for these contaminated areas. Among all the remediation cases of oil contamination, California was first analysed as an example. During the period 1983 to 2020, there were 114 sites in California, among which 20 sites were still undergoing site surveys and were not initiated with clean-up operations, so the 94 sites that have started clean-up operations were analysed to reveal the changes in pollutants and remediation methods during the 38-year time period.

3.1. Changes in Pollutants and Remediation Methods over Time

As shown in Figure 1, the 94 sites that had already been repaired are sorted chronologically. The histogram shows the change of the number of repaired sites over time, with five years as the interval, based on when each site was first repaired. The number of sites increased mainly between 1985 and 1995, with the greatest increase between 1990 and 1995 with 37 sites. During the last decade, the increase in the number of sites slowed down.

This might be due to the improvement of laws and regulations and people's increased awareness of environmental protection.

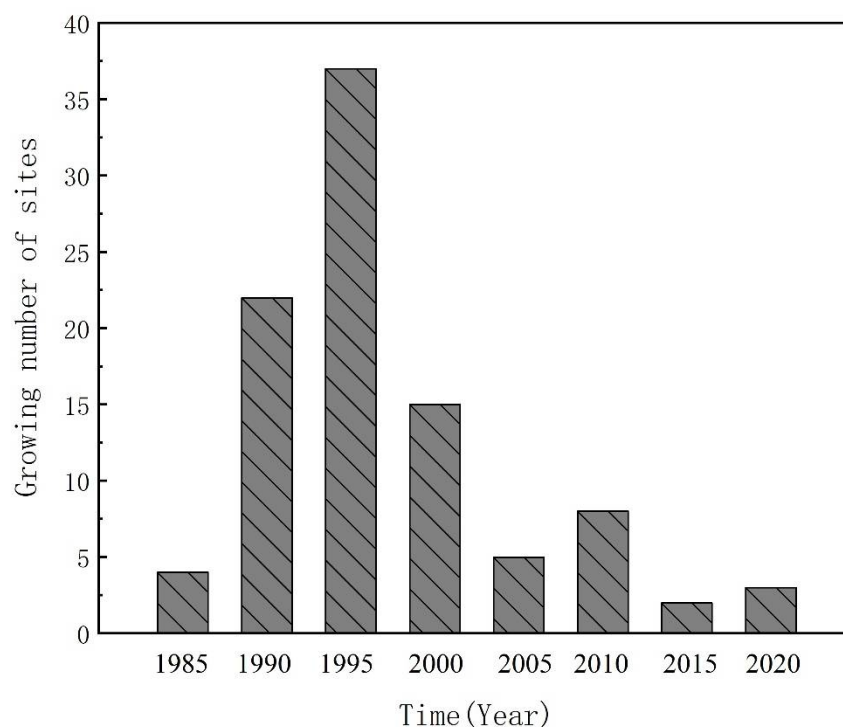


Figure 1. The number of sites growing over time.

As indicated by the black line and red line in Figure 2, the sites that contain highly toxic substances in the pollutant list were counted to draw the curve of the total number of sites and the number of sites that contain highly toxic substances in the pollutant list over time. The blue line in Figure 2 shows the number of sites with PAHs in the pollutant list and the curve of the number of sites over time. Because California is an oil-producing state in the U.S., this study focused on petroleum pollutants and benzene series such as BTEX; most benzene series in soil and surface water are derived from the leakage of petroleum products. The green line in Figure 2 shows the sites with BTEX in the pollutant list and the curve of the number of sites with BTEX in the pollutant list over time. In addition to the organic pollutants mentioned above, heavy metal pollution such as mercury, copper, zinc and cadmium was also studied. The purple line in Figure 2 shows the number of sites with heavy metals in the pollutant list and the curve of the number of sites with heavy metals contamination over time.

As shown in Table 1, the percentage of sites containing various pollutants relative to the total number of sites changed over time. Prior to 1985, the number of samples containing highly toxic contaminants was small, and the reference values were not large. Analysis of other data shows that since 1990, the percentage of sites containing highly toxic substances relative to the total number of sites are increasing, from 3.8% in 1990 to 25.5% in 2020. From 2000 onwards, however, there are marked declines in the rate of site growth, which may be due to the enforcement of different laws and regulations. The percentage of sites containing PAHs to the total number of sites fluctuates to some extent, but is around 30%. The percentage of sites containing BTEX is very high. In most cases, the percentage reaches 70%. As of 2020, the probability of metal pollution in the soil environment reached 50%, a relatively common type of pollution in California. This means that about half of the polluted areas in California are contaminated with heavy metals and one-third with PAHs. About 70% of the contaminated areas contain BTEX. This is consistent with the objective situation, as California has more oil pollution.

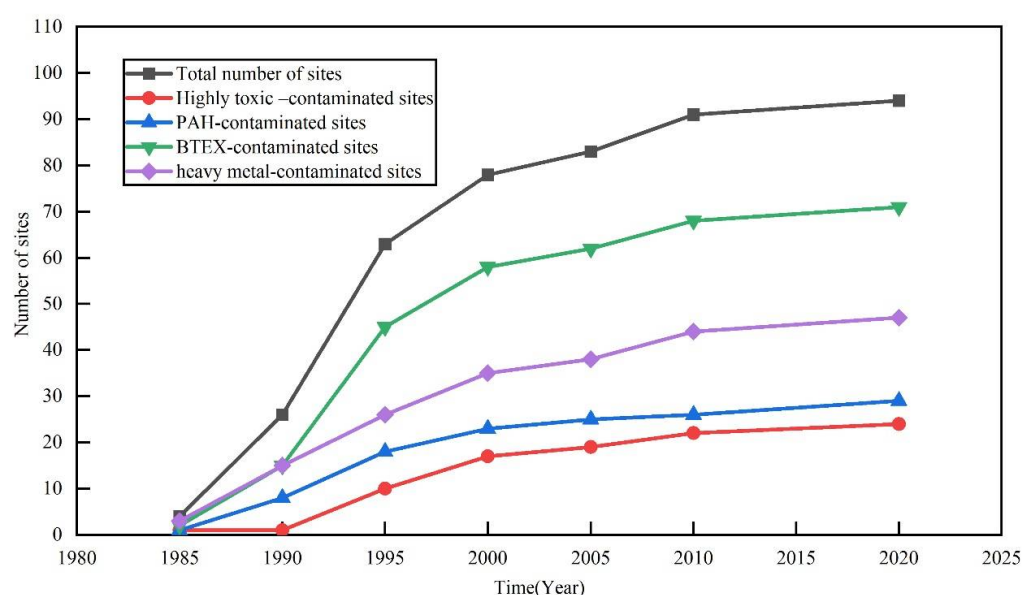


Figure 2. Changes in total sites and sites containing four types of pollutants over time.

Table 1. Number and percentage of the number of sites containing four types of pollutants.

Total Number of Sites	Highly Toxic-Contaminated Site	Highly Toxic-Contaminated Site/Total Number of Sites	PAH-Contaminated Sites	PAH-Contaminated Sites/Total Number of Sites	BTEX-Contaminated Sites	BTEX-Contaminated Sites/Total Number of Sites	Heavy Metal-Contaminated Sites	Heavy Metal-Contaminated Sites/Total Number of Sites
4	1	25.0%	1	25.0%	2	50.0%	3	75.0%
26	1	3.8%	8	30.7%	15	57.7%	15	57.7%
63	10	15.8%	18	28.6%	45	71.4%	26	41.3%
78	17	21.9%	23	29.5%	58	74.3%	35	44.9%
83	19	22.9%	25	30.1%	62	74.7%	38	45.8%
91	22	24.2%	26	28.5%	68	74.7%	44	48.3%
94	24	25.5%	29	30.9%	71	75.5%	47	50.0%

3.2. Changes in Pollutants and Remediation Methods in Space

Not only do contaminants and remediation methods change over time scales, but local remediation methods also change in response to differences in contaminant distribution among regions. Therefore, the distribution of pollutants in Texas, Alaska and Oklahoma was analysed and compared horizontally with the distribution of pollutants in California to obtain changes in pollutants on a spatial scale. Then, the changes in the restoration methods for Texas over time were analysed and compared horizontally with California's restoration methods to obtain the change in restoration practices on a spatial scale. Table 2 shows the distribution of pollutants and their proportions in California, Texas, Alaska and Oklahoma as of 2020.

Table 2. Distribution of Pollutants in the four states (as of 2020).

Pollutant Category	Number of Sites in California with Such Pollution	Percentage of Sites in California with Such Pollutants	Number of Sites in Texas with Such Pollution	Percentage of Sites in Texas with Such Pollutants	Number of Sites in Alaska with Such Pollution	Percentage of Sites in Alaska with Such Pollutants	Number of Sites in Oklahoma with Such Pollution	Percentage of Sites in Oklahoma with Such Pollutants
BTEX	71	75.5%	37	74.0%	5	71.4%	8	53.3%
PAHs	29	30.9%	20	40.0%	3	42.9%	5	33.3%
Highly Toxic	24	25.5%	7	14.0%	4	57.1%	2	13.3%
Heavy Metal	47	50.0%	28	56.0%	5	71.4%	12	80.0%

Texas and California are major oil-producing states in the U.S., and the proportion of sites containing BTEX reaches about 75%. However, the proportion of sites containing PAHs in Texas is about 10% higher than that in California. Sites containing highly toxic substances are more common in California, and the probability of the occurrence of sites containing highly toxic substances in Texas is only 14%. Heavy metal pollution is common in the two states, and the likelihood of its occurrence in Texas and California is about 50%.

Compared to those in Texas and California, the pollution of Alaska by highly toxic substances and heavy metals is more common, reaching 57.1% and 71.4%, respectively. The frequency of occurrence of BTEX and highly toxic substances are decreased in contaminated areas in Oklahoma compared to other states. However, the abundance of heavy metals reaches 80%, the highest among the four states. It is worth noting that the number of sites in Alaska and Oklahoma is relatively small, so the statistics may be accidental and are for reference only. Because sites in Alaska and Oklahoma are few, only the changes in the four types of pollutants in Texas over time were analysed, contrasting with the California situation. Figure 3A–D show that the number of sites with BTEX, PAHs, highly toxic substances and heavy metals in the list of pollutants changed over time in California and Texas, respectively.

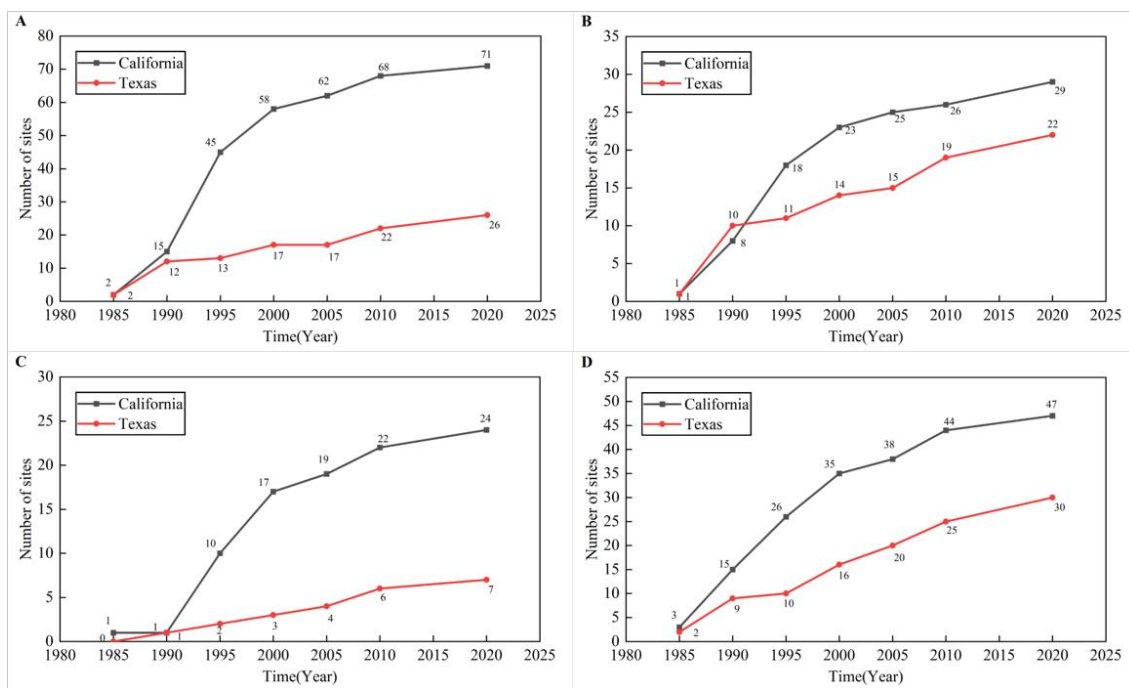


Figure 3. Changes in sites containing four types of pollutants in California and Texas over time. (A) the number of sites with BTEX pollutants changed over time in California and Texas. (B) the number of sites with PAHs pollutants changed over time in California and Texas. (C) the number of sites with highly toxic substances pollutants changed over time in California and Texas. (D) the number of sites with heavy metals pollutants changed over time in California and Texas.

3.3. Exploration of Remediation Methods by Machine Learning

The distribution of pollutants in California, Texas, Alaska and Oklahoma and their relevant spatial distribution can be obtained using traditional graphical analysis of the existing data. However, the traditional chart is not sufficiently comprehensive to guide the repair method. Therefore, Decision Tree Classifier as a machine learning model was used to build a decision tree model based on time, space, five types of pollutants and three remediation methods. These features were used to predict the remediation methods that should be considered under different pollution conditions and obtain the weight contributed by each feature when choosing the remediation methods. The model was built

only based on California, Texas and these two states combined; the number of cases in Oklahoma and Alaska was too small to build a model.

As shown in Figure 4 and Table 3, only the California pollution data are considered. Therefore, the correct cross-validation rate is 0.656, the location does not affect the choice of remediation methods, the weight is zero and the weight of other organic pollutants is also zero. This shows that other organic pollutants do not affect the choice of remediation methods. Of the remaining five characteristics, pollutants containing BTEX had the smallest impact on the choice of remediation method, with a weight of only 0.035. Conversely, pollutants containing highly toxic substances had the greatest impact on the choice of remediation method, with a weight of 0.274.

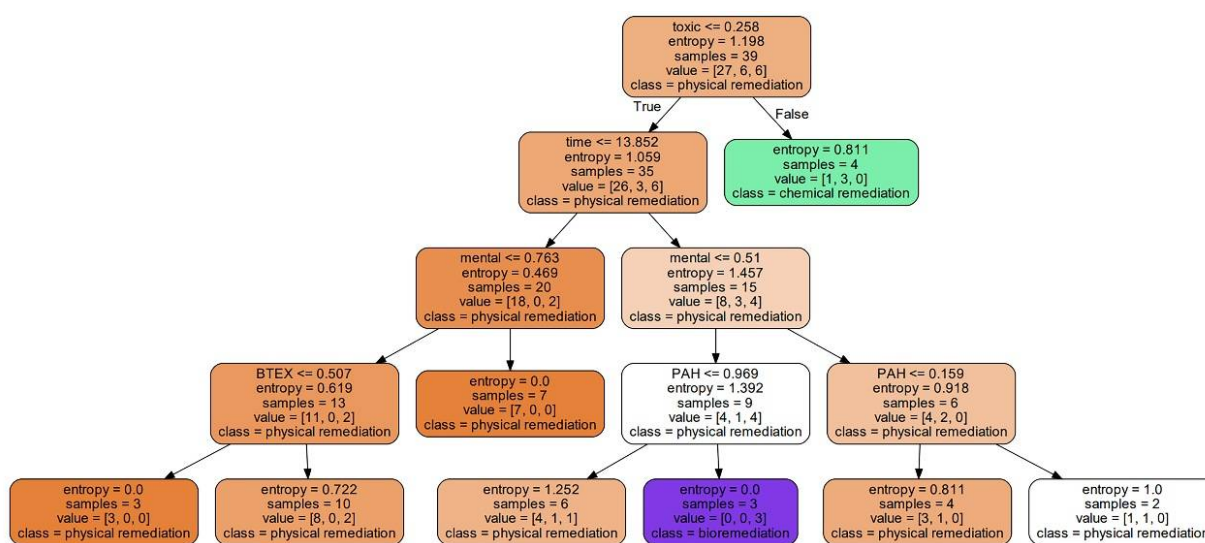


Figure 4. California's decision tree model.

Table 3. Weight for each feature.

Features	Weight
Time	0.248
Space	0.000
BTEX	0.035
PAHs	0.225
Heavy Metal	0.218
Highly Toxic	0.274
Organic pollutants	0.000

As shown in Figure 5 and Table 4, only the pollution data of Texas are considered. Therefore, the correct cross-validation rate is only 0.511, and the location weight is still zero. Contrary to California, Texas does not consider highly toxic substances when selecting remediation methods but focuses on metals and other organic pollutants. This may be because highly toxic substances are rare in Texas. Comparing the data of the two states, different factors are considered when choosing remediation methods. This may be due to differences in pollution, developments and laws and regulations in each state.

As shown in Figure 6 and Table 5, the data of Texas and California are modelled as datasets, and the correct cross-validation rate is 0.592. The decline in cross-validation accuracy may be due to the different choices of repair strategies from state to state. By adding Texas data, the location (space) information for each site in the dataset will not be the same. Furthermore, the weight of space increases to 0.288, ranking second among all features, explaining the significant weight difference between Texas and California. The weight of other organic pollutants will not be zero but only 0.045, indicating that it

hardly affects the choice of remediation methods. Of the remaining five characteristics, the pollutants containing BTEX had the greatest impact on the choice of remediation methods, with a weight of 0.301. The weight of highly toxic substances significantly decreased compared to California data. The weights of time, heavy metals and PAHs were also reduced.

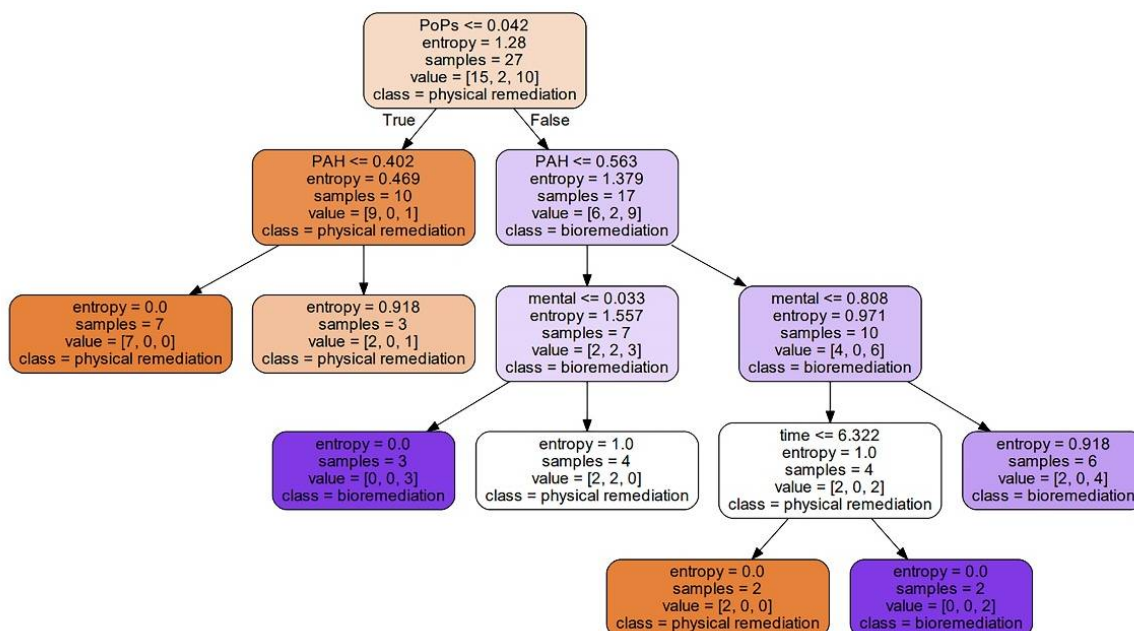


Figure 5. Decision tree model for Texas.

Table 4. Weight for each feature.

Features	Weight
Time	0.179
Space	0.000
BTEX	0.000
PAHs	0.214
Heavy Metal	0.318
Highly Toxic	0.000
Organic pollutants	0.288

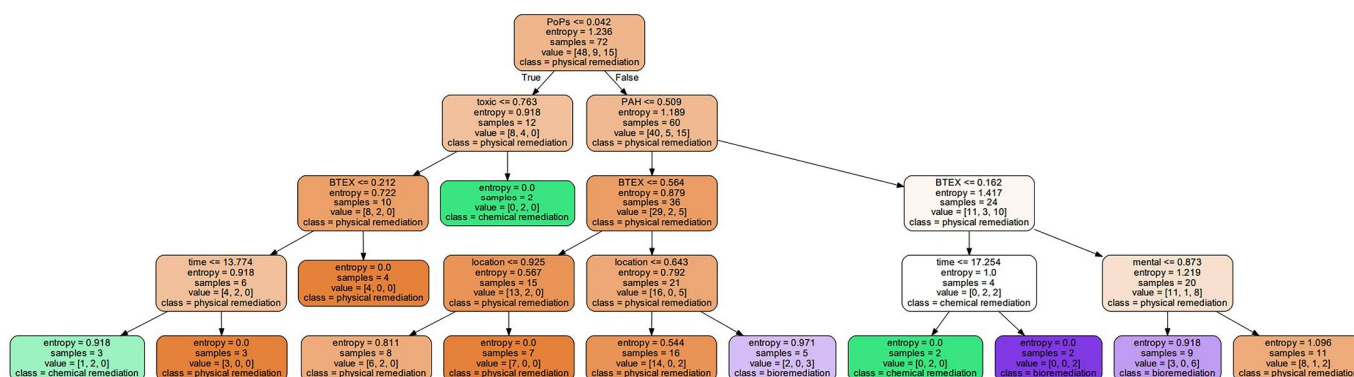


Figure 6. Decision tree model for the sum of California and Texas.

In terms of the weighting of features between California and Texas, in contrast to California, Texas does not consider highly toxic substances when selecting remediation methods, but instead focuses on metals and other organic pollutants. This might be

explained by the rarity of highly toxic substances in Texas. These two states consider different factors in the selection of their remediation methods, possibly reflecting differences in pollution, development and laws and regulations in each state.

Table 5. Weight of each feature.

Features	Weight
Time	0.147
Space	0.288
BTEX	0.301
PAHs	0.124
Heavy Metal	0.088
Highly Toxic	0.083
Organic pollutants	0.145

The three groups of decision tree models have the highest cross-validation accuracy rate of 0.656. The California single decision tree model has a maximum data volume of 144 sites. A model was obtained by adding California and Texas data, but the cross-validation accuracy rate is only 0.592. The decline in cross-validation accuracy may be due to the different choices of repair strategies from state to state. The reasons for the low accuracy rate can be as follows:

One hundred forty-four pieces of data of cases were not big enough to train a decision tree model—the model could be easily over-fitted or under-fitted. The reason for the lack of data is that the information in the database, in many cases, was incomplete and could not be analysed by machine learning. Furthermore, data were simplified at the pre-processing stage, and only the intrinsic characteristic variables in the analytical cases were selected for the convenience of this study, and the selection of variables might not be comprehensive enough to represent the whole case. In future studies, more quantitative specific characteristics such as strata, pollution distribution and restoration effect could be added for in-depth analysis.

4. Conclusions

Among 211 pollution project datasets for four states collected from the EPA Superfund CERCLA Database, 144 soil remediation projects in four major oil-producing states (California, Texas, Oklahoma and Alaska) were extracted for modelling by Decision Tree Classifier. The correlations among the pollutant species, pollutant site characteristics and the selected remediation methods were analysed.

- (1) Among the three repair methods, physical repair was the most commonly used one, with an application probability stable at over 80% at all time periods. Chemical remediation techniques were rarely used before 1990, and the frequency of chemical remediation techniques selected for most periods after 1990 reached more than 50%. The frequency of selecting bioremediation technology was the smallest, less than 50% at all times. However, the proportion of bioremediation applications has risen in general.
- (2) California and Texas are the major oil-producing states in the U.S. The proportion of sites containing metal pollution was about 50% in these two states. BTEX was more common in both states, and its occurrence probability in Texas and California was about 75%. There were more sites containing highly toxic substances in California than in Texas, reaching more than 25%. On the other hand, there were more sites containing PAHs in Texas than in California, reaching more than 40%.
- (3) Of the seven characteristic variables selected for this paper, considering California alone, the presence or absence of highly toxic substances has the greatest influence on the choice of remediation method, and the weight is 0.274, while in Texas the feature with the highest weight is heavy metal content, which has a weight of 0.318. The combined model of these two states showed that the presence or absence of BTEX

substances and the geographical location of the site largely influenced the choice of remediation method, and the weights reached 0.301 and 0.288, respectively.

Author Contributions: Conceptualization, T.L., H.L., Z.Z. and X.W.; methodology, Z.Z. and T.L.; data curation, Y.W.; formal analysis, writing—original draft preparation, Z.Z. and Y.W.; writing—review and editing, H.L., S.L., J.X. and X.W.; supervision, X.W.; funding acquisition, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Shandong Provincial Natural Science Foundation, China, grant number ZR2021JQ18; Shandong Provincial Natural Science Foundation, grant number ZX20220090 and the General Program of National Natural Science Foundation of China, grant number 52074337.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors sincerely acknowledge the assistance in editing from Lianjie Hou.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Wiener, J.B. Developments in the Law of Toxic Waste Litigation: Bankruptcy and Insurance Issues. *Harvard Law Review* **1985**, *99*, 1458.
- Hamilton, J.T.; Viscusi, W.K. How Costly Is “Clean”? An Analysis of the Benefits and Costs of Superfund Site Remediations. *J. Policy Anal. Manag.* **1999**, *18*, 2–27. [\[CrossRef\]](#)
- Ma, B.Q.; Tang, C.; Wang, X. Thought on the love canal moment and superfund for groundwater prevention and control in China. *Groundwater* **2021**, *43*, 18–21+55.
- Jasanoff, S. Review: Love Canal revisited: A Hazardous Inquiry: The Rashomon Effect at Love Canal. *Issues Sci. Technol.* **1998**, *14*, 15–31.
- U.S. Environmental Protection Agency. The Origins of EPA | US Environmental Protection Agency. Available online: www.epa.gov/history/origins-epa (accessed on 15 May 2021).
- Johnston, J.E.; Lim, E.; Roh, H. Impact of upstream oil extraction and environmental public health: A review of the evidence. *Sci. Total Environ.* **2019**, *657*, 187–199. [\[CrossRef\]](#) [\[PubMed\]](#)
- Jia, X.Y.; O'Connor, D.; Shi, Z.; Hou, D.Y. VIRS based detection in combination with machine learning for mapping soil pollution. *Environ. Pollut.* **2021**, *268*, 115845. [\[CrossRef\]](#) [\[PubMed\]](#)
- Wu, G.Z.; Kechavarzi, C.; Li, X.G.; Wu, S.M.; Pollard, S.J.T.; Sui, H.; Coulon, F. Machine learning models for predicting PAHs bioavailability in compost amended soils. *Chem. Eng. J.* **2013**, *223*, 747–754. [\[CrossRef\]](#)
- Bao, H.Y.; Wang, J.F.; Li, J.; Zhang, H.; Wu, F.Y. Effects of corn straw on dissipation of polycyclic aromatic hydrocarbons and potential application of backpropagation artificial neural network to prediction model for PAHs bioremediation. *Ecotoxicol. Environ. Saf.* **2019**, *186*, 109745. [\[CrossRef\]](#) [\[PubMed\]](#)
- Shadrin, D.; Pukalchik, M.; Kovaleva, E.; Fedorov, M. Artificial intelligence models to predict acute phytotoxicity in petroleum contaminated soils. *Ecotoxicol. Environ. Saf.* **2020**, *194*, 1010410. [\[CrossRef\]](#) [\[PubMed\]](#)
- Cipullo, S.; Snapir, B.; Prpich, G.; Campo, P.; Coulon, F. Prediction of bioavailability and toxicity of complex chemical mixtures through machine learning models. *Chemosphere* **2019**, *215*, 388–395. [\[CrossRef\]](#) [\[PubMed\]](#)
- Olawayin, R. Application of backpropagation artificial neural network prediction model for the PAH bioremediation of polluted soil. *Chemosphere* **2016**, *161*, 145–150. [\[CrossRef\]](#) [\[PubMed\]](#)
- Amin, J.S.; Kuyakhi, H.R.; Bahadori, A. Prediction of formation of polycyclic aromatic hydrocarbon (PAHs) on sediment of Caspian Sea using artificial neural networks. *Pet. Sci. Technol.* **2019**, *37*, 1987–2000. [\[CrossRef\]](#)
- Palansooriya, K.N.; Li, J.; Dissanayake, P.D.; Suvarna, M.; Li, L.Y.; Yuan, X.Z.; Sarkar, B.; Tsang, D.C.W.; Rinklebe, J.; Wang, X.N.; et al. Prediction of Soil Heavy Metal Immobilization by Biochar Using Machine Learning. *Environ. Sci. Technol.* **2022**, *56*, 4187–4198. [\[CrossRef\]](#) [\[PubMed\]](#)
- Lhafra, F.Z.; Abdoun, O. Hybrid Approach to Recommending Adaptive Remediation Activities Based on Assessment Results in an E-learning System Using Machine Learning. In Proceedings of the 3rd International Conference on Advanced Intelligent Systems for Sustainable Development (AI2SD), Tangier, Morocco, 21–26 December 2020; pp. 679–696.
- Zhu, X.Z.; Wang, X.N.; Ok, Y.S. The application of machine learning methods for prediction of metal sorption onto biochars. *J. Hazard. Mater.* **2019**, *378*, 120727. [\[CrossRef\]](#) [\[PubMed\]](#)
- Kaplan, G.; Aydinli, H.O.; Pietrelli, A.; Mieleve, F.; Ferrara, V. Oil-Contaminated Soil Modeling and Remediation Monitoring in Arid Areas Using Remote Sensing. *Remote Sens.* **2022**, *14*, 2500. [\[CrossRef\]](#)