Full Length Article

An efficient classification method for fuel and crude oil types based on m/z 256 mass chromatography by COW-PCA-LDA

Peiyan Sun[a,b,*], Kaiwen Bao[c], Haoshuai Li[d,e], Fujuan Li[a,b], Xinping Wang[a,b], Lixin Cao[a,b], Guangmei Li[a,b], Qing Zhou[a,b], Hongxia Tang[a,b], Mutai Bao[d,e,*]

[a] Key Laboratory of Marine Spill Oil Identification and Damage Assessment Technology, State Oceanic Administration, Qingdao 266033, China
[b] North China Sea Environmental Monitoring Center, State Oceanic Administration, Qingdao 266033, China
[c] Qingdao No. 2 Middle School of Shandong Province, Qingdao 266061, China
[d] Key Laboratory of Marine Chemistry Theory and Technology, Ministry of Education, Ocean University of China, Qingdao 266100, China
[e] College of Chemistry & Chemical Engineering, Ocean University of China, Qingdao 266100, China

ARTICLE INFO

ABSTRACT

Chemometric analysis was used to classify different oils based on the m/z 256 mass chromatogram. Three hundred oil samples comprised of 12 LFOs, 104 HFOs (13 weathered fuel oils), and 184 crude oil samples (63 weathered crude oils) were analyzed by GC–MS, and the m/z 256 mass chromatogram was chosen for classification. After normalization, the m/z 256 mass chromatograms were aligned using the correlation optimized warping (COW) method with segment lengths of 50 data points and slack parameters of 4 data points. They were then analyzed by principal component analysis (PCA). The score graphs of the PCs revealed that there was a good discriminant for HFOs, LHOs and oils. According to the changing tendency of the correct sample percentage, five PCs were chosen for the linear discriminant analysis (LDA). A good model with three discriminant equations for HFOs, LFOs and oils was obtained. For the training set with 172 samples, the correct percentage reached 100%, and 99% was obtained for the test set with 128 samples. This study proved that the COW-PCA-LDA method based on the m/z 256 mass chromatogram can be used in oil fingerprinting identification, especially for oil type classification.

1. Introduction

With the increase in sea transportation and the expansion of offshore crude oil exploitation and development, oil spills often occur, and those spills that are not quickly identified will be classified as "no source". For this type of oil spill, finding the true source is a crucial step in the oil spill investigation and subsequent emergency response. Screening for the types of oil present as soon as possible is helpful for identification.

Oil fingerprinting technology is a key oil spill screening and identification method, and gas chromatographic–mass spectrometry (GC–MS) has been widely used to analyze the fingerprint of oils, especially alkanes, steranes, terpane (biomarkers), and polycyclic aromatic hydrocarbons (PAHs). This method provides rich information reflecting oil fingerprinting features and sources [1,2]. Even so, fingerprinting is highly complex due to variations in the oil feedstock, refining processes and weathering, making visual comparisons alone difficult and ineffective for classifying different types of oils, especially for large quantities of oil samples.

In our last paper, we studied the sensitive indicators of fuels and crude oil and found that polycyclic aromatic hydrocarbons (PAHs) are sensitive to fuels, especially to heavy fuels [3]. In comparison, because the five major characteristic PAHs of oils have larger molecular weights and are less affected by weathering, the chrysenes series distribution may be better able to characterize the fuels. Moreover, from the visual observation spectrogram, there is a greater difference in C2 chrysenes mass chromatograms between certain fuels and crude oils. Thus, in this paper, the m/z 256 mass chromatogram, which mainly reflects C2 chrysenes, was selected to find a quick and effective identification method for oil types.

This paper will use the m/z 256 mass chromatograms of 300 samples analyzed at different times. Due to instrumental instability over time, the chromatograms between oil samples may change, especially with regard to the retention time shifting from one chromatogram to another. For data consistency, such retention time shifts should be corrected prior to a chemometric analysis. Many alignment techniques

---

* Corresponding authors at: Key Laboratory of Marine Spill Oil Identification and Damage Assessment Technology, State Oceanic Administration, Qingdao 266033, China (P. Sun). Key Laboratory of Marine Chemistry Theory and Technology, Ministry of Education, Ocean University of China, Qingdao 266100, China (M. Bao).
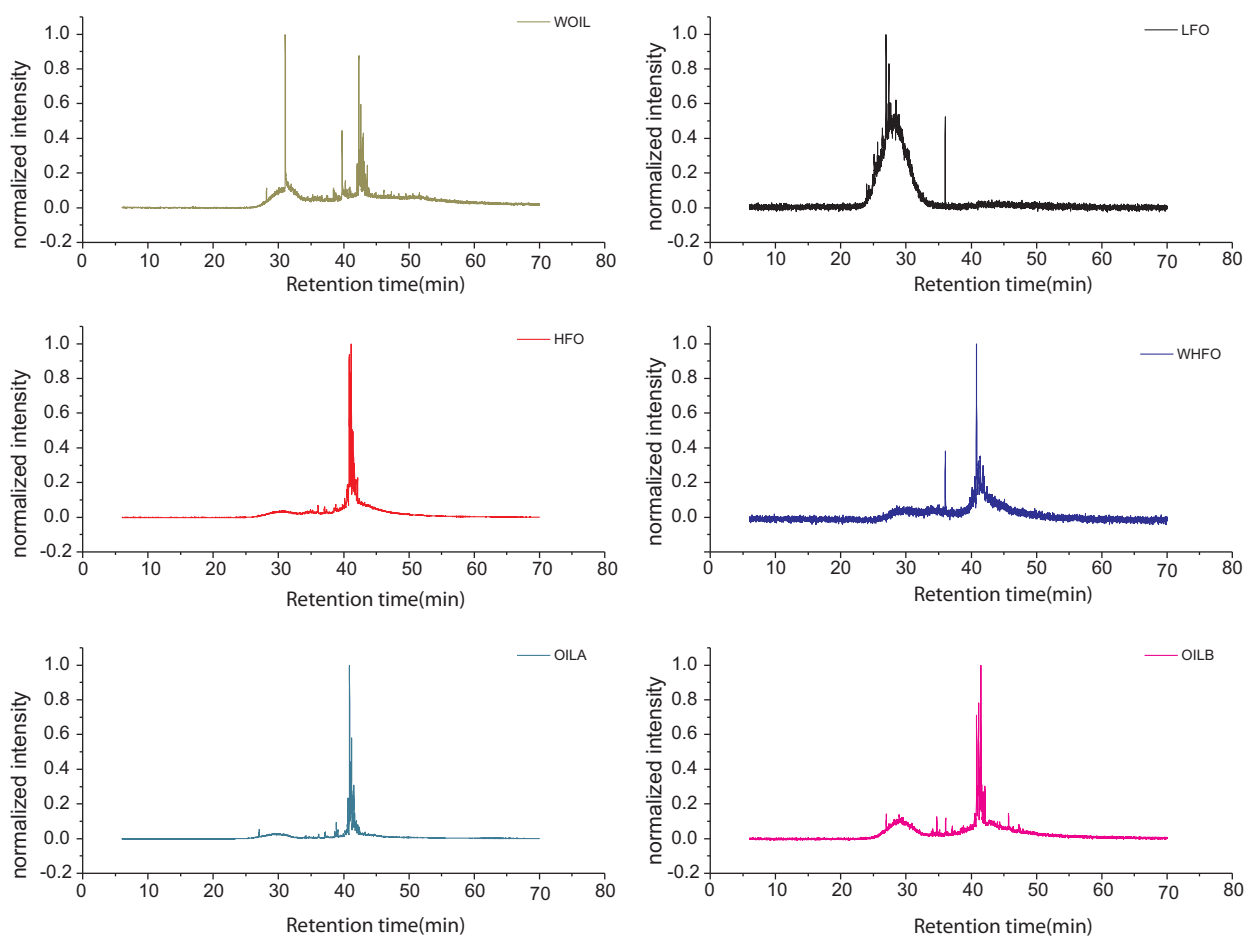*E-mail addresses:* sunpeiyan@bhfj.gov.cn (P. Sun), mtbao@ouc.edu.cn (M. Bao).

Fig. 1. *m/z* 256 mass chromatograms of different oil samples (top: full view; bottom: partial enlarged drawing; WOIL and WHFO refer to the weathered oil and weathered HFO, respectively).

have been used for time alignment, including correlation optimized warping (COW), which is widely used by chemists.

The *m/z* 256 mass chromatograms also contains noise and some less important signals, which will affect discrimination. We therefore sought to extract the most beneficial information for the classification of oil types, and principal component analysis (PCA) is a good method for this.

Linear discriminant analysis (LDA) is a classification tool that creates sample projections of high-dimensional patterns to best identify vector space, extract the classification information and compress the feature space dimension, which can guarantee that the samples in the new subspace projection have maximal differences between the groups and minimal differences within the groups. The combination of PCA and LDA is an effective classification tool.

Some studies have also shown the usefulness of the application of chemometric tools for oil classification and identification, such as determining vegetable oil and fat adulterants in diesel oil by high-performance liquid chromatography and multivariate methods [4], discovering the origin and type of diesel oil using physicochemical properties and multivariate analysis [5], determining gasoline origin by distillation curves and multivariate analysis [6], improving oil classification quality from oil spill fingerprint beyond the six sigma approach [7], and chemometric techniques in oil classification from oil spill fingerprinting [8]. These techniques can be used to find effective characteristic factors and to implement fast classification, which has been performed in the past. A new method was developed for the chemical fingerprinting of petroleum biomarkers consisting of a GC–MS analysis, preprocessing GC–MS chromatograms, and PCA of the selected regions; this method was applied to the chromatograms of *m/z* 217

(tricyclic and tetracyclic steranes) of oil spill samples and source oils [9]. A framework for the use of chemometric approaches was reviewed in tiered oil spill fingerprinting and clear-cut practical details and experiences that can be used by forensic chemists [10]. The practicality of classifying the sources and types of crude oils from different oil fields in several countries and regions with terahertz time-domain spectroscopy (THz-TDS) and multivariate statistical methods, including cluster analysis (CA) and principal component analysis, was evaluated and used to build models [11]. The principal component analysis of the elemental composition data obtained by high resolution FT-ICR mass spectrometry was used to correctly identify the source of environmental contamination caused by the unintended release of heavy fuel oil (HFO) [12]. Gas chromatography–mass spectrometry (GC–MS) time warping and principal component analysis were applied to assess oil weathering [13]. HPLC–MS and linear discriminant analysis were used to determine the cultivar and aging of Sicilian olive oils [14]. A two-dimensional (2-D), correlation-optimized warping algorithm was developed to align 2-D gas chromatography coupled with time-of-flight mass spectrometry (GC-GC/TOF-MS) data [15]. A new alignment procedure was presented for chromatographic signals with bilinear decomposition based on the alignment of chromatographic profiles [16]. PCA and LDA were used to establish the origin of gasoline samples with three different grades based on chromatographic data [17]. GG–MS and multivariate pattern recognition techniques were presented for petroleum-based accelerant detection and classification, and PCA was used to accelerate classification. A soft, independent model classification analogy model was then developed to evaluate the evaporation, sample size, and sample charring [18]. The extraction of the target information contained in the GC GC–QMS data was evaluated for discriminating