Hindawi
Advances in Materials Science and Engineering
Volume 2020, Article ID 2025072, 13 pages
https://doi.org/10.1155/2020/2025072

Research Article

Crude Oil Source Identification of Asphalt via ATR-FTIR Approach Combined with Multivariate Statistical Analysis

Ruibo Ren,[1] Wenmiao Fan,[1] Pinhui Zhao ,[1] Hao Zhou,[1] Weikun Meng,[1] and Ping Ji[2]

[1]*Shandong Provincial Key Laboratory of Road and Traffic Engineering in Colleges and Universities, School of Transportation Engineering, Shandong Jianzhu University, Jinan 250101, China*
[2]*Shandong Hi-Speed Engineering Testing Co., Ltd., Jinan 250101, China*

Correspondence should be addressed to Pinhui Zhao; zhaopinhui08@163.com

Received 14 February 2020; Revised 15 May 2020; Accepted 3 June 2020; Published 6 July 2020

Guest Editor: Meng Guo

The types of crude oil for producing asphalt have a decisive influence on various performance measures (including aging resistance and durability) of asphalt. To discriminate and predict the crude oil source of different asphalt samples, a discrimination model was established using 12 greatly different infrared (IR) characteristic absorption peaks (CAPs) as predictive variables. The model was established based on diverse fingerprint recognition technologies (such as principal component analysis (PCA) and multivariate logistic regression analysis) by using attenuated total reflectance-Fourier transform infrared spectroscopy (ATR-FTIR). In this way, the crude oil source of different asphalt samples can be effectively discriminated. At first, by using PCA, the 12 CAPs in the IR spectra of asphalt samples were subjected to dimension reduction processing to control the variables of key factors. Moreover, the scores of various principal components in asphalt samples were calculated. Afterwards, the scores of principal components were analysed through modelling based on multivariate logistic regression analysis to discriminate and predict the crude oil source of different asphalt samples. The result showed that the logistic regression model shows a favourable goodness of fit, with the prediction accuracy reaching 93.9% for the crude oil source of asphalt samples. The method exhibits some outstanding advantages (including ease of operation and high accuracy), which is important when controlling the source and quality and improving the performance of asphalt.

1. Introduction

Asphalt pavements are widely used: as a black binding material produced from oil, asphalt is widely used as the binder in asphaltic mixtures [1–3]. Due to the differences in origins and production modes of crude oil for producing asphalt, the properties of crude oil exert important influences on the performance of asphalt mixtures, which also lead to significant differences in the performance of the various asphalt produced therewith [4–8].

The conventional performance of the same grade of asphalt is very similar; however, different asphalt exhibit large differences in various aspects, including high- and low-temperature performance, durability, and fatigue properties, which are considered as external expressions of chemical composition, molecular structure, and transformation of asphalt [9–11]. Furthermore, the study shows that the differences in the composition and structure of asphalt mainly depend on the source of crude oil and refining process of asphalt production. Due to the differences in the geological structure, oil generation conditions, and age, the nature and composition of crude oil in different regions are very different. However, crude oil with similar properties and composition in the same region has similar processing, storage, and transportation options. At the same time, most of the petroleum asphalt is produced by distillation currently, and the molecules in the asphalt retain their original state in the crude oil. Therefore, most of the composition and structure of asphalt are inherited from crude oil; that is to say, the structural performance of asphalt mainly depends on the source of crude oil. Because the asphalt is produced by different types of crude oil, the physical and chemical composition information about asphalt is unique. Just like the human fingerprint information, these components which can express the unique structure of asphalt can be called the "oil fingerprint" of asphalt. It is because of the uniqueness of "oil fingerprint" information of asphalt that it is feasible to discriminate the oil fingerprints of asphalt from different crude oil sources [12–16].

At present, as the composition and structure of asphalt are extremely complex, the characterization of its structure requires more high-resolution and high-throughput analysis means and equipment, so there are few reports on the identification and analysis of asphalt oil fingerprints [17]. However, the identification and analysis of marine oil spill fingerprints has always been an issue of widespread concern. Similar to the method and purpose of identifying "oil fingerprints" of oil spills at sea, the purpose of recognising oil fingerprints of asphalt is to attain oil fingerprint information of asphalt through different methods such as physical, chemical, and biological methods [18]. Moreover, by applying multivariate statistical methods (including principal component analysis (PCA) and regression analysis), the chemical composition variables of oil fingerprints are summarised, classified, and discriminated [19, 20]. On this basis, qualitative and quantitative relationships between data are obtained to distinguish the crude oil source of asphalt, thus effectively controlling their qualities. Meanwhile, some testing methods used in the "oil fingerprint" identification of marine oil spills have been successfully used to analyse the composition and structure of asphalt [21–23]. For example, a gas chromatograph-mass spectrometer (GC-MS) was used to explore the chemical compositions of smoke released by asphalt materials during heating [24, 25]. Gel permeation chromatography (GPC) and thin-layer chromatography (TLC) were used to measure the molecular weights and the composition distributions of asphalt [26–28]. Nuclear magnetic resonance (NMR) and Fourier transform infrared spectroscopy (FTIR) were used to investigate the compositions, structures, and functional groups of asphalt [29, 30]. In all analytical techniques, compared with other methods (including GC-MS and NMR), which generally show some disadvantages (including high cost, damage to samples, and being laborious and time consuming during analysis), infrared (IR) spectroscopy is the most widely used technique in investigating asphalt materials. The reason is that IR spectroscopy shows many outstanding advantages, including being label-free, rapid, nondestructive, and low-cost, with simple sample preparation [31–33]; however, in the above analysis, the chemical structures of asphalt are qualitatively analysed, mainly aiming at those of a certain or multiple specific asphalt samples while lacking quantitative research into the types of asphalt. The research into discrimination of the types of asphalt, tracing of the production area, and quality control of asphalt has not yet been reported.

Therefore, by utilising attenuated total reflectance-Fourier transform infrared spectroscopy (ATR-FTIR), the characteristic functional groups of asphalt from different crude oil source were discriminated and quantitatively analysed. Based on multivariate statistics, PCA and logistic regression analysis were conducted on IR spectral data to establish a discriminant function. An accurate, nondestructive, stable method of discriminating the crude oil source of asphalt samples was explored, which provides a scientific basis for realising reasonable selection, supervision quality, and guaranteed origins of asphalt.

2. Experimental Raw Materials and Methods

2.1. Experimental Materials. During the experiment, 33 asphalt samples were purchased from factories in China for producing asphalt. Before being applied, the asphalt samples were sealed in original oxygen-free containers at 5°C to prevent the samples from being oxidised. Additionally, all asphalt samples were unprocessed before use. As mentioned in Section 1, the differences in the "oil fingerprint" of asphalt are determined by the crude oil from which it is produced. Due to the same geological structure, oil generation conditions, and age in the same region, the composition and chemical structure of crude oil are also very similar. Therefore, the "oil fingerprints" of asphalt produced by crude oil from the same region are very similar, such as crude oil from the Middle East Gulf region, including Saudi Arabia, Iran, Kuwait, Iraq, and United Arab Emirates, crude oil from South America, including Marry, Poscan, Maya, and Castilla, and crude oil from the Bohai Rim region of China, such as Bohai Bay, Huanxiling, and Caofeidian. The crude oil of 33 asphalt samples came from the above three regions. According to the names of the three regions, the crude oil source of asphalt is divided into three categories: Middle East, South America, and the Bohai Rim region of China. The basic performance measures (penetration ratio (ASTM D5), ductility ratio (ASTM D113), and softening point (ASTM D36)) of asphalt and the crude oil source of asphalt are listed in Table 1. It is worth noting that the last digit of the asphalt number listed in Table 1 represents different sampling batches of the same asphalt.

2.2. FTIR Analysis. Through ATR-FTIR (using a Cary 630 FTIR microscope), the IR spectra of asphalt samples were explored. Within the range of 400–4,000 cm^{-1}, 64 scans were conducted, each at a resolution of 1 cm^{-1}. The samples were placed on the horizontal ATR crystal made of zinc selenide, being subjected to multiple reflections. After each operation, the ATR crystal was cleaned using acetone.

The original spectrum data were first subjected to baseline correction by applying the OMNIC software to eliminate baseline effects. Afterwards, based on the standardised variation diagram of preprocessed spectrum data, the difference in masses of different samples was eliminated.

2.3. Multivariate Statistical Analysis. Through the combination of principal component analysis (PCA) and multiple logistic regression analysis, the infrared spectrum data are analysed to establish the discrimination model of the crude oil source of asphalt. Logistic regression analysis is a multivariate analysis method to analyse and predict attribute-dependent variables based on single or multiple continuous or attribute-independent variables. Furthermore, each

Table 1: Basic properties and crude oil source of asphalt.

Asphalt number	Asphalt name[1]	Penetration (0.1 mm)	Ductility (cm, 10°C)	Softening point (°C)	Source	Category
1	MM-1	79.6	65.7	47.3	Middle East (Saudi Arabia)	1
2	Q-1	75.4	>100	46.5	South America (Marry & Poscan)	2
3	SK (BY)-1	69.7	52.2	46.9	Middle East (Saudi Arabia)	1
4	Q-2	61.5	39.0	48.5	South America (Marry & Poscan)	2
5	SL-1	63.4	47.5	47.0	Middle East (Saudi Arabia)	1
6	CMR-1	67.5	>100	47.2	South America (Marry)	2
7	LH-1	76.6	>100	45.6	South America (Marry & Poscan)	2
8	QPK-1	61.8	>100	46.8	Middle East (Iran)	1
9	MM-2	61.1	16.0	49.6	Middle East (Saudi Arabia)	1
10	QP-1	68.9	33.2	48.7	Middle East (Kuwait)	1
11	JB-1	67.3	74.0	47.5	South America (Marry & Poscan)	2
12	SK (XY)-1	71.5	47.9	47.2	Middle east (Saudi Arabia)	1
13	ZH-1	67.4	13.5	48.1	South America (Maya & Castilla)	2
14	JL-1	64.2	35.6	48.2	Middle East (Kuwait)	1
15	HR-1	63.2	85.2	48.6	South America (Marry & Poscan)	2
16	AS-1	62.1	84.3	47.4	Middle East (Saudi Arabia)	1
17	XT-1	62.7	23.9	48.8	Middle East (Saudi Arabia)	1
18	ZH-2	64.0	>100	50.1	South America (Maya& Castilla)	2
19	LH-2	69.0	>100	47.0	South America (Marry & Poscan)	2
20	QL-1	64.0	21.3	50.0	Middle east (Kuwait)	1
21	KL-1	79.8	>100	47.2	Bohai Rim region of China (Bohai Bay)	3
22	KL-2	61.0	57.8	49.5	Bohai Rim region of China (Bohai Bay)	3
23	SL-2	65.0	30.0	49.2	Middle East (Saudi Arabia)	1
24	SK (NB)-1	71.0	47.9	48.4	Middle East (Saudi Arabia)	1
25	Q-3	75.0	>100	47.3	South America (Marry & Poscan)	2
26	DSZ-1	69.0	>100	50.0	Bohai Rim region of China (Huanxiling)	3
27	CMR-2	68.2	>100	47.9	South America (Marry)	2
28	JB-2	64.6	74.0	46.3	South America (Marry & Poscan)	2
29	ZH-3	63.5	>100	51.2	South America (Maya & Castilla)	2
30	LH-3	68.7	>100	47.9	South America (Marry & Poscan)	2
31	DSZ-2	78.4	>100	48.6	Bohai Rim region of China (Huanxiling)	3
32	SZ-1	61.9	57.8	49.3	Bohai Rim region of China (Bohai Bay)	3
33	SZ-2	68.0	>100	50.9	Bohai Rim region of China (Bohai Bay)	3

[1] In this column, 1, 2, and 3 indicate the different sampling batches of the same asphalt.

variable is required to be independent of each other in variable screening and parameter estimation. In many studies, there is a certain degree of linear dependence between their variables, which is called multicollinearity. This multiple collinear relationship may increase the mean square error and standard error of the estimated parameters, which leads to the instability of the analysis results of the logistic regression model. The main reason for the problem of multicollinearity is the overlap of information. However, PCA can reduce the repeatability of information and achieve the purpose of eliminating multicollinearity by extracting independent principal components from explanatory variables.

For this reason, this study used a multinomial logistic regression model based on PCA to improve the discrimination accuracy of the model. First of all, the PCA was used to reduce the dimension of the CAPs variables of the infrared spectrum, so that the variables with strong correlation were integrated into the same principal components. The principal components were independent of each other; thus, the multiple collinear relationship between variables was eliminated. Then, by using these principal components as independent variables, the discriminant model of crude oil source of asphalt was obtained by logistic regression analysis.

2.3.1. PCA Analysis. PCA refers to a simplification of multidimensional data to several relevant variables (principal components) through a dimension reduction approach. Each principal component reflects most of the information of original variables, and the contained information is not repeated. PCA can compress countless information and simplify complex problems [34]. The modelling process of PCA is as follows:

(1) Calculation of the correlation coefficient matrix:

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}, \quad (1)$$

where $r_{ij}\,(i, j = 1, 2, \ldots, p)$ refers to the correlation coefficient of original variables X_i and X_j, $r_{ij} = r_{ji}$, which can be calculated by using the following formula:

$$r_{ij} = \frac{\sum_{k=1}^{n} \left(X_{ki} - \overline{X}_i\right)\left(X_{kj} - \overline{X}_j\right)}{\sqrt{\sum_{k=1}^{n} \left(X_{ki} - \overline{X}_i\right)^2 \sum_{k=1}^{n} \left(X_{kj} - \overline{X}_j\right)^2}}. \quad (2)$$

(2) Calculating eigenvalues and eigenvectors:

The characteristic equation $|\lambda I - R| = 0$ was solved. Generally, the eigenvalues were calculated by using the Jacobi method and, in descending order are $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. The eigenvectors $e_i\,(i = 1, 2, \ldots, p)$ corresponding to eigenvalue λ_1 were separately calculated, satisfying $\|e_i\| = 1$, that is, $\sum_{j=1}^{p} e_{\mathrm{ij}}^2 = 1$, where e_{ij} denotes the j^{th} component of vector e_i.

(3) Calculating contribution and cumulative contribution of principal components:

$$\text{contribution}: \frac{\lambda_i}{\sum_{k=1}^{p} \lambda_k}, \quad (i = 1, 2, \ldots, p), \quad (3)$$

$$\text{cumulative contribution}: \frac{\sum_{\mathrm{k}=1}^{i} \lambda_k}{\sum_{k=1}^{p} \lambda_k}, \quad (i = 1, 2, \ldots, p). \quad (4)$$

In general, the eigenvalues with the cumulative contribution not lower than 70% are taken. $\lambda_1, \lambda_2, \ldots, \lambda_m$ are the corresponding first, second, . . ., m^{th} $(m \leq p)$ principal components.

(4) Calculating the loads of principal components:

$$l_{ij} = p\left(Z_i, X_j\right) = \sqrt{\lambda_i e_{ij}}. \quad (i = 1, 2, \ldots, p). \quad (5)$$

(5) Scores of various principal components:

$$Z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ \vdots & \vdots & & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nm} \end{bmatrix}. \quad (6)$$

2.3.2. Logistic Regression Analysis. Logistic regression is a multivariate analysis method for investigating the relationship between binominal or multinomial observation results (dependent variable) and influencing factors (independent variable), belonging to probabilistic nonlinear regression methods. The logistic regression when the dependent variable only shows two or more states belongs to binomial logistic regression and multinomial logistic regression, respectively [35, 36]. For discriminating and classifying the crude oil of asphalt, multinomial logistic regression is applied to conduct data analysis, owing to the crude oil of asphalt being sourced from the Bohai Rim region of China, South America, and the Middle East.

(1) Model fitting:

For multinomial logistic regression, a certain level of dependent variables is defined as the reference level herein. Compared with the other levels, i-1 (i refers to the number of dependent variables) generalised logistic regression models were fitted. By taking three-level dependent variables as an example, it is supposed that the values of dependent variables are 1, 2, and 3: the probabilities corresponding to the values are π_1, π_2, and π_3, respectively. Based on m-independent variables, two models are fitted as follows:

$$\operatorname{logit}\frac{\pi_1}{\pi_3} = \alpha_1 + \beta_{11}X_1 + \beta_{12}X_2 + \ldots + \beta_{1m}X_m, \quad (7)$$

$$\operatorname{logit}\frac{\pi_2}{\pi_3} = \alpha_2 + \beta_{21}X_1 + \beta_{22}X_2 + \ldots + \beta_{2m}X_m. \quad (8)$$

(2) Meaning of regression parameters:

For multinomial logistic regression, each independent variable contains $(m - 1)$ parameters. The parameter β_{1m} represents an independent variable x_m that changes one unit on the premise that other independent variables remain unchanged, and it reflects the variation of the log-odds ratio (OR) of class i. The OR is subjected to logarithmic transformation to obtain the linear mode $(\ln(p_i/1 - p_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)$ of the logistic regression model.

3. Results and Discussion

3.1. Establishment of Discrimination Indices for Crude Oil Source of Asphalt. FTIR is an important means of identifying organic compounds. When irradiating organics using the IR light, the molecules absorb the IR light leading to vibrational energy level transition, and different chemical bonds or functional groups show diverse absorption frequencies. The contents of various materials are reflected in their IR absorption spectra, which can be quantitatively analysed according to peak location and absorption intensity. The structural composition of asphalt is complex, and asphalt shows significant differences in behaviour. For these reasons, it fails to effectively characterize the difference of behaviours of asphalt from different crude oil only by quantitatively comparing the peak areas of IR spectrograms. Therefore, by observing the shapes and locations of IR spectrograms, 12 significant characteristic absorption peaks (CAPs) were selected to analyse the transmittances of absorption peaks.

The IR absorption spectra of 33 asphalt samples are similar. By using the mean value method, the mutual mode of the IR spectrogram of all asphalt samples was constructed (Figure 1): the assignments of 12 characteristics peaks are as follows: the strong absorption peaks around 2850 cm^{-1} and 2920 cm^{-1} are triggered by the stretching vibration of CH_2, and a very weak absorption peak around 1700 cm^{-1} is induced by the stretching vibration of C=O. Moreover, the vibration of the benzene ring leads to the absorption peak in the vicinity of 1600 cm^{-1}, and the absorption peaks at 1380 cm^{-1} and 1460 cm^{-1} are caused by the bending vibration of CH_3. The fingerprint region appears below 1300 cm^{-1}, in which the absorption peaks at 1166 cm^{-1} and 1032 cm^{-1} are triggered by the stretching vibrations of C=S and S=O, respectively. The stretching vibration of CH results in a weak absorption peak around 969 cm^{-1}, while the absorption peaks at 872 cm^{-1} and 812 cm^{-1} are induced by vibrations of an isolated hydrogen and two adjacent hydrogen atoms on the benzene ring, respectively. Additionally, the absorption peak at 723 cm^{-1} is also caused by the stretching vibration of CH_2.

3.2. Analysis of Predictive Variables Based on Descriptive Statistics. The IR spectra of all asphalt samples are similar, and it is difficult to distinguish the differences among asphalt samples by comparing spectrograms alone. Hence, 12 significantly different CAPs were selected from the spectrograms to describe the transmittances of absorption peaks based on descriptive statistics. From two aspects of centralised location (including indices such as average and median) and degree of dispersion (including indices such as extreme value), the samples are described so as to reflect spectrographic data (Table 2).

In Table 2, according to the analysis result of descriptive statistics on the transmittances of 12 CAPs, it can be seen that the asphalt produced by crude oil from the Bohai Rim region of China showed a larger transmittance. By contrast, the transmittances of asphalt produced by crude oil from the Middle East and South America were consistently low. However, it is impossible to distinguish the oil source of asphalt based on the descriptive statistics of infrared spectral transmittance of asphalt. Therefore, it is necessary to introduce multivariate statistical analysis methods, such as multinomial logistic regression analysis based on PCA described in Section 2.3.

3.3. Correlation Analysis of Predictive Variables. Correlation analysis aims to explore the correlation among multiple variables, which is also an important parameter for evaluating the fingerprint variables of asphalt [37]. In order to further evaluate whether the selected 12 variables were of sufficient significance to the prediction model, a correlation analysis of the 12 CAP variables was required. Generally, correlation analysis is conducted by applying Pearson and Spearman correlation coefficients. The Pearson correlation coefficient is generally applicable to data satisfying a normal distribution, and the Spearman correlation coefficient is employed for data that do not satisfy a normal distribution. Therefore, before the correlation test, it is necessary to test the normal distribution of 12 variables to determine the appropriate correlation test method.

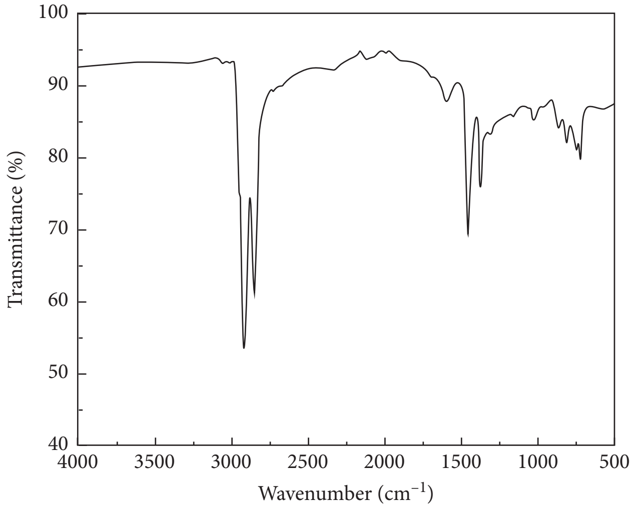

FIGURE 1: The common mode of infrared fingerprints of the asphalt.

By using the skewness-kurtosis test method, whether the transmittances of the 12 CAPs of 33 asphalt samples conform to a normal distribution was assessed, and through the K-S test as an auxiliary analysis method, the accuracy of the test results was ensured [38, 39]. The 12 variables were processed by importing them into SPSS19 (Tables 3 and 4).

It can be seen from Table 3 that the values of skewness and kurtosis of transmittances of the 12 CAPs of all asphalt samples produced by three origins of oil fluctuate within a certain small positive and negative range around zero. It can be further seen from Table 4 that the asymptotic significances of the 12 variables all exceed 0.05. Moreover, based on the result of the skewness-kurtosis test, it can be considered that the 12 variables of 33 asphalt samples all conform to a normal distribution, which provides a basis for determining the method for testing correlation among variables. Therefore, the Pearson correlation coefficient is used to analyse the correlation between variables (Table 5).

As shown in Table 5, the IR CAP at 2850 cm^{-1} showed a significant correlation with those at 2920, 1460, and 723 cm^{-1}, respectively. Additionally, there are significant correlations between each IR CAP at 1700, 1600, 1460, 1380, 1166, 1032, 969, 872, 812, and 723 cm^{-1}. Moreover, multiple CAPs exhibited a high correlation. The aforementioned CAPs with high correlation covered all 12 CAPs. This showed that the 12 selected CAPs contained most of the fingerprint information about the asphalt, thus providing a basis for selecting variables capable of discriminating the different crude oil sources of asphalt.

Table 2: Descriptive statistics analysis of the 12 CAPs.

Statistical indicators	Oil source	Average value	Median value	Maximum value	Minimum value
2920^{-1}	Middle East	52.10	51.62	55.46	49.83
	South America	55.85	54.20	61.02	53.29
	Bohai Rim region of China	54.39	53.29	56.64	53.23
2850^{-1}	Middle East	59.82	59.51	63.60	58.43
	South America	63.26	62.39	67.38	60.74
	Bohai Rim region of China	62.97	62.36	64.88	61.69
1700^{-1}	Middle East	91.07	91.14	91.96	89.28
	South America	91.46	91.50	92.29	90.83
	Bohai Rim region of China	92.92	92.41	94.06	92.29
1600^{-1}	Middle East	87.31	87.37	88.25	85.10
	South America	87.67	87.49	88.74	87.20
	Bohai Rim region of China	90.66	90.26	91.63	90.08
1460^{-1}	Middle East	68.71	68.64	70.10	67.64
	South America	69.81	69.81	71.68	68.27
	Bohai Rim region of China	72.03	72.18	72.24	71.67
1380^{-1}	Middle East	75.51	75.48	76.95	74.08
	South America	76.11	76.12	77.26	74.77
	Bohai Rim region of China	78.51	78.50	78.96	78.07
1166^{-1}	Middle East	85.26	85.41	86.33	83.43
	South America	85.88	85.91	86.84	84.65
	Bohai Rim region of China	89.48	89.10	90.25	89.08
1032^{-1}	Middle East	84.37	84.49	86.35	82.29
	South America	85.78	86.30	86.80	82.57
	Bohai Rim region of China	90.10	89.81	90.68	89.81
969^{-1}	Middle East	86.53	86.72	87.73	84.97
	South America	87.28	87.22	88.10	85.98
	Bohai Rim region of China	90.33	90.15	90.69	90.14
872^{-1}	Middle East	83.74	83.93	84.64	81.00
	South America	84.86	84.67	86.09	83.79
	Bohai Rim region of China	88.97	88.53	89.86	88.50
812^{-1}	Middle East	80.73	81.12	81.85	77.23
	South America	82.70	82.79	83.88	80.40
	Bohai Rim region of China	88.63	88.24	89.43	88.21
723^{-1}	Middle East	78.18	78.14	80.74	76.17
	South America	81.32	81.38	82.70	78.25
	Bohai Rim region of China	85.61	84.90	87.15	84.78

Table 3: The skewness-kurtosis test of the 12 CAPs.

Statistical indicators (cm^{-1})	Skewness	Kurtosis
2920	1.123	1.279
2850	0.797	0.130
1700	0.713	1.241
1600	1.120	1.379
1460	0.509	−1.149
1380	0.509	−0.728
1166	1.129	0.600
1032	0.737	−0.213
969	0.926	0.074
872	1.052	0.735
812	0.975	0.245
723	0.651	0.459

3.4. Establishment of Logistic Regression and Discriminant Model Based on PCA

3.4.1. PCA on All Variables. According to the correlation analysis of variables, it can be seen that the information contained in the 12 CAPs shows a certain repeatability. PCA not only can remove repeated information but can retain key information, thus realising dimension reduction. Furthermore, it makes the modelling for logistic regression and discrimination more reliable due to reducing the disturbance caused by accidental factors.

The transmittances of the 12 CAPs of 33 asphalt samples are input into the SPSS19 software for PCA. The results are displayed in Table 6 and Figure 2. As shown in Table 6, there

Table 4: The Kolmogorov–Smirnov test of a single sample.

Statistical indicators	2920 cm^{-1}	2850 cm^{-1}	1700 cm^{-1}	1600 cm^{-1}	1460 cm^{-1}	1380 cm^{-1}	1166 cm^{-1}	1032 cm^{-1}	969 cm^{-1}	872 cm^{-1}	812 cm^{-1}	723 cm^{-1}
Number of samples	33	33	33	33	33	33	33	33	33	33	33	33
Kolmogorov–Smirnov	1.14	0.77	0.68	1.33	0.99	0.75	1.30	0.98	1.08	1.35	1.14	0.86
Asymptotically significant (bilateral)	0.15	0.59	0.73	0.06	0.27	0.63	0.07	0.28	0.19	0.05	0.14	0.43

Table 5: Pearson correlation coefficients of transmittances of the 12 CAPs.

Correlation coefficient (cm^{-1})	2920 cm^{-1}	2850 cm^{-1}	1700 cm^{-1}	1600 cm^{-1}	1460 cm^{-1}	1380 cm^{-1}	1166 cm^{-1}	1032 cm^{-1}	969 cm^{-1}	872 cm^{-1}	812 cm^{-1}	723 cm^{-1}
2920	1.00											
2850	0.94	1.00										
1700	0.15	0.35	1.00									
1600	0.22	0.41	0.91	1.00								
1460	0.49	0.61	0.71	0.84	1.00							
1380	0.30	0.40	0.72	0.85	0.96	1.00						
1166	0.13	0.32	0.89	0.96	0.84	0.88	1.00					
1032	0.07	0.28	0.85	0.88	0.75	0.79	0.97	1.00				
969	0.12	0.30	0.86	0.92	0.83	0.88	0.99	0.98	1.00			
872	0.25	0.46	0.89	0.97	0.86	0.87	0.98	0.94	0.96	1.00		
812	0.23	0.43	0.84	0.93	0.86	0.87	0.98	0.96	0.98	0.99	1.00	
723	0.42	0.63	0.79	0.86	0.83	0.77	0.90	0.91	0.90	0.94	0.95	1.00

Table 6: Result of principal component analysis.

Number of principal components	Eigenvalue	Variance contribution (%)	Cumulative variance contribution (%)
1	9.319	77.658	77.658
2	1.860	15.498	93.156
3	1.017	3.508	96.664
4	0.255	2.123	98.787
5	0.081	0.677	99.464
6	0.031	0.256	99.720
7	0.019	0.159	99.879
8	0.006	0.051	99.931
9	0.004	0.033	99.964
10	0.003	0.021	99.985
11	0.002	0.013	99.998
12	0.000	0.002	100.000

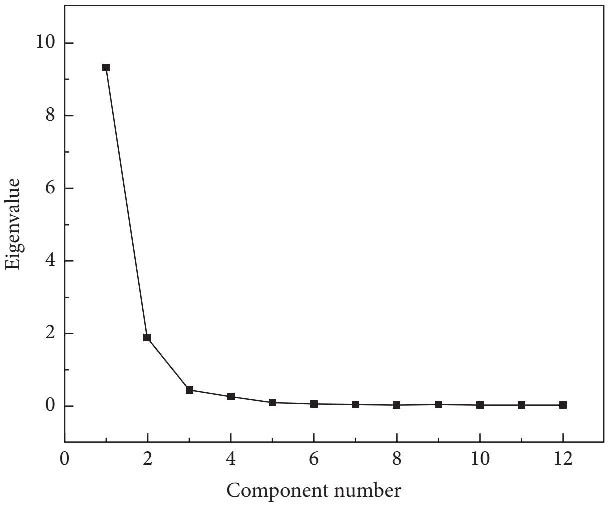

Figure 2: Analysis of the scree plot of principal components.

are three principal components whose eigenvalues exceed one. The first, second, and third principal components explain 77.658%, 15.498%, and 3.508% of the nature of the original variables, respectively. The cumulative variance contribution of the three principal components is 96.664% (research shows that there is a high explanation rate when the cumulative contribution is higher than 70%). It can be seen from the scree plot (Figure 2) that the broken lines of the first three principal components are steep while later tending to become shallower. This further indicates that it is appropriate to extract the three principal components (PCA1, PCA2, and PCA3). According to the correlation coefficients between the principal component and the original variables, the principal components Y_1, Y_2, and Y_3 are separately expressed as follows:

$$\begin{aligned} Y_1 = {} & 0.035x_1 + 0.055x_2 + 0.095x_3 + 0.103x_4 + 0.098x_5 + 0.097x_6 + 0.104x_7 + 0.100x_8 + 0.104x_9 \\ & + 0.106x_{10} + 0.106x_{11} + 0.102x_{12}, \end{aligned} \tag{9}$$

$$\begin{aligned} Y_2 = {} & 0.503x_1 + 0.449x_2 - 0.088x_3 - 0.056x_4 + 0.128x_5 - 0.001x_6 - 0.115x_7 - 0.138x_8 - 0.120x_9 \\ & - 0.042x_{10} - 0.049x_{11} + 0.069x_{12}, \end{aligned} \tag{10}$$

$$\begin{aligned} Y_3 = {} & -0.025x_1 + 0.393x_2 + 0.527x_3 + 0.049x_4 - 0.781x_5 - 0.997x_6 + 0.004x_7 + 0.305x_8 - 0.025x_9 \\ & + 0.129x_{10} + 0.069x_{11} + 0.471x_{12}, \end{aligned} \tag{11}$$

where $x_1, x_2, \ldots, x_{12}$ represent the transmittances of CAPs at 2920, 2850, 1700, 1600, 1460, 1380, 1166, 1032, 969, 872, 812, and 723 cm^{-1}, respectively.

3.4.2. The Process and Result of Multinomial Logistic Analysis. By substituting the transmittances of the 12 CAPs of 33 asphalt samples into formulae (9–11), the scores of the

Table 7: The scores of each principal component.

Asphalt number	Asphalt name[2]	PCA1	PCA2	PCA3
1	MM-1	−0.446	−0.840	−0.024
2	Q-1	−0.537	0.876	1.484
3	SK (BY)-1	−0.870	−0.225	−0.392
4	Q-2	−0.383	0.118	−1.233
5	SL-1	−0.986	0.240	−0.471
6	CMR-1	−1.796	0.963	−0.386
7	LH-1	−0.865	−1.151	0.438
8	QPK-1	−0.123	−0.995	−0.555
9	MM-2	−1.042	−0.455	−0.089
10	QP-1	−0.593	−1.212	−0.046
11	JB-1	−0.742	−1.037	−0.711
12	SK (XY)-1	−1.152	−0.335	0.338
13	ZH-1	−0.470	−0.919	−0.641
14	JL-1	−0.113	−0.717	−1.306
15	HR-1	−0.523	−0.757	0.423
16	AS-1	−0.316	−0.098	1.739
17	XT-1	−0.268	0.156	2.187
18	ZH-2	0.418	2.666	−0.827
19	LH-2	−0.434	2.149	−1.062
20	QL-1	−0.451	0.336	0.688
21	KL-1	0.175	−0.225	−0.545
22	KL-2	−0.153	0.221	0.312
23	SL-2	0.627	1.081	0.147
24	SK (NB)-1	−0.268	0.156	2.187
25	Q-3	0.418	2.666	−0.827
26	DSZ-1	0.175	−0.225	−0.545
27	CMR-2	−0.153	0.221	0.312
28	JB-2	1.536	−0.789	−0.679
29	ZH-3	1.626	−0.611	−1.122
30	LH-3	2.273	0.072	1.503
31	DSZ-2	1.536	−0.789	−0.679
32	SZ-1	1.626	−0.611	−1.122
33	SZ-2	2.273	0.072	1.503

[2] In this column, 1, 2, and 3 indicate the different sampling batches of the same asphalt.

three principal components can be calculated (Table 7). Moreover, the scores of the principal components are taken as factors, and three kinds of origins of asphalt are considered as dependent variables. Among them, the crude oil from the Bohai Rim region of China is regarded as a reference group to establish a multinomial logistic regression model based on principal components. On this basis, the parameters of the three principal components used for the logistic regression model are obtained.

Based on the parameter from regression, the logistic regression model can be obtained as follows:

$$\begin{aligned}
\text{logit}\,p_1 &= 17.130 - 40.988Y_1 + 5.758Y_2 - 1.788Y_3,\\
\text{logit}\,p_2 &= 22.543 - 21.163Y_1 + 9.509Y_2 + 0.024Y_3,\\
\text{logit}\,p_3 &= 0\,(\text{reference group}),
\end{aligned} \tag{12}$$

where $p_1, p_2,$ and p_3 refer to the probabilities of crude oil sources (the Middle East, South America, and Bohai Rim region of China) of asphalt and Y_1, Y_2, and Y_3 denote the first, second, and third principal components, respectively.

By substituting expressions (9), (10), and (11) into expression (12), the expression (formula (13)) for characterising the relationship between the logistic regression model and the 12 variables can be acquired. During discrimination and prediction, the probabilities of crude oil sources of asphalt can be separately acquired by substituting the transmittances of the 12 CAPs of the asphalt. The maximum probability corresponds to the predicted origin:

Table 8: The test of the logistic regression model.

Test of goodness of fit	Coefficient	Chi-square	Significance level
	Pearson	5.502	1
	Deviance	6.078	1
Test of pseudo R-squared	Cox and Snell	Nagelkerke	McFadden
	0.849	0.971	0.971
Test of likelihood ratio	Parameter	Chi-square	Significance level
	B	13.521	0.001
	F1	47.335	0.000
	F2	8.554	0.014
	F3	3.389	0.042

Table 9: The discrimination result of multinomial logistic regression.

Observed value	Predicted value			
	Middle East	South America	Bohai Rim region of China	Discrimination accuracy (%)
Middle East	14	1	0	93.3
South America	1	11	0	91.7
Bohai Rim region of China	0	0	6	100.0
Percentage (%)	45.5	36.4	18.2	93.9

$$\begin{aligned}
\operatorname{logit} p_1 &= 17.130 + 1.506x_1 - 0.372x_2 - 5.343x_3 - 4.632x_4 - 1.883x_5 - 2.199x_6 \\
&\quad - 4.932x_7 - 5.439x_8 - 4.909x_9 - 4.817x_{10} - 4.750x_{11} - 4.626x_{12}, \\
\operatorname{logit} p_2 &= 22.543 + 4.042x_1 + 3.115x_2 - 2.835x_3 - 2.711x_4 - 0.876x_5 - 2.086x_6 \\
&\quad - 3.294x_7 - 3.421 - 3.343x_9 - 2.640x_{10} - 2.708x_{11} - 1.491x_{12}, \\
\operatorname{logit} p_3 &= 0\,(\text{reference group}).
\end{aligned} \tag{13}$$

Additionally, to validate whether the model shows adequate practical meaning, it is necessary to test the goodness of fit, pseudo R-squared, and likelihood ratio of the model. The tests (including the Pearson chi-square test and the deviance chi-square test) of goodness of fit can test whether the model fits the original data, or not. If the significance level exceeds 0.05, the fitting effect is favourable. The pseudo R-squared value can verify the degree of explanation offered by the model for information contained in its original variables, which is shown in Cox, Nagelkerke and McFadden pseudo R-squared values. The closer the result is to 1, the better the explanation. The likelihood ratio test measures the contribution of original variables to the model. If the significance level is lower than 0.05, the contribution of original variables is high.

According to the test result (Table 8) obtained through use of the logistic regression model, the goodness of fit, pseudo R-squared, and likelihood ratio of the model all satisfy test requirements. This indicates that the extracted principal components PCA1, PCA2, and PCA3 also retain key information about the data while effectively realising dimension reduction, which makes a significant contribution to the construction of the logistic regression model. The final result obtained through model regression is also meaningful.

3.4.3. Validation of Discriminatory Effect of the Model. By taking IR CAPs of 33 original asphalt samples as verification samples, the discrimination effect obtained through the multinomial logistic regression model in multivariate statistical analysis was evaluated by applying formula (13). The discrimination result of multinomial logistic regression in multivariate statistical analysis is shown in Table 9.

As shown in Table 9, discrimination accuracies of 15, 12, and six asphalt samples separately produced by crude oil sourced from the Middle East, South America, and Bohai Rim region of China are 93.3%, 91.7%, and 100%, respectively. The comprehensive discrimination accuracy is 93.9%. The above result showed that multivariate logistic regression analysis based on PCA can rapidly discriminate the origins of asphalt.

4. Conclusions

Based on ATR-FTIR technology, the infrared spectra of 33 kinds of asphalt produced by crude oil from the Middle East, South America, and Bohai Rim region of China were collected. Furthermore, the 12 selected CAPs of infrared spectra were analysed by multivariate statistics. The comprehensive accuracy of the logistic regression model based on PCA in discriminating asphalt, which were produced by crude oil from three different regions reached 93.9%. The results indicated that the combination of ATR-IR spectral analysis and multivariate statistics can accurately and nondestructively discriminate between different crude oil source of asphalt. Moreover, the method shows some remarkable advantages, including ease of operation, rapidity, and high accuracy, which is important when controlling the origins and quality of asphalt and improving the performance thereof.

The method provided in this paper is suitable for the oil source identification of base asphalt produced by crude oil from different regions and can also provide reference for other kinds of asphalt, such as polymer-modified asphalt. However, the accuracy and applicability of this method need to be further improved. In particular, whether the asphalt produced by crude oil mixing from different regions can be effectively identified needs further research.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was financially supported by the National Natural Science Foundation of China (no. 51608511), Key R&D Project of Shandong Province (2018GGX105013), and Project of Science and Technology Support for Youth Entrepreneurship in Colleges and Universities of Shandong Province (2019KJG004). The authors would like to acknowledge many coworkers, students, and laboratory assistants for providing technical help on instrument analysis.

References

[1] S. Zhao and J. Liu, “Using recycled asphalt pavement in construction of transportation infrastructure: Alaska experience,” *Journal of Cleaner Production*, vol. 177, pp. 155–168, 2018.

[2] Y. Tan and M. Guo, “Using surface free energy method to study the cohesion and adhesion of asphalt mastic,” *Construction and Building Materials*, vol. 47, pp. 254–260, 2013.

[3] P. Li, J. Liu, and S. Zhao, “Implementation of stress-dependent resilient modulus of asphalt-treated base for flexible pavement design,” *International Journal of Pavement Engineering*, vol. 19, no. 5, pp. 439–446, 2018.

[4] F. C. G. Martinho and J. P. S. Farinha, “An overview of the use of nanoclay modified bitumen in asphalt mixtures for enhanced flexible pavement performances,” *Road Materials and Pavement Design*, vol. 20, no. 3, pp. 671–701, 2019.

[5] F. Xiao, D. Ma, J. Wang, D. Cai, L. Lou, and J. Yuan, “Impacts of high modulus agent and anti-rutting agent on performances of airfield asphalt pavement,” *Construction and Building Materials*, vol. 204, pp. 1–9, 2019.

[6] Y. Sun, W. Wang, and J. Chen, “Investigating impacts of warm-mix asphalt technologies and high reclaimed asphalt pavement binder content on rutting and fatigue performance of asphalt binder through MSCR and LAS tests,” *Journal of Cleaner Production*, vol. 219, pp. 879–893, 2019.

[7] M. Guo, Y. Tan, D. Luo et al., “Effect of recycling agents on rheological and micromechanical properties of SBS-modified asphalt binders,” *Advances in Materials Science and Engineering*, vol. 2018, Article ID 5482368, 12 pages, 2018.

[8] W. Y. Fan, X. Xin, M. Liang et al., “Study on rheology and storage stability of modified asphalt,” *Journal of China University of Petroleum*, vol. 39, pp. 165–170, 2015.

[9] M. Lagos-Varas, D. Movilla-Quesada, J. P. Arenas et al., “Study of the mechanical behavior of asphalt mixtures using fractional rheology to model their viscoelasticity,” *Construction and Building Materials*, vol. 200, pp. 124–134, 2019.

[10] P. Wang, Z. J. Dong, Y. Q. Tan et al., “Effect of multi-walled carbon nanotubes on the performance of styrene-butadiene-styrene copolymer modified asphalt,” *Materials and Structures*, vol. 50, no. 1, p. 17, 2017.

[11] M. Guo, Y. Tan, L. Wang, and Y. Hou, “Diffusion of asphaltene, resin, aromatic and saturate components of asphalt on mineral aggregates surface: molecular dynamics simulation,” *Road Materials and Pavement Design*, vol. 18, no. 3, pp. 149–158, 2017.

[12] Y. Xue, Z. Ge, F. Li, S. Su, and B. Li, “Modified asphalt properties by blending petroleum asphalt and coal tar pitch,” *Fuel*, vol. 207, pp. 64–70, 2017.

[13] V. Y. Pivsaev, P. E. Krasnikov, A. A. Pimenov, and D. E. Bykov, “Enhancement of adhesive properties of road asphalts, waste oil processing products,” *Petroleum Chemistry*, vol. 55, no. 1, pp. 80–83, 2015.

[14] M. N. Siddiqui, “NMR fingerprinting of chemical changes in asphalt fractions on oxidation,” *Petroleum Science and Technology*, vol. 28, no. 4, pp. 401–411, 2010.

[15] M. N. Siddiqui, “NMR finger printing of chemical changes in asphalt fractions on oxidation,” *Petroleum Science and Technology*, vol. 27, no. 17, pp. 2033–2045, 2009.

[16] R. Ren, K. Han, P. Zhao et al., “Identification of asphalt fingerprints based on ATR-FTIR spectroscopy and principal component-linear discriminant analysis,” *Construction and Building Materials*, vol. 198, pp. 662–668, 2019.

[17] C. M. Xu, Y. Liu, S. Q. Zhao et al., “Analysis of heteroatomic compounds in petroleum asphaltenes by high resolution mass spectrometry,” *Journal of China University of Petroleum*, vol. 37, pp. 190–195, 2013.

[18] Q. F. Wu and Z. C. Tang, “Study on the determinants of economic compensation for ecological damage caused by oil spill at sea,” *Journal of China University of Petroleum*, vol. 34, pp. 1–8, 2018.

[19] Y. Zhong, H. Wang, G. Lu, Z. Zhang, Q. Jiao, and Y. Liu, “Detecting functional connectivity in fMRI using PCA and regression analysis,” *Brain Topography*, vol. 22, no. 2, pp. 134–144, 2009.

[20] M. Kharbach, R. Kamal, M. A. Mansouri et al., “Selected-ion flow-tube mass-spectrometry (SIFT-MS) fingerprinting versus chemical profiling for geographic traceability of Moroccan argan oils,” *Food Chemistry*, vol. 263, pp. 8–17, 2018.

[21] P. Y. Zhou, C. S. Chen, J. J. Ye et al., “Combining molecular fingerprints with multidimensional scaling analyses to identify the source of spilled oil from highly similar suspected oils,” *Marine Pollution Bulletin*, vol. 93, no. 1-2, pp. 121–129, 2015.

[22] A. Ismail, M. E. Toriman, H. Juahir et al., “Chemometric techniques in oil classification from oil spill fingerprinting,” *Marine Pollution Bulletin*, vol. 111, no. 1-2, pp. 339–346, 2016.

[23] J. M. Bayona, C. Domínguez, and J. Albaigés, “Analytical developments for oil spill fingerprinting,” *Trends in Environmental Analytical Chemistry*, vol. 5, pp. 26–34, 2015.

[24] E. C. Y. Chan, P. K. Koh, M. Mal et al., “Metabolic profiling of human colorectal cancer using high-resolution magic angle spinning nuclear magnetic resonance (HR-MAS NMR) spectroscopy and gas chromatography mass spectrometry

(GC/MS)," *Journal of Proteome Research*, vol. 8, no. 1, pp. 352–361, 2009.

[25] Y. Li, S. Wu, and S. Amirkhanian, "Investigation of the graphene oxide and asphalt interaction and its effect on asphalt pavement performance," *Construction and Building Materials*, vol. 165, pp. 572–584, 2018.

[26] G. Xu and H. Wang, "Molecular dynamics study of oxidative aging effect on asphalt binder properties," *Fuel*, vol. 188, pp. 1–10, 2017.

[27] N. Nciri and N. Cho, "A thorough study on the molecular weight distribution in natural asphalts by gel permeation chromatography (GPC): the case of Trinidad lake asphalt and asphalt ridge bitumen," *Materials Today: Proceedings*, vol. 5, no. 11, pp. 23656–23663, 2018.

[28] D. O. Larsen, J. L. Alessandrini, A. Bosch, and M. S. Cortizo, "Micro-structural and rheological characteristics of SBS-asphalt blends during their manufacturing," *Construction and Building Materials*, vol. 23, no. 8, pp. 2769–2774, 2009.

[29] H. Yao, Q. Dai, and Z. You, "Fourier transform infrared spectroscopy characterization of aging-related properties of original and nano-modified asphalt binders," *Construction and Building Materials*, vol. 101, pp. 1078–1087, 2015.

[30] F. Zhang, J. Yu, and J. Han, "Effects of thermal oxidative ageing on dynamic viscosity, TG/DTG, DTA and FTIR of SBS-and SBS/sulfur-modified asphalts," *Construction and Building Materials*, vol. 25, no. 1, pp. 129–137, 2011.

[31] A. Dazzi, C. B. Prater, Q. Hu, D. B. Chase, J. F. Rabolt, and C. Marcott, "AFM-IR: combining atomic force microscopy and infrared spectroscopy for nanoscale chemical characterization," *Applied Spectroscopy*, vol. 66, no. 12, pp. 1365–1384, 2012.

[32] M. Lopes, V. Mouillet, L. Bernucci, and T. Gabet, "The potential of attenuated total reflection imaging in the mid-infrared for the study of recycled asphalt mixtures," *Construction and Building Materials*, vol. 124, pp. 1120–1131, 2016.

[33] Y. Xiong, G. Chen, S. Guo, and G. Li, "Lifetime prediction of NBR composite sheet in aviation kerosene by using nonlinear curve fitting of ATR-FTIR spectra," *Journal of Industrial and Engineering Chemistry*, vol. 19, no. 5, pp. 1611–1616, 2013.

[34] G. Y. Chen and S. E. Qian, "Denoising of hyperspectral imagery using principal component analysis and wavelet shrinkage," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 3, pp. 973–980, 2010.

[35] S. Tonidandel and J. M. LeBreton, "Determining the relative importance of predictors in logistic regression: an extension of relative weight Analysis," *Organizational Research Methods*, vol. 13, no. 4, pp. 767–781, 2010.

[36] R. Azen and N. Traxel, "Using dominance analysis to determine predictor importance in logistic regression," *Journal of Educational and Behavioral Statistics*, vol. 34, no. 3, pp. 319–347, 2009.

[37] J. Adler and I. Parmryd, "Quantifying colocalization by correlation: the Pearson correlation coefficient is superior to the Mander's overlap coefficient," *Cytometry Part A*, vol. 77A, no. 8, pp. 733–742, 2010.

[38] A. Soberón and W. Stute, "Assessing skewness, kurtosis and normality in linear mixed models," *Journal of Multivariate Analysis*, vol. 161, pp. 123–140, 2017.

[39] A. D. Ho and C. C. Yu, "Descriptive statistics for modern test score distributions," *Educational and Psychological Measurement*, vol. 75, no. 3, pp. 365–388, 2015.