TRACEABILITY OF OIL SPILL FROM BAYESIAN CLASSIFICATION

Dan Huang[1, 2,*] , Ya Zhang[1, 2], Wenqing Yu[1, 2]

[1] College of Geomatics, Shandong University of Science and Technology, Qingdao, China; (1508643238, 2291403018, 944943566) @qq.com
[2] The Key Lab of Surveying and Mapping Technology on Island and Reef, Ministry of Natural Resources, China;

Commission VI, WG VI/4

KEY WORDS: Bayesian classification; marine oil spill; cross-validation; Naive Bayesian; oil spill source determination

ABSTRACT:

Oil spills over the sea resulted in the destruction of the marine environment. Therefore, the Bayesian algorithm was developed to determine the source of spilled oil for reducing the damage to the marine environment. Based on the flow data of the offshore waters of the Bohai Sea, the influencing factors of oil spill drift were used for algorithm. The main factors affecting the traceability of marine oil spills are considered: the distance from the oil point to the source point, the flow direction of water, the type of the spilled oil source, and the scale of the source. The algorithm is completed on the basis of a feasible simulation database. We not consider the relationship between the attributes by naive Bayesian classification. In this paper, we determine the result by the minimum error rates of each source. Then the performance evaluation of the model was done by cross-validation method. The experimental result shows that the Bayesian algorithm can be used to determine the source of spilled oil. It is easier and faster to determine the source by the method raised.

* Corresponding author

1. INTRODUCTION

The oil spills have caused damage to the economic development and human of the affected coastal areas. In the event of marine oil spills, if the source of the oil spill cannot be directly identified, it is difficult to confirm the person responsible for the accident. Penalties for those responsible will cause public concern about the marine environmental pollution. The existing methods are based on the "Lagrange" particle tracking method to establish a model for predictive analysis (Cao, 2016), or the use of oil fingerprint library from oil components (Zhang, Wang, 2016), or the simple distance judgment. To meet the needs for the source determination of oil spill, so we propose a naive Bayesian classifier for this practical problem in this paper.

The structure of this paper is as follows. First, we discuss the theory of Bayesian classification. Bayesian classification was applied to determine the oil spill source after analysing the factors of oil spill drift. Second, the superior performance of Bayesian classification is proved by experiments, and the correct rate of classifier is evaluated by cross-validation. Finally, the performance of the method application is summarized.

2. BAYESIAN CLASSIFICATION

The Bayesian classification is characterized by the use of probabilities to represent all forms of uncertainty, and learning or reasoning is done using probability rules. When the amount of sample data is large, the frequency of each sample obtained from a large number of repeated tests tends to a stable value. Therefore, the frequency of occurrence of an event is often treated as the probability of occurrence of an event in actual problems, and there may be more than two categories in practice, which may be more complicated. Then, the Bayesian formula in the complex case is as follows:

$$P(B_i|A)=\frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{n}P(B_j)P(A|B_j)} \quad (1)$$

where $P(B_i)$ = the probability of event B_i

$P(A|B_i)$ =the probability of even A under the condition that event B_i has occurred

$P(B_i|A)$ =the probability of event B_i under event A

We applied the minimum error rate for naive Bayesian classification. When the prior probability of a category appears and the conditional probability density of the sample distribution in each class, a posterior probability of each class to be classified can be obtained. For a given item to be classified $X=\{a_1,a_2,...a_n\}$, the set category $C=\{y_1,y_2,...y_m\}$, find the probability of occurrence of each category y_i under the conditions, and thus the items to be classified. Expressed as follows:

$$\text{if } P(Y=y_j|X=x)=\max\begin{Bmatrix}P(Y=y_1|X=x),\\P(Y=y_2|X=x)\\,,,\\P(Y=y_m|X=x)\end{Bmatrix} \text{ then } X\in y_j.$$

3. NAIVE BAYESIAN CLASSIFIER MODEL

3.1 Study Area

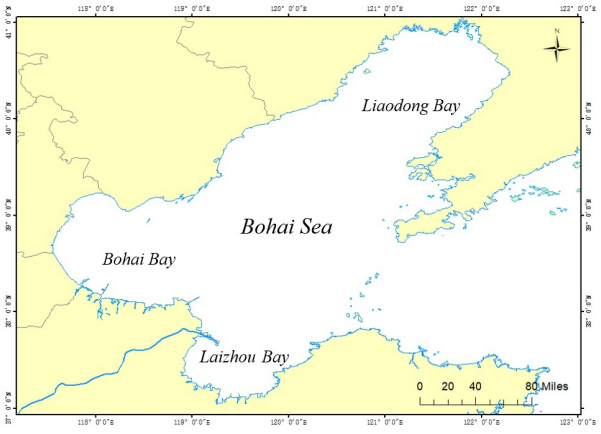

Figure 1. Distribution of oil spills in the Bohai area

The Bohai Sea is the highest latitude territorial water in China and is located at 37°07′-41°00′N and 117°35′-121°10′E. The Bohai Sea consists of Liaodong Bay, Laizhou Bay, Bohai Bay and the central Bohai Sea, as shown in Figure 1. The north, west and south sides of the sea are surrounded by land and the terrain is complex. There are many ports and oil fields in the Bohai Sea, and the incidents of marine oil spills occurs. The Bohai Sea urban agglomeration is an important economic zone in China, and the regional GDP has reached 3.8 trillion yuan, accounting for 28.2% of the national total. Therefore,

monitoring of marine oil spills is important for offshore environmental protection and economy.

3.2 Factors for Oil Spill Drift

In offshore waters, tidal currents and wind currents are important factors in determining oil spill drift. Due to the limited statistics on marine oil spill accidents, the sources of oil spill have been determined by the type and scale of the possible oil spill sites. Therefore, the analysis of the oil spill source identification based on the simulation data is reliable in this paper. The oil spill movement affected by the oil spill source and the ocean flow field. We use the distance from the oil source to oil spill point, the current direction, the type of oil source, and the size of the oil source to make a determination.

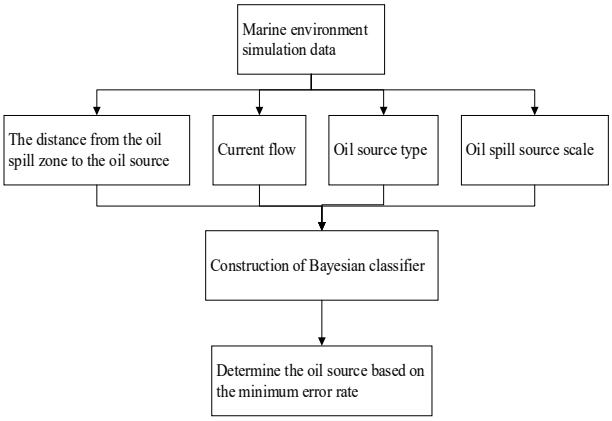

Figure 2. Flow chart to classify by using the simulation data

We grasp the environment of oil spill on the sea firstly, then the simulation data of the characteristic database is prepared by various indicators, and the oil source is determined by Bayesian classifier. The technical route of oil spill source determination based on Bayesian classifier is shown in the Figure 2. Taking the marine environment simulation data as the background, four indicators for Bayesian classifier model including the distance from the oil spill point to the oil source, the angle between the current flow direction and the oil spill source to the oil point, the type of oil source (such as ship accident, oil field leakage), oil spill source size (on the order of magnitude of oil spills). The oil spill source is determined based on the minimum error rate.

3.3 Representation of Environmental Simulation Data

3.3.1 Current Flow Direction: The flow field of the ocean is changing all the time. There are differences in flow velocity and flow direction at different points in the same moment, as shown in Figure 3. The vector direction of the arrows represents the direction of the current flow. The ocean flow field is affected by various factors such as the season and the wind field, thus we need to simplify the flow velocity and direction of the sea area. We take the flow field of a certain day in mid-June as an example to carry out model construction and verification.

It is worth noting that the angle value used to construct the Bayesian classifier is the declination value between the vector direction of the oil spill point to the oil spill source and the main direction of the flow field. Because of other unstable factors, oil spills cannot simply move in one direction.

This paper mainly estimates the ocean current field based on a day in June (8:00 on June 10 to11:00 on June 11,2013). Because the amount of data in the area is very large, and the flow field changes complexly in the offshore area, we mainly select the flow velocity and direction of the twenty-four moments of the eight points dispersed in the Bohai Sea from the flow field attributes. Then we vectorize the flow rate for each moment. After calculating the declination value between the vector direction of the oil spill point and the starting point of the oil spill and the main direction of the flow field, and then used as the angle value of the database for the naive Bayesian classifier.

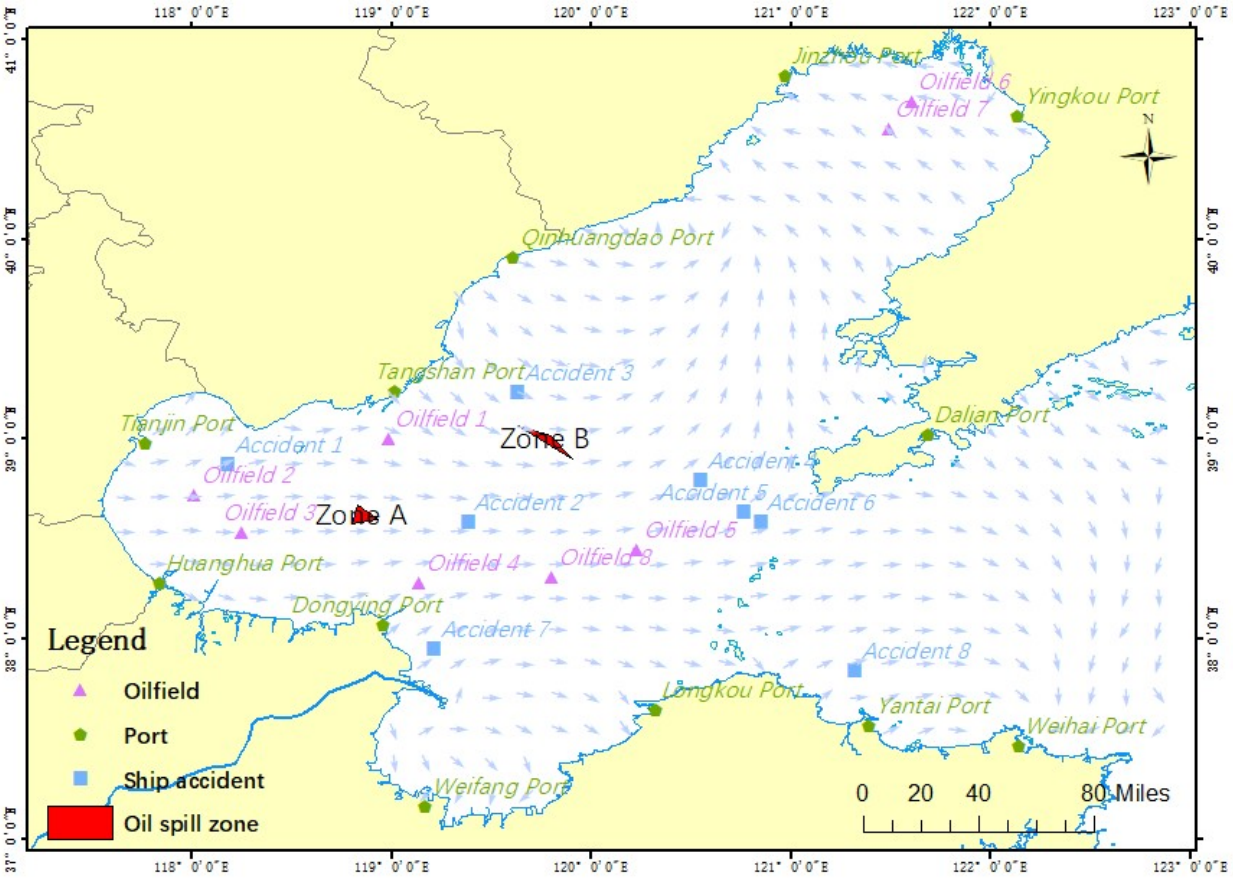

Figure 3. Ocean flow field diagram at 10:00 on June 10,2013

3.3.2 Distance from Oil Spill Point to Oil Source: The distance between the oil spill point and the oil source is also related to the time of movement. The following distances are calculated based on the drift time of one day. The day is divided into 24 parts at one-hour intervals, and the corresponding flow field speed can be obtained for each time period.

Figure 4 shows the average flow velocity of the current over the course of the day varies greatly with time. We use the flow rate at a certain moment as the flow rate of one-hour period, such as the speed at 8 o'clock as the average speed from 8 o'clock to 9 o'clock, and thus we accumulate the sum to get the drift distance of the oil spill one day.

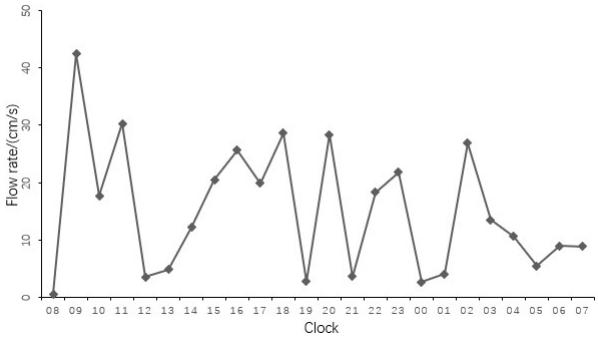

Figure 4. Flow chart of sea current flow rate with time

3.3.3 Oil source type: The causes of oil spill accidents are basically ship collision, oil pipeline leakage, oil tank leakage, etc. Therefore, this paper can mainly classify the causes of oil spills at sea into three major categories for simplifying the data set: ship accidents, oilfield leaks, and port leaks.

3.3.4 Oil source scale: Oil spill accidents are graded internationally based on their size required. Level 1 refers to a small oil spill that can be handled and controlled by using the oil spill reaction resources of the area. Level 2 refers to other

oil spill response resources in the area to assist in the treatment and control of larger oil spills. Level 3 refers to large or catastrophic oil spills that require domestic or even international oil spill response forces to assist with handling and control.

The research is mainly based on domestic oil spill grading, and the oil source scale is divided into three categories for simulation data processing, which respectively represent the three levels of large, medium and small oil spills.

4. BAYESIAN CLASSIFIER PERFORMANCE EVALUATION

We use cross-validation to test the performance of the naive Bayes classifier. Cross-validation is a method that can test the performance of model classification. The idea is to take most of the data as a training set for model and then see if the remaining data fits the model (Ren, 2015). In a given model, take most of the samples to build the model, leaving a small portion of the sample to be forecasted with the model just created, and the prediction error of the small part of the sample, then record the sum of their squares. This process continues until all samples are forecasted once and only once.

The data used in this paper is the flow field data from 8:00 on June 10, 2013 to 8:00 on June 11,2013, and we can calculate from the flow field data of the day: distance D=1.1796km, flow field main direction θ= - 171°54′40″, draw as shown in Figure 5.

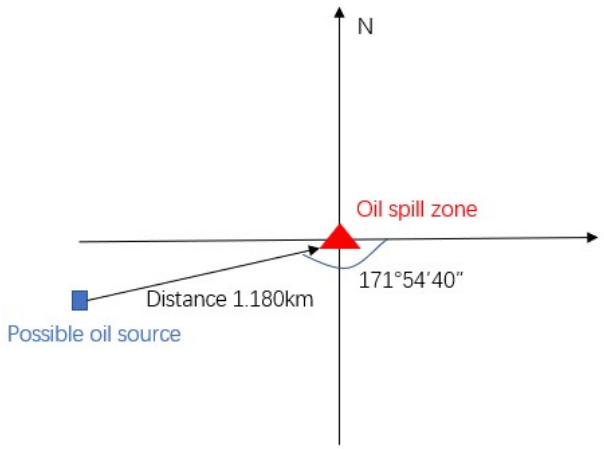

Figure 5. Flow field data simplified diagram

Assuming an existing example, the oil spill zone B was found in July of that year. Because it is considered that monitoring of marine oil spills will not be carried out every day, we use the flow rate of the day to establish the sample simulation data which making the model closer to the actual situation. The flow direction derived from the properties of the flow field map is the counter clockwise angle from the east direction, and the deviation value is obtained by combining the angle obtained in the above Figure 5. Table 1 shows the declination values of each possible oil source and main flow. And the formula is as follows:

$$\Delta = |\alpha - \theta| \tag{2}$$

where Δ = declination values of each possible oil source and main flow

α =the difference between the normal north and the oil spill zone to the possible oil source

θ = the flow direction to the main direction

Possible oil source	α/°	Declination/°
Tianjin Port	179.326	7.782
Huanghua Port	159.540	12.004
Dongying Port	131.135	40.409
Weifang Port	108.380	63.164
Longkou Port	68.373	103.171
Yantai Port	41.647	129.897
Qinhuangdao Port	280.872	87.584
Ship accident 1	174.710	3.166
Ship accident 2	133.363	38.181
Ship accident 3	205.396	63.060
Ship accident 4	15.888	155.656
Ship accident 5	21.084	150.460
Ship accident 6	22.466	149.078
Ship accident 7	118.073	53.471
Ship accident 8	37.224	134.320
Oil field 1	179.176	7.632
Oil field 2	170.502	1.042
Oil field 3	162.648	8.896
Oil field 4	131.906	39.638
Oil field 5	52.365	119.179
Oil field 8	88.624	82.920

Table 1. Deviation angles from the main stream

Possible oil source	Distance /km	Possible oil source	Distance /km
Tianjin Port	172.330	Ship accident 5	97.590
Huanghua Port	184.856	Ship accident 6	107.000
Dongying Port	126.470	Ship accident 7	129.140
Weifang Port	212.175	Ship accident 8	189.750
Longkou Port	160.535	Oil field 1	66.720
Yantai Port	216.131	Oil field 2	154.390
Qinhuangdao Port	100.000	Oil field 3	140.700
Ship accident 1	136.780	Oil field 4	98.150
Ship accident 2	57.850	Oil field 5	76.790
Ship accident 3	26.370	Oil field 8	79.200
Ship accident 4	73.220		

Table 2. Distances of oil spill point to oil source

The distance can be directly taken out from the map, and Table 2 shows the distance from the oil spill point to each possible oil source. Based on the established Bayesian classifier to determine the source of the oil spill zone B, it is concluded that the oil spill location is 66.72km, which is 7.63 degrees away from the general direction of the current flow, so oilfield 1 is the source of the oil identified this time. According to the actual situation, the location is closer and the declination value from the main direction is smaller. It is in line with the flow rate of the day, indicating that it is feasible to determine oil spill source with the Bayesian analysis. Similarly, it is determined that the oil spill source of the oil spill zone A is determined by the Bayesian classifier probability. Three locations may be oil spill sources, namely ship accident 1, oilfield 2 and oilfield 3, as shown in Table 3. We can see that the distances and declination values of the three places are relatively close. We should make a judgment considering the actual uncontrollable factors. Here we determine that the oilfield 3 with the lowest declination value is the source of the oil spill, and then conducts investigation and control of the oil spill.

Possible oil spill areas	Distance from oil spill area/km	Declination/°
Ship accident 1	64.48	19.57
Oil field 2	73.61	10.85
Oil field 3	52.76	0.23

Table 3. Oil spill source determination of zone A

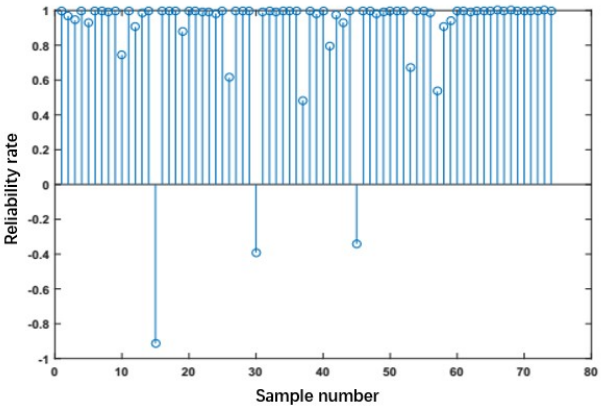

Figure 6. Result of Bayesian classifier cross-validation

The source of oil spill zone A is oilfield 3, the source of oil spill zone B is oilfield 1, and the drift of oil spill zone A and B is shown by the red arrow in Figure 7. Through the analysis of the flow field of the zone during the period, the oil traceability judgment is in line with reality, and it can be concluded that: Utilizing the Naive Bayes classifier, with the principle of minimum error rate, the source of marine oil spill can be determined.

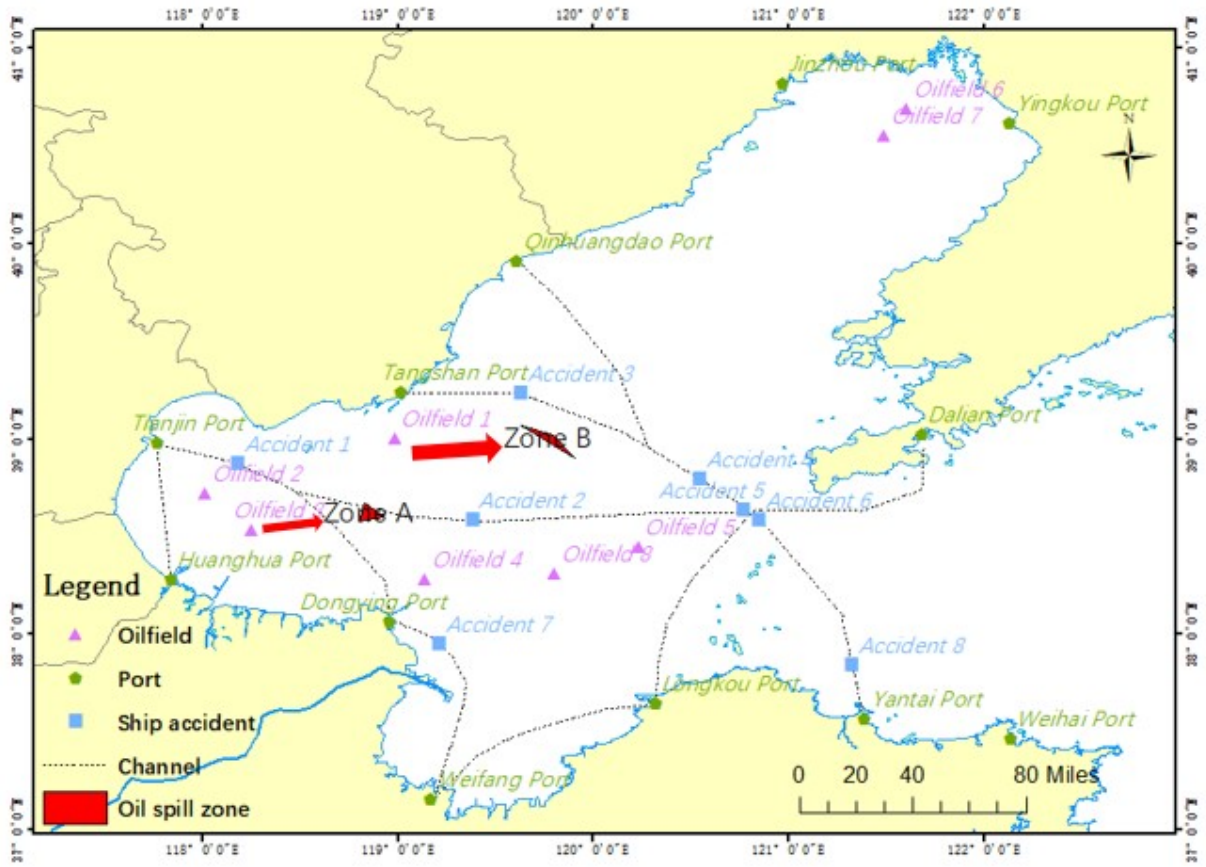

Figure 7. Schematic diagram of the directions of oil spill drift

Using the cross-validation test Bayesian classifier performance, the ten-fold cross-validation showed that there was a classification error in the total sample with a ratio of 0.0405, indicating that the Bayesian classifier has excellent application performance. Figure 6 is a plot of the results of cross-validation of the Bayesian model. It can be seen that among the 74 samples, most of the samples can obtain better classification results, but a few samples are not correctly classified. The abscissa of Figure 7 represents the training sample number, and the ordinate represents the reliability rate. The upper part of the figure indicates correct classification and the lower part indicates classification error.

5. CONCLUSIONS

We analyse the current research status of oil spill traceability, and discusses the principle of Bayesian classification. Then the Bayesian classifier and ocean flow field data are discussed, and the minimum error rate naive Bayes classifier is constructed to solve the oil spill traceability in the Bohai area. We analyse the factors affecting the oil spill drift: the distance from the oil spill source, the current flow, the type of oil source, and the scale of the oil source, then determine the final oil source. According to the simulation data, the source of the oil spill zone A and B is analysed. The source of the oil spill zone A is the oilfield 3, and the source of the oil spill zone B is the oilfield 1. Judging from the actual flow field data, the determination of the oil traceability is confirmed by the situation. The cross-validation function is used to verify the classification accuracy of the data, and the error rate is only 0.0405. The classification can be used in practice. And the model of this paper isn't carried out by redundant data collection and processing, and also makes up for the shortcomings of the original method. This method can find the source of oil spill more easily and quickly under the condition of data support for reducing the damage to the marine environment.

6. REFERENCES

C. Lännergren,2009. Net- and Nanoplankton: Effects of an Oil Spill in the North Sea. *Botanica Marina*,21(6).

F. R. Engelhardt,2013. Remote Sensing for Oil Spill Detection and Response. *Pure and Applied Chemistry*,71(1).

Haijiang Zhang, Chuanyuan Wang, Ruxiang Zhao, Xiaonan Yin, Hongyang Zhou, Liju Tan, Jiangtao Wang, 2016. New diagnostic ratios based on phenanthrenes and anthracenes for effective distinguishing heavy fuel oils from crude oils. *Marine Pollution Bulletin*,10: 58-61.

Hong Qi, Yuanye Ling, Jun Zhu, Wanqing Li, 2018. Evaluation of the Cross Validation on the Digital Soil Mapping of Microelements in Bo Zhou of North Anhui. *Chinese Journal of Soil Science*, 49(1): 9-15.

Huifeng Shi, 2013. The Load Forecasting Model which Parameters Optimized by MCMC Based on Bayesian Theory. *North China Electric Power University*.

Li Y, Cui C, Liu Z, et al, 2017. Detection and Monitoring of Oil Spills Using Moderate/High-Resolution Remote Sensing Images. *Archives of Environmental Contamination and Toxicology*, 73(1):154-169.

Menzies N A, Djøra I. Soeteman, Pandya A, et al, 2017. Bayesian Methods for Calibrating Health Policy Models: A Tutorial. *Pharmaco Economics*, 35(6):613-624.

Mohajerani H, Kholghi M, Mosaedi A, et al, 2017. Application of Bayesian Decision Networks for Groundwater Resources Management Under the Conditions of High Uncertainty and Data Scarcity. *Water Resources Management*, 31(6):1859-1879.

Newton R, Wernisch L, 2017. A comparison of machine learning and Bayesian modelling for molecular serotyping. *BMC Genomics*, 18(1).

Scott Farrow, Douglas M. Larson,2012. News and Social Cost: The Case of Oil Spills and Distant Viewers. *Journal of Benefit-Cost Analysis*,3(4).

Shuyi Yang, Xueyi You,2016. Determination of the oil spill removal area by oil particle tracking in a harbor. *Oceanological and Hydrobiological Studies*,45(2).

Skiba Y N, Parra-Guevara D, 2017. Application of Adjoint Approach to Oil Spill Problems. *Environmental Modeling & Assessment*, 22(4):379-395.

Song Gao, Juan Huang, Tao Bai, Yajing Cao, Jiangling Xu,2014. Study on the ensemble forecast of oil spilled sources. *Marine Sciences*, 38(3): 42-45.

Takeshita K, Tanikawa K, Kaji K, 2017. Applicability of a Bayesian state-space model for evaluating the effects of localized culling on subsequent density changes: sika deer as a case study. *European Journal of Wildlife Research*, 63(4):71.

Yajing Cao, Song Gao, Yi Ding, Tao Bai, 2016. Study on Tracing Model of the Oil Spill Sources in Bohai Sea. *Journal of Guangxi Academy of Sciences*, 32(2): 83-87.

Yanfang Li, Yu Wang, Jihong Li, 2013. The Replicability of Several Cross-validated Tests. *Journal of Taiyuan Normal University (Natural Science Edition)*, 12(4): 46-49.

Yang Y, Li Y, Liu G, et al, 2017. A hindcast of the Bohai Bay oil spill during June to August 2011. *Acta Oceanologica Sinica*.

Yunlong Guo, Yongtao Xi, Shenping Hu, Ling Xu, 2013. Dynamic Bayesian Network-based Prediction of Ship Oil Spill Risk. *China Safety Science Journal*, 23(11): 53-59.

Zhe Ren, Huailiang Chen, Lianxi Wang, Yin Li, Qi Li, 2015. Research on inversion model of wheat LAI using cross-validation. *Remote Sensing for Land & Resources*, 27(4): 34-40.

Zhou Y, Wang Z, Jin J, et al, 2017. Uncertainty analysis of designed flood on Bayesian MCMC algorithm: a case study of the Panjiakou Reservoir in China. *Environmental Earth Sciences*, 76(23):788.