摩擦学学报
Tribology
ISSN 1004-0595,CN 62-1095/O4

《摩擦学学报》网络首发论文

题目：基于组合优化特征波长的润滑油基础油成分定量分析方法研究(英文)

作者：夏延秋，NAY MINAUNG，王裕兴，冯欣

收稿日期：2024-10-10
网络首发日期：2025-04-14
引用格式：夏延秋，NAY MINAUNG，王裕兴，冯欣．基于组合优化特征波长的润滑油基础油成分定量分析方法研究(英文)[J/OL]．摩擦学学报．
https://link.cnki.net/urlid/62.1095.o4.20250411.1451.010

Published online：2025-04-14 11:22:53 URL：https://link.cnki.net/urlid/62.1095.o4.20250411.1451.010

基于组合优化特征波长的润滑油基础油成分定量分析方法研究

夏延秋*,NAY MINAUNG，王裕兴,冯欣

（华北电力大学 能源动力与机械工程学院，北京 102206）

摘 要：针对润滑油基础油成分的定量分析，选取矿物油(KN4010)、碳氢基合成油(PAO40)和合成酯(PriEco 3000)这 3 种油品成分作为定量分析对象，采集不同配比配制的润滑油基础油样品的中红外光谱数据，采用 SiPLS-BGWO 组合优化方法在光谱全范围内筛选特征波长，剔除大量冗余无效信息，降低搜索空间维数。试验结果表明：对于矿物油、碳氢基合成油和多元醇酯的含量预测，组合优化模型的误差指标明显改善，与采用所有光谱波长相比，均方根误差(RMSE)降低幅度最大可达 60.58%，拟合指标的 R^2 值均高于 99%。此外，SiPLS-BGWO 方法将特征波长数量减少至 40 个以下，显著降低了运算负担，有效提高了多物质组分定量分析模型的准确性和适用性。

关键词：润滑油基础油；中红外光谱；特征波长筛选；区间带筛查；组合优化模型

中图分类号：TH117.1 **文献标志码**：A

Quantitative Analysis Method of Lubricant Base Oil Composition Based on Combined Optimized Characteristic Wavenumber

XIA Yanqiu*, NAY MINAUNG, WANG Yuxing, FENG Xin

(*School of Energy, Power and Mechanical Engineering, North China Electric Power University, Beijing 102206*)

Abstract: For the quantitative analysis of lube oil base oil components, three oil components, mineral oil (KN4010), hydrocarbon-based synthetic oil (PAO40), and synthetic ester (PriEco 3000) were selected as quantitative analysis objects, and then the mid-infrared spectral data of lube oil base oil samples formulated in different ratios were collected. The synergy interval partial least squares-binary grey wolf optimization algorithm (SiPLS-BGWO) combination optimization method was used to screen the characteristic wavenumbers in the full range to eliminate redundant invalid information and reduce the search space dimension. By optimizing the selection of characteristic wavenumbers, the SiPLS-BGWO approach not only enhanced the prediction accuracy

Received 10 October 2024, revised 4 December 2024, accepted 12 December 2024

*Corresponding author. E-mail: xiayq@ncepu.edu.cn, Tel: +86-10-61772251.

This project was supported by the Beijing Natural Science Foundation (2232066) and the Open Project Foundation of State Key Laboratory of Solid Lubrication (LSL-2212).

北京市自然科学基金项目(2232066)和固体润滑国家重点实验室开放课题(LSL-2212)资助.

but also demonstrated its ability to address challenges associated with overlapping spectral features in complex mixtures. The test results showed that the combined optimization model's error indexes were significantly improved for the content prediction of mineral oil, hydrocarbon-based synthetic oil, and polyol ester. The RMSE (root mean square error) was reduced by up to 60.58% compared to using all spectral wavenumbers, and the fit indexes' R2 values were higher than 99%. The significant reduction in RMSE underscored the method's capability to identify and eliminate irrelevant or noisy spectral information, ensuring that the predictive model focused only on relevant features. In addition, the SiPLS-BGWO method had reduced the number of characteristic wavenumbers to less than 40, significantly reducing the operational burden and effectively improving the accuracy and applicability of the quantitative analysis model for multi-matter components. The ability to reduce the number of characteristic wavenumbers to below 40 demonstrated the algorithm's efficiency in dimensionality reduction while retaining essential predictive information. The results affirmed that the SiPLS-BGWO model was a powerful tool for predictive modeling, providing a balance between accuracy and efficiency in the quantitative analysis of multi-component systems. And a novel framework for bridging the gap between spectral data complexity and actionable chemical insights, setting a precedent for future developments in the field.

Key words: lubricant base oil; mid-infrared spectroscopy; feature wavenumber screening; interval band screening; combinatorial optimization model

1. Introduction

As we all know, lubricants are composed of base oils and additives, and there are three main types of base oils: including mineral-based oils, synthetic-based oils, and vegetable-based oils. Mineral-based oils are widely used[1], with more than 90% of the usage, but some applications must use synthetic or vegetable oils in combination with mineral oils. Usually, the base oil determines the appearance, density, viscosity, flash point, freezing point, pour point, moisture, mechanical impurities, residual carbon, etc. Of course, some additives can also improve and enhance the above properties. Although lubricant base oil accounts for 95%~99% of the total weight of oil products, the class

of oil products is diverse and complex. The lubricant product's performance and appearance cannot identify the lubricant class nor determine its composition content[2]. The standard measure to identify and analyze the content of lubricant base oils is to use traditional physical and chemical tests and other methods to determine and analyze. Still, it is relatively complex to identify the types of mixed base oils, especially when determining the content of base oil components, which is more complicated. The samples' mid-infrared spectral information was obtained using Fourier transform infrared radiation (FT-IR) and processed using first and second derivatives, complementing methods such as Raman spectroscopy, performance testing, and gas chromatography[3]. These methods are mostly affected by human factors such as high time and labour costs, and the results have significant errors[4-6], so an accurate and effective oil analysis technique is needed. Mid-infrared spectroscopy has the advantage of being non-destructive and rapid as well as large and low-cost, so it is widely used for both qualitative and quantitative analysis of the composition of substances [7-8]. Machine-learning data analysis processing methods are required for determination and analysis to find the relevant characteristic peaks of infrared spectra quickly, eliminate interfering factors, and complete the task of substance identification and content prediction. However, most studies in lubricant analysis have primarily focused on oil identification tests[9]. Only a limited number of studies have been conducted on the identification and content analysis of base oils within the latest categories of oils[10].

Lubricant composition content has an essential impact on its performance, and traditional lubricant development has been carried out mainly using repeated extensive design experiments and performance testing[11]. Only by choosing the right proportional content of blended lubricants can they meet the requirements, so quantitative analysis and control of the constituent components in the oil development process is required. In this paper, we start with infrared spectroscopy and combine the machine learning method of intelligent optimization algorithm to construct a quantitative analysis model of the lubricant composition. The commonly used algorithm for creating predictive models is partial least squares (PLS)[12-13], a classical linear modelling method that can overcome the covariance problem and reduce noise interference, and the analysis of

spectral data is widely used. For mid-infrared spectral-type data with many features, the training process will occupy a lot of memory, prolong the computing time, and seriously degrade the prediction quality[14]. Therefore, screening out redundant information and reducing the search dimension is necessary, so filtering the characteristic wavenumbers of mid-infrared spectra is essential. Spectral feature wavenumber screening methods are divided into feature band screening algorithms and feature point screening algorithms. The commonly used feature band screening algorithms include interval partial least squares (iPLS)[15-17], Synergy interval partial least squares (SiPLS)[18], and backward interval partial least squares (BiPLS)[15, 19]. The use of these band screening algorithms alone for the interference in the band interval information cannot be removed and is heavily influenced by interval segmentation. However, the commonly used algorithms for feature wavenumber point screening are generally the genetic algorithm (GA)[20], the ant colony algorithm (ACO)[21], and the emerging swarm intelligence search algorithm binary particle swarm algorithm (BPSO)[10,22], binary bat algorithm (BBA)[23-24] and binary grey wolf optimization algorithm (BGWO)[25-26] in recent years, and these methods are either more obsolete when used singly complex or have poor operational efficiency, too much interference information appears in the full range of spectral wavenumbers to locate the characteristic wavenumbers accurately, and the calculation of complex operations is too long. As can be seen, single optimization methods have advantages in terms of computational efficiency, global searchability, and generality, but each has unavoidable problems. Therefore, this paper combines the complementary strategies solution and proposes an optimization method of mid-infrared spectral feature wavenumber screening based on combining feature band and feature-wavenumber point screening methods.

In this paper, for the quantitative analysis of three oil components in lubricating base oils, a combined SiPLS-BGWO optimization method was designed and implemented to perform feature wavenumber screening of spectral full-area feature data. Firstly, the feature band screening method was used to screen the feature bands and built a single optimization model. Then, single and combined optimization models were created for all wavenumbers and the selected feature bands. Finally, the results of

the single and combined optimization models were compared with those obtained using complete wavenumber data and principal component analysis to compress the feature data into the model. The differences in wavenumber range and prediction results between the single optimization model, the combined optimization model, and the base model were analyzed by example tests to examine the effect of infrared spectral feature wavenumber screening and the prediction accuracy of the quantitative analysis model and to verified the effectiveness of the SiPLS-BGWO combined optimization method.

2. Infrared spectroscopy sample extraction

2.1 Lubricant sample preparation

The test sample was formulated with three oil components: mineral oil (Kramer KN4010), hydrocarbon-based synthetic oil (Mobil PAO40), and polyol ester (NACO PriEco 3000) to form the lubricant base oil, and this blend was used as the base oil for formulating customized industrial equipment lubricants, Table 1. Changsha Zhongcheng Lubricant Co, Ltd., supplied the base oils, and the relevant data, including physical and chemical properties, were obtained from the product specifications provided by the manufacturer.

Table 1 Shows the typical physical and chemical properties of the three oils.

Property	Mineral oil (KN4010)	Hydrocarbon-based synthetic oil (PAO40)	Synthetic ester (PriEco 3000)
Viscosity at 40~100 ℃ (cSt)	40~5.8	40~7.5	45~8
Viscosity index (VI)	95~100	140~150	160~170
Pour point/℃	-10~-15	-50~-60	-40~-45
Flash point/℃	200~220	240~260	250~270
Density at 15℃ /(g/cm ^{3})	0.87~0.89	0.82~0.84	0.92~0.95
Thermal oxidation stability	Moderate	High	Very High
Lubricity	Moderate	Excellent	Superior
Biodegradability	Low	Low to Moderate	High
Additive solubility	Moderate	Good	Excellent
Sulfur content/(r/min)	300~500	<10	<5
Evaporation loss (% weight)	2%~5%	<1%	<0.5%
Copper corrosion (3 h at 100℃)	1b	1a	1a
Foaming tendency (Seq. I, mL)	30/0	Oct~00	Oct~00
Water separability (min at 54℃)	30~60	< 10	< 5

The sample blends were mixed and developed by the previous permutations to form

30 samples, and the sample design scheme is shown in Table 2.

Table 2 Sample Design Solutions

Factor / Number	Mineral oil/%	Hydrocarbon-based synthetic oils/%	Polyol esters /%
1	55	25	20
2	50	20	30
3	40	45	15
4	45	30	25
5	45	25	30
6	35	35	30
7	60	20	20
8	35	45	20
9	30	45	25
10	55	30	15
11	45	40	15
12	45	35	20
13	60	25	15
14	50	35	15
15	30	55	15
16	30	40	30
17	40	40	20
18	40	35	25
19	50	25	25
20	55	20	25
21	35	30	35
22	45	20	35
23	30	50	20
24	50	30	20
25	40	30	30
26	50	15	35
27	35	40	25
28	30	35	35
29	35	50	15
30	40	25	35

2.2 Acquisition of raw sample spectral data

A Thermo Scientific Nicolet iS5 Fourier transforms infrared spectrometer was used as the sample data acquisition instrument with a spectral range of 7 800~350 cm^{-1} and a KBr window sheet with a transmission wavenumber of 7 800~400 cm^{-1} and a transmission rate of >92%. Acquisition settings: 16 scans, resolution 16, data interval

1.928 cm^{-1}. Spectral data were collected once for each sample after reloading to simulate the manual errors generated by different collection personnel during IR spectrum collection. Four spectral data were collected for each sample, for a total of 120 spectral data.

2.3 Sample set division

Machine learning methods for building infrared spectral analysis models required sufficient and representative samples, and selecting representative samples required experienced experts[27]. Still, most people have yet to gain advanced experience, so it was necessary to select representative samples from the many samples collected to build training models using relevant sample partitioning methods. The commonly used methods for sample set partitioning included random partitioning, sample set partitioning based on joint *x-y* distances (SPXY)[28], and Kolmogorov-Smirnov (K-S)[29]. Still, the random selection method could not ensure whether the selected models met the requirements of the training set. In contrast, K-S partitioning of sample sets only considered the relationship between sample spectra and ignored the relationship between the spectra and the corresponding chemical values. The SPXY algorithm fully considered the relationship between the spectral information of the sample and the corresponding physicochemical properties. Based on K-S, it calculated the joint distance between spectral and chemical values. The method could effectively cover the multidimensional space and significantly avoided the problem of samples with weak spectral information and low chemical value content needing to be more responsive to the K-S algorithm. K-S algorithm could effectively improve the model's prediction performance. Therefore, the SPXY method was used in this study to divided the collected infrared spectral data into training and prediction sets in a 3:1 ratio 90 samples in the training set and 30 samples in the prediction set). The statistical results of the content of each component in the training and prediction sets of lubricant base oils were shown in Table 3.

Table 3 Statistics on the content of each component in the training and prediction sets of lubricant base oils

Sample division		Sample size	Average value /%	Maximum value /%	Minimum value /%	Standard deviation /%
Mineral oil	Training set	90	43.055 6	60	30	9.109 0
	Test set	30	42.166 7	60	30	9.348 0
Hydrocarbon-based synthetic oils	Training set	90	32.666 7	55	15	10.117 3
	Test set	30	34.666 7	55	15	10.333 5
Polyol esters	Training set	90	24.277 8	35	15	7.053 6
	Test set	30	23.166 7	35	15	7.007 8

3 Model construction and evaluation criteria

3.1 Spectral Preprocessing

To avoid the measuring instrument's zero drift and significant differences in data values, the method, as shown in formula (1), is selected for normalization processing. The minimum value of all infrared spectrum data is set to 0, and the maximum value is set to (1).

$$\boldsymbol{y_{ij}} = \frac{\boldsymbol{x_{ij}} - \boldsymbol{x_{min}}}{\boldsymbol{x_{max}} - \boldsymbol{x_{min}}} \in Y \quad (1)$$

In formula (1), $\boldsymbol{x_{ij}}$ was the original spectrum data, $\boldsymbol{x_{min}}$ was the minimum value of the spectrum data, and $\boldsymbol{x_{max}}$ was the maximum value of the spectrum data. $\boldsymbol{y_{ij}}$ denotes the normalized spectral value corresponding to $\boldsymbol{x_{ij}}$, and Y represents the set of all normalized spectral data.

3.2 Combinatorial optimization of characteristic wavenumber screening schemes

Since the infrared spectra had many feature dimensions[30], Some bands were too correlated, and the absorption peaks overlapped; the spectral data should have been filtered by band optimization, which searches for characteristic band regions corresponding to the relevant substances and eliminated overlapping and redundant information. Feature wavenumber filtering algorithms could compress spectral data features, improve computing efficiency, and improve model performance, broadly divided into feature band filtering algorithms and feature-wavenumber point filtering algorithms[31].

The commonly used feature band screening algorithms mainly included SiPLS and BiPLS, two effective feature band screening methods proposed based on iPLS that were

widely used in fields such as infrared spectral analysis. iPLS divided the spectrum into *k* intervals and performed partial least squares regression on each interval separately to obtain *k* regression models. The cross-validation method was used to calculate the calculated root mean square error of each *k* model and compare the error values of each model. The optimal model was the regression model corresponding to the interval with the smallest error. In contrast, both BiPLS and SiPLS operated on the divided subintervals based on iPLS: BiPLS first eliminated the interval with the worst correlation among the *k* intervals and built a PLS model for the remaining *k-1* intervals. Then, the worst correlation interval among the remaining *k-1* intervals was eliminated again, a PLS model was built for the remaining *k-2* intervals, and so on, until only one interval remains. The root-mean-square error value of each PLS model was used as the evaluation index, where the combination of intervals corresponding to the minimum value of the root-mean-square error was the optimal interval. SiPLS was a joint interval of *j* ($2 \leqslant j \leqslant k$) intervals randomly selected among the *k* intervals delineated by iPLS to build a PLS model, and a total of C_k^j PLS models are built, and the combination of *j* intervals corresponding to the minimum root mean square error value was the optimal interval. The computation volume of SiPLS was highly dependent on the values of *k* and *j*. When the value of *k* was specific, the computation volume would increase exponentially with the increase of the value of *j*. Therefore, the value of *j* should have been manageable during the computation of SiPLS, which was generally less than 5.

The wavenumber point screening algorithm could be chosen from the group intelligent search algorithm, which has emerged in recent years, and the algorithm was mainly used to solve the continuous space function optimization problem at the beginning of the proposed algorithm. Later, to decode the practical issues of feature selection and combinatorial optimization in work, different discrete discretization processing schemes have been proposed one after another to transform the continuous problem into a 0~1 planning problem, which could thus be used as a method to deal with large-scale feature engineering preferences, proposed the Binary Gray Wolf Optimization algorithm (BGWO), which transforms the gray wolf position by updating

the position of the gray wolf using the Sigmoid function as:

$$\vec{X}_d(t+1)=\begin{cases}1, & if\ sigmoid\left(\dfrac{\overrightarrow{X_1}+\overrightarrow{X_2}+\overrightarrow{X_3}}{3}\right)\geqslant r\\ 0, & otherwise\end{cases} \tag{2}$$

$$sigmoid(a)=\frac{1}{1+e^{-10(x-0.5)}} \tag{3}$$

Where r was a random value between $\{0,1\}$; $\overrightarrow{X_1}$, $\overrightarrow{X_2}$ and $\overrightarrow{X_3}$ do the three wolves give the optimal prey position information; $\vec{X}_d(t+1)$ was the binary position of the wolf updated after t iterations of the search in d-dimensional space. BGWO was tested to set the relevant parameters; its population number was set to 30, and the maximum number of iterations was set to 500.

In IR spectral feature wavenumber preference processing, the band screening algorithms (BiPLS and SiPLS) usually select one or several consecutive intervals. There was still a large amount of redundant information within the interval. The screening results of such methods were seriously affected by the interval division, so the results obtained by using them alone could be more satisfactory. To further remove the interference information, reduce the data dimensionality, and improve the prediction ability of the model, based on the above two feature band selection algorithms, the feature-wavenumber point screening algorithm Binary Swarm Intelligence Algorithm (BGWO) was introduced to screen the selected spectral data twice and compared and analyzed the processing effect with Binary Particle Swarm Algorithm (BPSO) and Binary Bat Algorithm (BBA).

3.3 Quantitative analysis model construction for lubricant base oil composition

The characteristic wavenumber expression regions of the mid-infrared spectra of mineral oils, hydrocarbon-based synthetic oils, and polyol esters, the three main components of lubricant base oils, were preferentially selected using a combination of optimized characteristic wavenumbers, respectively. Using PLS as a basis, the optimized characteristic wavenumbers of infrared spectra in lubricating oil were input into the model to construct a quantitative analysis model of lubricating oil base oil

composition, and the overall processing and analysis steps were as follows：

Step 1: By dividing SiPLS and BiPLS into subintervals from 10 to 30, the prediction performance of the models with different numbers of subintervals was observed and counted. In particular, for SiPLS, the number of subintervals chosen was set to 2, 3 and 5 to prevent the model computation from skyrocketing and to find out the optimal state for dividing the number of different intervals as well as in the case of its choice of different subintervals so that the model performance was optimal.

Step 2: The preferred performance models of SiPLS and BiPLS for different compositions (mineral oil, hydrocarbon-based synthetic oil, and polyol ester) were compared and selected for the next step of secondary screening of characteristic wavenumbers. Each characteristic wavenumber point of the IR spectrum had only two states, so the characteristic wavenumber point screening could be said to be the problem of finding a suitable 0/1 string, the length of which was the number of wavenumber points of the original spectral data (a total of 1869 wavenumber points), where 0 is not selected, and one was selected. The normalized raw spectral data were substituted into BPSO, BBA, and BGWO to optimize the characteristic wavenumber points, and the training accuracy and the number of incoming wavenumber points of their models were observed and counted.

Step 3: For the quantitative analysis model of each lubricant component, the selected feature band screening method, the chosen feature-wavenumber point screening method, and the combined optimized feature screening method were compared to analyzing the applicability issues in different situations. We also compared the full spectral characteristic wavenumber and the traditional PCA compressed characteristic wavenumber method in the PLS model to investigate the performance and computational efficiency improvement.

The workflow diagram was shown in Fig 1:

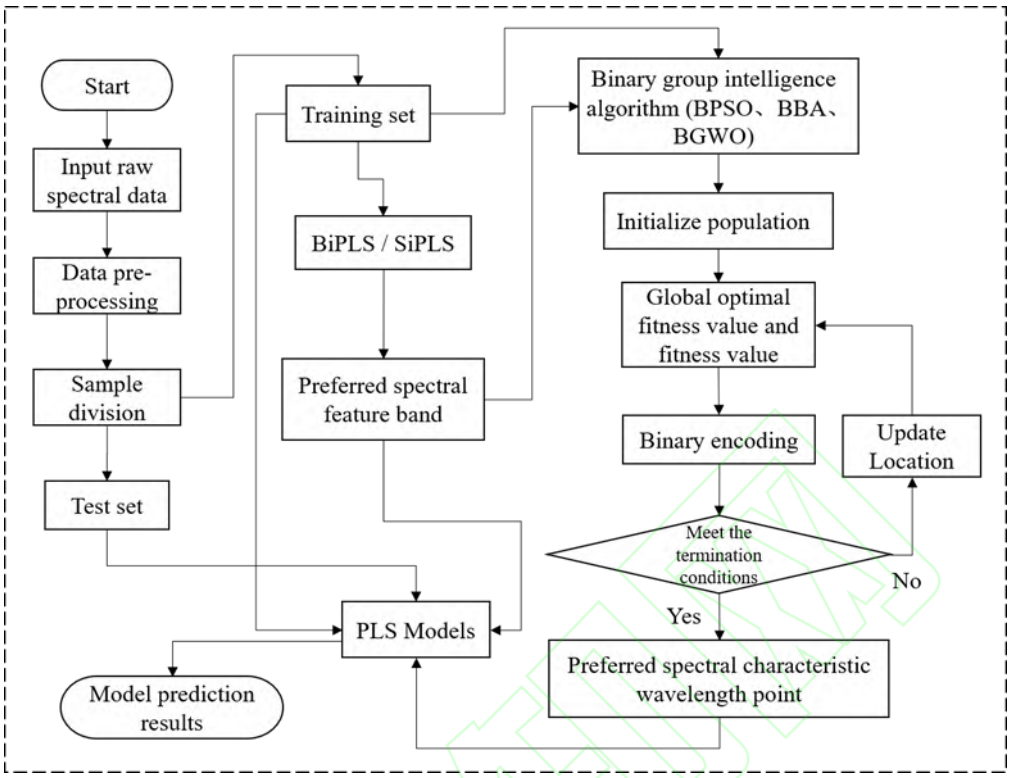

Fig 1. Characteristic wavenumber screening program workflow diagram

BiPLS: Backward Interval Partial Least Squares; SiPLS: Synergy Interval Partial Least Squares; PLS: Partial Least Squares; BPSO: Binary Particle Swarm Optimization Algorithm; BBA: Binary Bat Algorithm; BGWO: Binary Grey Wolf Optimization Algorithm

3.4 Evaluation Criteria

In the paper, mean absolute percentage error (MAPE), root mean square error (RMSE), and coefficient of determination (R^2) were used as evaluation criteria for the comprehensive performance of the model, and the formulae were calculated as follows:

$$MAPE = 100 \times \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{5}$$

$$R^2 = \frac{\left(n \sum_{i=1}^{n} y_i \hat{y}_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} \hat{y}_i \right)^2}{\left[n \sum_{i=1}^{n} \hat{y}_i^2 - \left(\sum_{i=1}^{n} \hat{y}_i \right)^2 \right] \left[n \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i \right)^2 \right]} \tag{6}$$

Where n was the total number of samples, $\hat{y}_i$ was the predicted value of the test sample, and y_i was the actual value. The coefficient of determination R^2 was between 0 and 1, and the closer it is to 1, the better the model fits and the better the performance.

4 Results and Analysis

4.1 Infrared spectral data pre-processing

The original data were linearly transformed to map the processed data between 0 and 1. The results of normalizing the raw spectral data were shown in Fig 2：

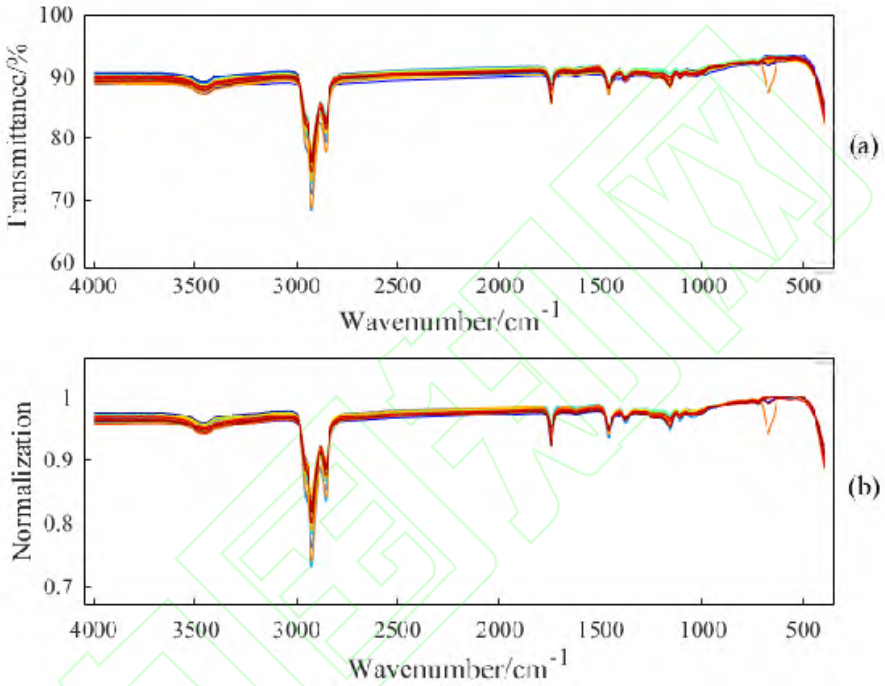

Fig 2. Normalization of raw spectral data: (a) raw spectra; (b) normalized processing

4.2 Comparison of BiPLS and SiPLS screening feature bands

The pre-processed full spectral data range (4 000~400 cm^{-1}) was divided into 10 to 30 subintervals, respectively, and the BiPLS characteristic spectral interval screening model was established to build a quantitative analysis model of lubricant components with the preferred spectral interval and to make content predictions. The spectral screening results were shown in Fig 3.

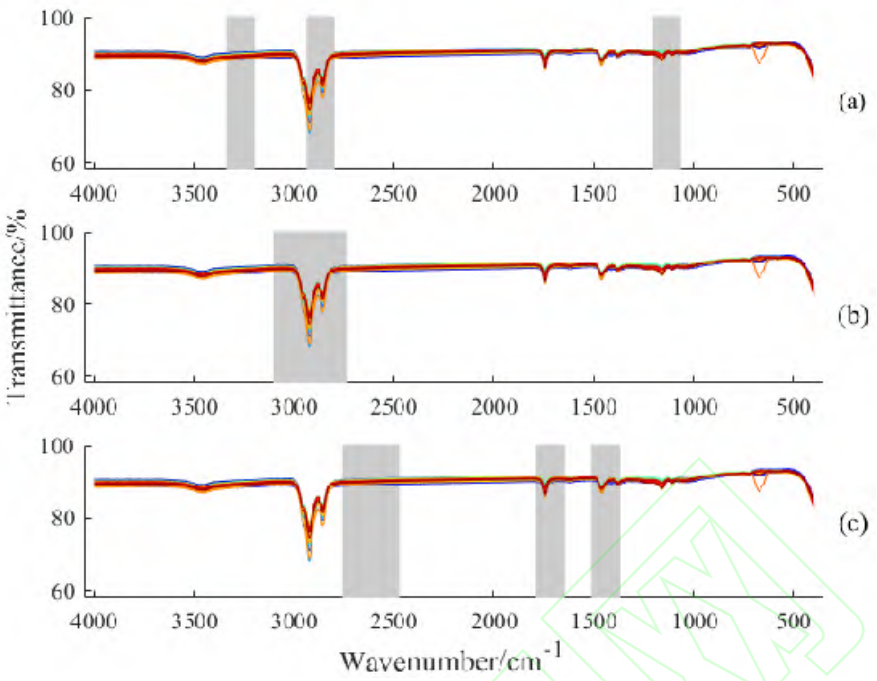

Fig 3. BiPLS screening of characteristic wavenumber distribution of (a) mineral oil, (b) hydrocarbon-based synthetic oil, and (c) polyol ester

As seen in Table 4, among all the corresponding BiPLS spectral interval screening models for mineral oil components, when the whole spectrum was divided into 27 subintervals, the best modelling results were selected for the combination of {6, 19, 22} subintervals with a training set RMSECV=1.270 6 and 208 selected wavenumber points; For hydrocarbon-based synthetic oil components, the best modelling results were selected for the combination of {14-15} subintervals when the entire spectrum was divided into 20 subintervals with a training set RMSECV=0.956 and 186 selected wavenumber points; For the polyol ester component, when the whole spectrum was split into 26 subintervals, the best modelling result was selected for the combination of {8, 10, 16 ,17} subintervals with a training set RMSECV= 0.697 8 and 288 wavenumber chosen points.

Table 4 BiPLS Preferred Model Results

Lubricant composition	Total number of intervals	Selected sub-interval	Cross-validation root-mean-square error	Selected wavenumber points
Mineral oil	27	{6, 19, 22}	1.270 6	208

Hydrocarbon-based synthetic oils	20	{14, 15}	0.956 0	186
Polyol esters	26	{8, 10, 16, 17}	0.697 8	288

The preprocessed full spectral data range (4 000~400 cm^{-1}) was divided into 10 to 30 subintervals, and the SiPLS characteristic spectral interval screening model was established by selecting 2, 3, and 4 joint intervals. The preferred spectral interval was used to establish the quantitative analysis model of lubricant composition and content prediction, and the spectral screening results were shown in Fig 4.

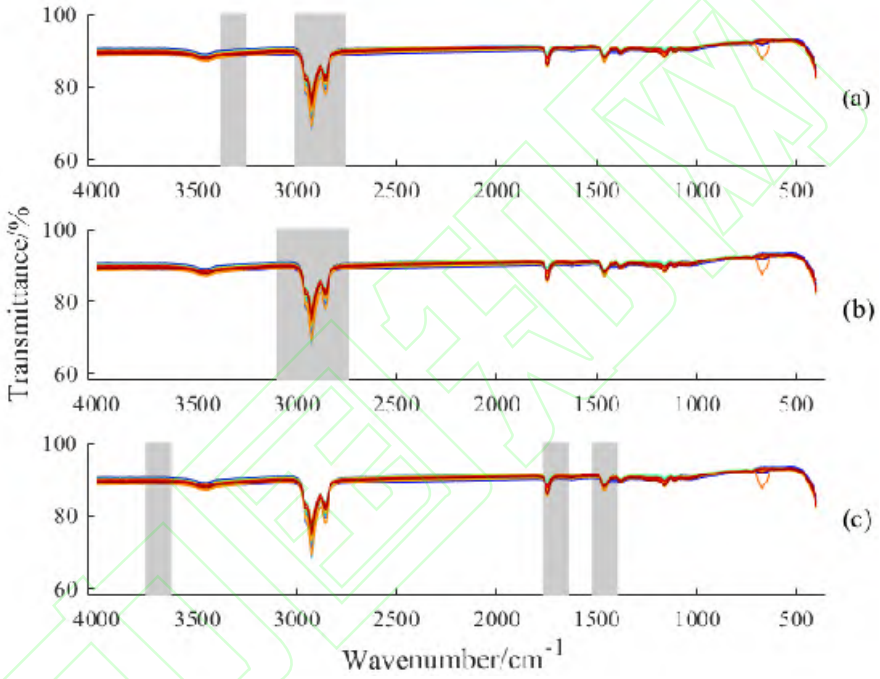

Fig 4. SiPLS screening of characteristic wavenumber distribution of (a) mineral oil, (b) hydrocarbon-based synthetic oil, and (c) polyol ester

As seen in Table 5, among all the corresponding SiPLS spectral interval screening models, For the mineral oil component, when the whole spectrum was divided into 29 subintervals, the best modelling results were selected for the combination of three joint subintervals {20, 21, 24} with a training set RMSECV=1.223 and 192 selected wavenumber points; For hydrocarbon-based synthetic oil components, when the entire spectrum was divided into 20 subintervals, the best modelling results were selected for the combination of 2 joint subintervals {14, 15} with a training set RMSECV= 0.956 and 186 selected wavenumber points; For the polyol ester component, when the whole

spectrum was divided into 29 subintervals, three joint subinterval combinations {9, 11, 27} were selected for the best modelling results with a training set RMSECV= 0.686 2 and 194 wavenumber chosen points.

Table 5 SiPLS Preferred Model Results

Lubricant composition	Total number of intervals	Selected sub-interval	Cross-validation root-mean-square error	Selected wavenumber points
Mineral oil	24	{17, 20}	1.236 0	156
	29	{20, 21, 24}	1.223 0	192
	30	{7, 15, 18, 21}	1.223 0	249
Hydrocarbon-based synthetic oils	20	{14, 15}	0.956 0	186
	27	{17, 19, 22}	0.972 3	207
	30	{18, 19, 21, 22}	0.962 6	248
Polyol esters	22	{6, 9}	0.715 3	170
	29	{9, 11, 27}	0.686 2	194
	26	{8, 10, 13, 16}	0.688 6	288

Among the preferred BiPLS and SiPLS models, SiPLS was better than BiPLS for screening the characteristic spectra of both mineral oil and polyol ester components, with lower RMSECV and fewer enrolled wavenumber points. The screening results for the hydrocarbon-based synthetic oil were consistent with the spectral characteristic band of 2 746.136~3 104.831 cm^{-1}. Therefore, the screening results of the SiPLS model could be used for the secondary screening of characteristic wavenumbers so that the model could achieve better prediction and further eliminate the invalid spectral band points.

4.3 Comparison of binary swarm intelligence algorithms for filtering feature wavenumber points

Three binary intelligent search algorithms, BPSO, BBA, and BGWO, were used for feature point screening of the full-spectrum data (a total of 1 869 wavenumber points) preprocessed in the range of 4 000~400 cm^{-1}, respectively, and the iteration curves of the three algorithms were shown in Fig 5 below. When searching for the corresponding

spectral characteristic wave numbers for different components (mineral oil, hydrocarbon-based synthetic oil, and polyol ester) in lubricants, BGWO was significantly better than the other two algorithms regarding convergence accuracy and stability. At the same time, BPSO had the poorest search capability and tended to fall to the local optimum.

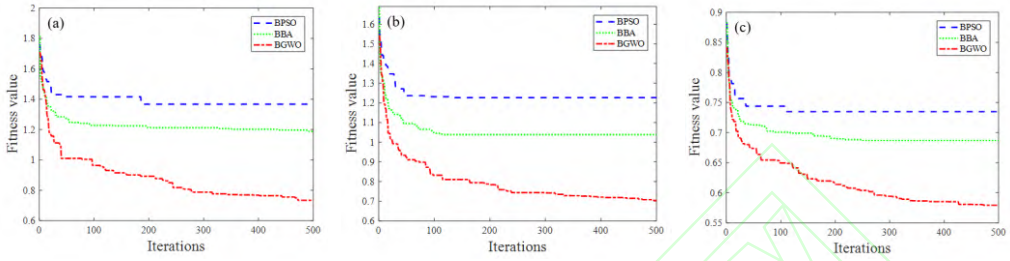

Fig 5. Three iteration curves of binary group intelligence algorithms:

(a) Mineral oil; (b) Hydrocarbon-based synthetic oils; (c) Polyol esters

The results of the three search algorithms for the quantitative analysis of the different components of the lubricant by substituting the selected spectral features wavenumbers into the PLS model were shown in Table 6:

Table 6 Binary intelligence algorithm model prediction results

Lubricant composition	Algorithm	RMSE	Selected wavenumber points
Mineral oil	BPSO	1.366 6	919
	BBA	1.187 4	881
	BGWO	0.733 2	314
Hydrocarbon-based synthetic oils	BPSO	1.226 3	923
	BBA	1.038 2	879
	BGWO	0.703 6	308
Polyol esters	BPSO	0.734 8	884
	BBA	0.686 8	893
	BGWO	0.579 1	346

Table 6 showed that BGWO has a more outstanding search capability, eliminating more invalid feature information than the remaining two methods. The RMSE comparison between BPSO, BBA, and BGWO for different lubricant compositions

showed that BGWO consistently outperforms the others across all lubricant types. For Mineral oil, BGWO achieved the lowest RMSE (0.733 2) compared to BPSO (1.366 6) and BBA (1.1874), with similar trends observed for Hydrocarbon-based synthetic oils (BGWO: 0.703 6, BPSO: 1.226 3, BBA: 1.038 2) and Polyol esters (BGWO: 0.579 1, BPSO: 0.734 8, BBA: 0.686 8). This indicated that BGWO was the most effective for reducing prediction errors, regardless of the lubricant composition were shown in Fig 6.

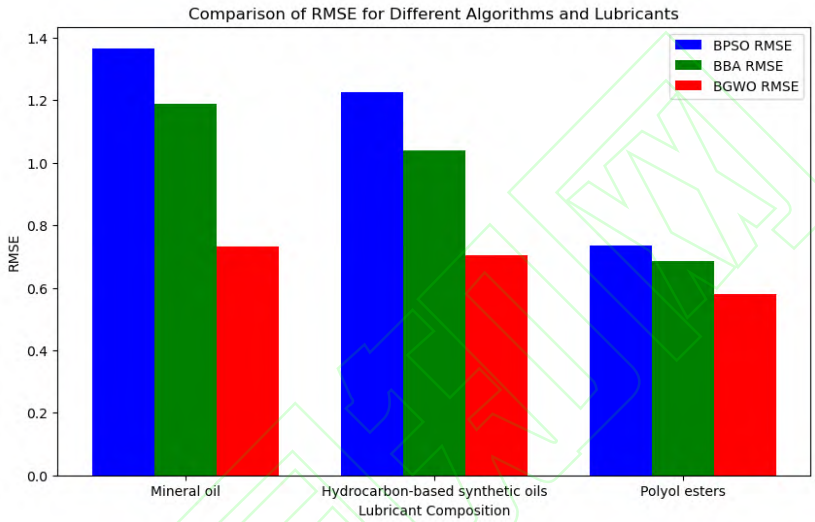

Fig 6. Evaluate and compare RMSE (Root Mean Square Error) values for various algorithms and lubricant types.

BPSO: Binary Particle Swarm Optimization Algorithm; BBA: Binary Bat Algorithm; BGWO: Binary Grey Wolf Optimization Algorithm

The number of selected wavenumber points was below 350, its spectral feature wavenumber distribution was shown in Fig 7. BGWO's prediction was better for quantitative analysis of all three components. BGWO was chosen as a secondary screening method for feature wavenumber finding in the subsequent combination optimization.

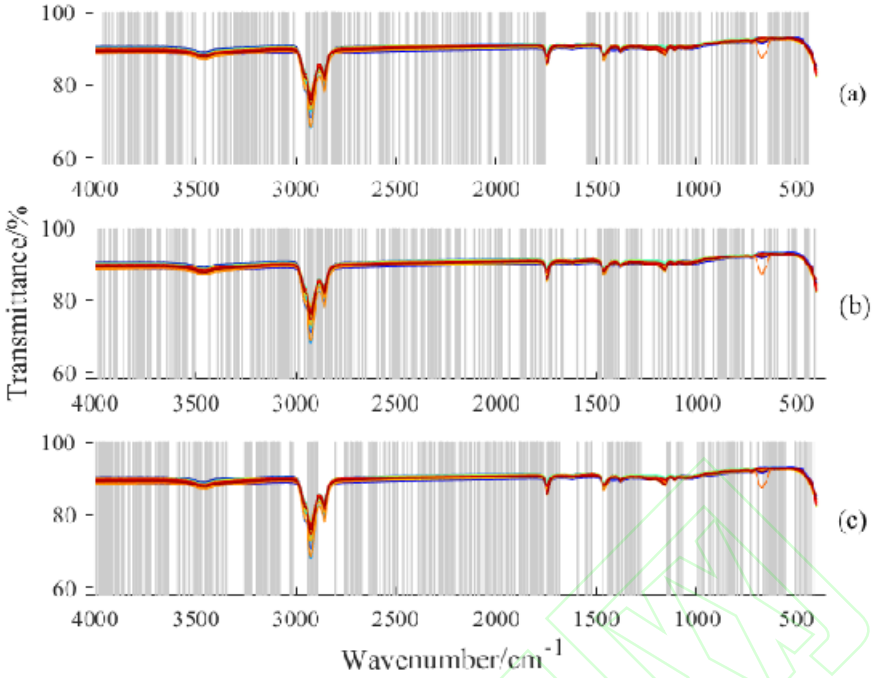

Fig 7. BGWO screening of characteristic wavenumber distribution of (a) mineral oil, (b) hydrocarbon-based synthetic oil, and (c) polyol ester

4.4 Combinatorial optimization screening of characteristic wavenumbers

A mixture of interval band screening and wavenumber point screening of the characteristic wavenumbers of infrared spectra was used to substitute the pre-processed raw spectral data for dimensional reduction and compression. Then, a PLS content prediction model was established. SiPLS-BGWO was used for the characteristic wavenumber screening for the three lubricant components to examine the prediction effect for subsequent computational analysis. As seen in Table 7, the predicted results of SiPLS-BGWO performed well, and the selected spectral features were more obvious in the wavenumber. The use of the SiPLS-BGWO secondary feature wavenumber screening method had improved to varying degrees over the use of a single process, eliminating more invalid feature information and creating a more accurate and effective model.

Table 7 Combinatorial optimization model prediction results

Lubricant composition	Screening Method	RMSE	Selected wavenumber points
Mineral oil	SiPLS-BGWO	0.706 9	39
Hydrocarbon-based	SiPLS-BGWO	0.740 8	30

synthetic oils			
Polyol esters	SiPLS-BGWO	0.654 1	38

The screening of feature wavenumbers using feature compression algorithms such as SiPLS was based on interval division selection, and there was still some interference information selected in the band interval to eliminate redundant information further and reduce the data dimensionality. Based on the above feature band selection algorithm, the selected spectral band data were screened twice by using BGWO, and the distribution of spectral feature wavenumbers was obtained as in Figure 8, 20.3% of the original feature wavenumbers for mineral oil, 16.1% of the original feature wavenumbers for PAO, and 19.6% of the original feature wavenumbers for mineral oil.

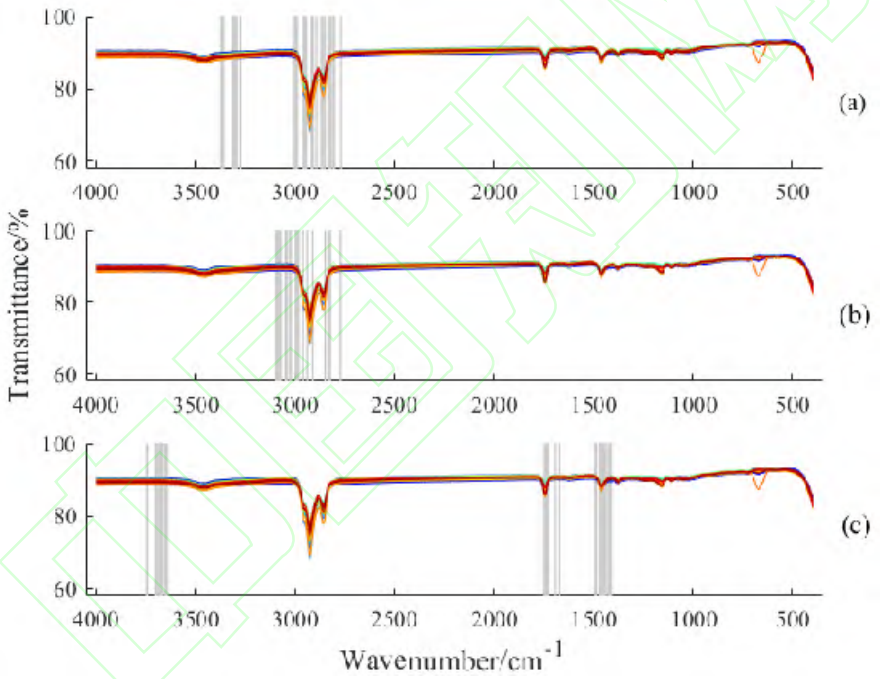

Fig 8. SiPLS-BGWO screening of characteristic wavenumber distribution of (a) mineral oil, (b) hydrocarbon-based synthetic oil, and (c) polyol ester

As a result, the secondary feature wavenumber screening using combinatorial optimization could be used to select the features needed for the model more quickly and effectively and compress the feature dimension. Further analysis of Fig 8, and based on the results of past physical and chemical methods of testing, known aromatic ring C —H stretching vibrations at 3 080 cm^{-1} and —CH_3 and —CH_2 stretching vibrations at 2 924~2 854 cm^{-1} for mineral oils; —CH_3 and —CH_2 stretching vibrations at 2 924~2

854 cm^{-1} for synthetic oil-based; The polyol ester absorbed the aromatic ring C＝O stretching vibration at 1 731 cm^{-1} and the －CH_3 with －CH_2 deformation vibration at 1 465 cm^{-1}. These vibration points are compared with the characteristic wavenumber points in the shaded part of the figure, and a good correspondence could be found. The selected spectral characteristic wavenumber points could reflect the material information of lubricant composition more realistically.

4.5 Analysis of model test results

Using the preferred method model, the SiPLS preferred band and BGWO preferred wavenumber points, and SiPLO-BGWO combined preferred wavenumber points were compared with the full spectral characteristic wavenumbers. The traditional principal component analysis (PCA) compressed characteristic wavenumbers by substituting them into the PLS model. The calculation results were shown in Table 8 below.

Table 8 Comparison of model prediction results

Lubricant composition	Screening Method	MAPE	RMSE	R^2	Selected wavenumber points
Mineral oil	All wavenumbers	3.72%	1.793 3	0.965 6	1 869
	PCA	5.24%	2.885 6	0.916 0	-
	SiPLS	2.23%	1.128 3	0.986 1	192
	BGWO	1.42%	0.733 2	0.993 8	314
	SiPLS-BGWO	1.40%	0.706 9	0.994 3	39
Hydrocarbon-based synthetic oils	All wavenumbers	4.78%	1.836 6	0.973 9	1 869
	PCA	7.09%	2.693 5	0.939 8	-
	SiPLS	2.34%	0.986 8	0.990 6	186
	BGWO	1.79%	0.703 6	0.995 3	308
	SiPLS-BGWO	1.89%	0.740 8	0.994 8	30
Polyol esters	All wavenumbers	3.47%	0.921 5	0.982 3	1 869
	PCA	3.42%	0.975 3	0.980 6	–
	SiPLS	3.11%	0.840 4	0.985 7	194
	BGWO	2.02%	0.579 1	0.993 5	346

	SiPLS-BGWO	2.32%	0.654 2	0.991 9	38

Combining the results of all tests, it could be seen that the data feature dimensionality reduction method using PCA does not improve the model's performance but significantly reduced the model's prediction accuracy. This was because although the use of the PCA algorithm greatly reduces the data dimensionality and improves the model training speed, it also caused a serious loss of information in the original spectral data, which did not better reflect the effective features of the model input. SiPLS can quickly find the feature bands reflecting the corresponding lubricant components and built a relatively well-performing model with fewer inputs. Still, there is a significant difference in various prediction accuracy metrics compared to the model built using BGWO screening features alone. However, the single use of BGWO also has many problems. Firstly, the model training time was long, and the computational load was large; furthermore, the optimized feature wavenumber still contains a lot of irrelevant information, which could not better identify the wave peaks corresponding to the lubricant components on the spectrogram and reflect the correspondence between the spectrum and the substance. These problems could be avoided using the combined Si-BGWO optimization method, which could filter spectral features faster and input more streamlined and effective spectral wavenumbers. However, in the quantitative analysis of some components, the prediction accuracy was slightly different in some indexes compared to the single use of the BGWO method. It was foreseen that the potential of the combinatorial optimization method will be exploited to a greater extent if infrared spectral data with larger data feature sizes were input into the model.

The comparison of different screening methods (All wavenumbers, PCA, SiPLS, BGWO, and SiPLS-BGWO) across three lubricant compositions (Mineral oil, Hydrocarbon-based synthetic oils, and Polyol esters) showed that the SiPLS-BGWO method consistently achieved the best performance. It had the lowest MAPE (1.40% for Mineral oil, 1.89% for Hydrocarbon-based oils, and 2.32% for Polyol esters), the lowest RMSE (0.706 9, 0.740 8, and 0.654 2, respectively), and the highest R^2 values (0.994 3, 0.994 8, and 0.991 9), outperforming the other methods in terms of accuracy

and prediction efficiency were shown in Fig 9.

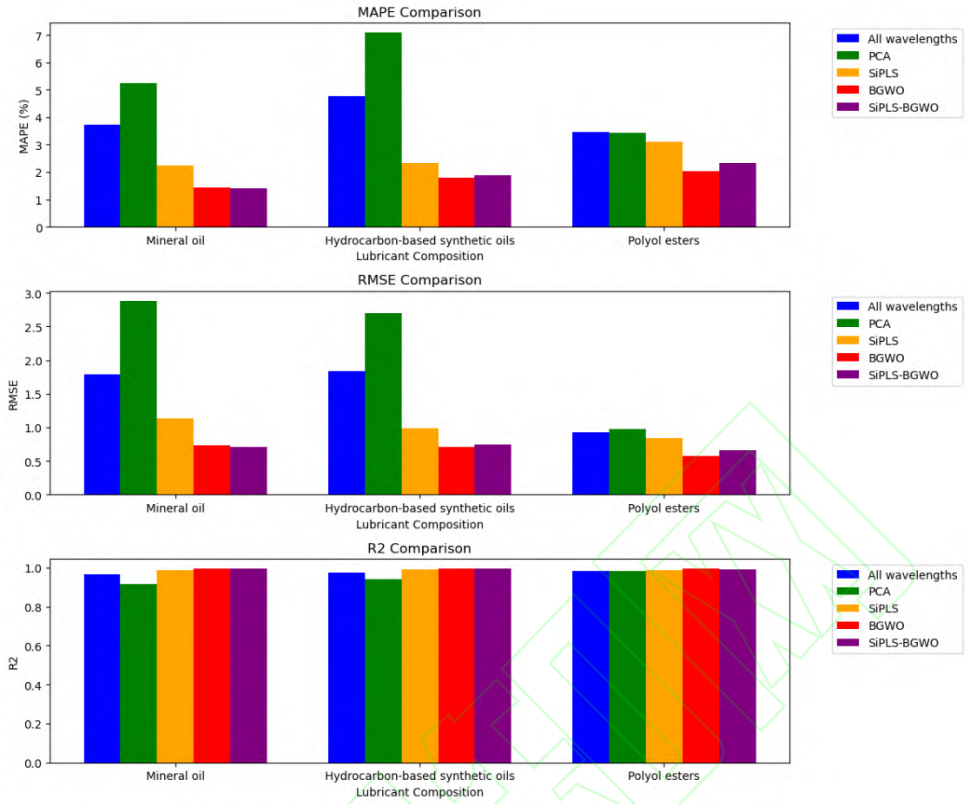

Fig 9. Comparing the screening methods' performance for various lubricant compositions. mean absolute percentage error (MAPE), root mean square error (RMSE), and coefficient of determination (R2).

Thus, for the quantitative analysis of each lubricant component, the SiPLS-BGWO combination optimization method was chosen to compress the features and find better spectral wavenumber points as input to build the model. The predicted results for each component (mineral oil, hydrocarbon-based synthetic oil, and polyol ester) were shown in Fig 10.

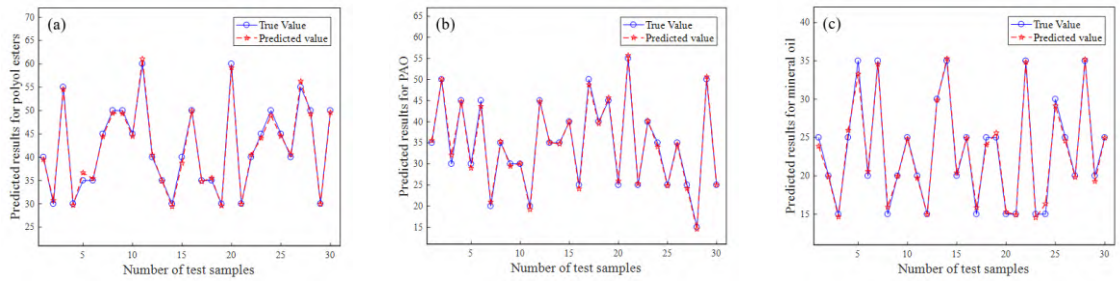

Fig 10. SiPLS-BGWO model prediction results:

(a) Mineral oil; (b) Hydrocarbon-based synthetic oils; (c) Polyol esters

Combining Figure 8 and the above table, it could be seen that the infrared spectral

wavenumber feature point input model optimized and filtered by the SiPLS-BGWO combination had significantly improved the error indicators for the content prediction of mineral oil, hydrocarbon-based synthetic oil, and polyol ester, with the mean relative percentage error (MAPE) all below 2.5%. The root mean square error (RMSE) was reduced by up to 60.58% compared with all spectral wavenumbers. The coefficients of determination, R2, were above 99%. The resulting combined and optimized quantitative analysis model of lubricant composition could quickly find the corresponding spectral bands of the relevant components, which could accomplish the task of content prediction.

5 Conclusion

In this paper, through the problem of rapid and accurate analysis of the content of each component in the infrared spectral test data of lubricating oil, the combination of the interval band screening method (SiPLS) and binary characteristic wavenumber point screening method (BGWO) was proposed to optimize the spectral characteristic wavenumber and the following conclusions were obtained through modelling analysis:

a. The interval band screening method could quickly locate the corresponding spectral bands of the constituent substances and massively reduce the input amount of feature wavenumbers for the model, in which SiPLS could optimize the feature spectra more effectively than BiPLS, with fewer selected feature wavenumber points and more obvious model improvement.

b. Among the three binary group intelligent search algorithms, BGWO had high and stable iteration accuracy and better optimization capability than BBA and BPSO, which tended to converge prematurely with suboptimal solutions. The SiPLS-BGWO approach offered a powerful and efficient technique for quantitatively analyzing lubricant composition. It reduced computational complexity by up to 60.58% and achieved prediction accuracies with R^2 values exceeding 99%.

c. The combined SiPLS-BGWO optimization method proved to be an efficient and powerful technique for quantitative analysis of lubricant composition. By integrating the strengths of both SiPLS and BGWO, the method significantly improved predictive accuracy, as seen with the reduction in RMSE for predicting mineral oil from 1.1283 to

0.7069. Additionally, the number of selected wavenumber points was reduced from 192 to 39, enhancing model simplicity and efficiency. These improvements demonstrate the capability of the SiPLS-BGWO method to optimize feature selection effectively and deliver accuracy for complex spectral data analysis.

References

[1] Zhang Enhui, Li Weimin, Zhao Gaiqing,, et al., The influence of viscosity of mineral oil in the microstructure and performances of lithium-based greases[J]. Tribology, 2022, 42(6): 1258-1266.doi: 10.16078/j.tribology.2021242.

[2] Xia Yanqiu, Xu Dayi, Feng Xin, et al., Identification and prediction of lubricant additives based on extreme learning machines and optimization algorithms[J]. Tribology, 2020, 40(1): 97-106. doi: 10.16078/j.tribology.2019107.

[3] Xie Peiyuan, Feng Xin, Xia Yanqiu. Multi-Label one-dimensional convolutional neural network based on infrared spectroscopy for the identification of additives in lubrication oils[J]. Tribology, 2025. 45(5): 1-12. doi: 10.16078/j.tribology.2024031.

[4] Chiriu D, Pisu F A, Ricci P C, et al. Application of Raman spectroscopy to ancient materials: models and results from archaeometric analyses[J]. Materials, 2020, 13(11): 2456. doi:10.3390/ma13112456.

[5] Pedro F, Daniela H. P, Lucas R. Experimental data and prediction of the physical and chemical properties of biodiesel[J]. Chemical Engineering Communications, 2019, 206(10): 1273-1285. doi: 10.1080/00986445.2018.1555533.

[6] Lazzari E, Souza Silva É A, Bjerk T R, et al. Evaluation of the matrix effect in the quantitative bio-oil analysis by gas chromatography[J]. Fuel, 2021, 290(2): 119866. doi:10.1016/j.fuel.2020.119866.

[7] Feng Xin, Xia Yanqiu, Xie Peiyuan, et al. Classification and spectrum optimization method of grease based on infrared spectrum[J]. Friction, 2024, 12(6): 1154–1164. doi:10.1007/s40544-023-0786-y.

[8] Tan Chao, Chen Hui, Lin Zan. Brand classification of detergent powder using near-infrared spectroscopy and extreme learning machines[J]. Microchemical Journal, 2021, 160(7): 105691. doi:10.1016/j.microc.2020.105691.

[9] Xu Jigang, Liu Shujun, Gao Ming, et al. Classification of lubricating oil types using mid-infrared spectroscopy combined with linear discriminant analysis–support vector machine algorithm[J]. Lubricants, 2023, 11(6): 268. doi:10.3390/lubricants11060268.

[10] Xia Yanqiu, Wang Chen, Feng Xin. GA-BPSO hybrid optimization mid-infrared spectral characteristic band screening lubricant additive type identification technology[J]. Tribology, 2022, 42(1): 142-152. doi: 10.16078/j.tribology.2020164.

[11] Wang Yuesen, Chen Yu, Liang Xingyu, et al. Impacts of lubricating oil and its formulations on diesel engine particle characteristics[J]. Combustion and Flame, 2021, 225: 48–56. doi:10.1016/j.combustflame.2020.10.047.

[12] Costa M C A, Morgano M A, Ferreira M M C, et al. Quantification of mineral composition of Brazilian bee pollen by near infrared spectroscopy and PLS regression[J]. Food Chemistry, 2019, 273: 85–90. doi:10.1016/j.foodchem.2018.02.017.

[13] Zhao Na, Wu Zhisheng, Wu Chunying, et al. Performance evaluation of variable selection methods coupled with partial least squares regression to determine the target component in solid samples[J]. Journal of Near Infrared Spectroscopy, 2022, 30(4): 171–178. doi:10.1177/09670335221097236.

[14] Feng Xin, Xia Yanqiu. Research progress on intelligent identification of lubricating oil based on infrared spectroscopy technology[J]. lubricating oil, 2024. 39(1): 38-42. doi：10.19532/j.cnki.cn21-1265/tq.2024.01.008

[15] Song Xiangzhong, Du Guorong, Li Qianqian, et al. Rapid spectral analysis of agro-products using an optimal strategy: dynamic backward interval PLS–competitive adaptive reweighted sampling[J]. Analytical and Bioanalytical Chemistry, 2020, 412(12): 2795–2804. doi:10.1007/s00216-020-02506-x.

[16] Jiang Weiwei, Lu Changhua, Zhang Yujun, et al. Molecular spectroscopic wavelength selection using combined interval partial least squares and correlation coefficient optimization[J]. Analytical Methods, 2019, 11(24): 3108–3116. doi:10.1039/C9AY00898E.

[17] Li Yuanpeng, Fang Tao, Zhu Siqi, et al. Detection of olive oil adulteration with waste cooking oil via Raman spectroscopy combined with iPLS and SiPLS[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2018, 189: 37–43. doi:10.1016/j.saa.2017.06.049.

[18] Abou El-Alamin M M, Sultan M A E, Hegazy M, et al. Pure component contribution (PCCA) and synergy interval partial least squares (siPLS) algorithms for efficient resolution and quantification of overlapped signals; an application to novel antiviral tablets of daclatasvir, sofosbuvir and ribavirin[J]. European Journal of Chemistry, 2019, 10(4): 350–357. doi:10.5155/eurjchem.10.4.350-357.1899.

[19] Ran Zhiyong, Sun Laijun, Liu Yangyang, et al. Forward and backward interval partial least squares method for quantitative analysis of frying oil quality[J]. Infrared Physics & Technology, 2020, 105: 103207. doi:10.1016/j.infrared.2020.103207.

[20] Tan Jiahua, Sun Yan, Ma Li, et al. Knowledge-based genetic algorithm for resolving the near-infrared spectrum and understanding the water structures in aqueous solution[J]. Chemometrics and Intelligent Laboratory Systems, 2020, 206: 104150. doi:10.1016/j.chemolab.2020.104150.

[21] Mrityunjay G, Nivedita D, Debdeep M et al., A novel quantum algorithm for ant colony optimisation[J]. IET Quantum Communication, 2022. 3(1): 13-29. doi: 10.1049/qtc2.12023.

[22] Abbes W, Kechaou Z, Hussain A, et al. An enhanced binary particle swarm optimization (E-BPSO) algorithm for service placement in hybrid cloud platforms[J]. Neural Computing and Applications, 2023, 35(2): 1343–1361. doi:10.1007/s00521-022-07839-5.

[23] Al-Dyani W Z, Ahmad F K, Kamaruddin S S. Binary bat algorithm for text feature selection in news events detection model using Markov clustering[J]. Cogent Engineering, 2022, 9(1). doi:10.1080/23311916.2021.2010923.

[24] Kaur A, Kumar Y. Recent developments in bat algorithm: a mini review[J]. Journal of Physics: Conference Series, 2021, 1950(1): 012055. doi:10.1088/1742-6596/1950/1/012055.

[25] Momanyi E, Segera D. A master-slave binary grey wolf optimizer for optimal feature selection in biomedical data classification[J]. BioMed Research International, 2021, 2021 (3):1-12. doi:10.1155/2021/5556941.

[26] Xia Yanqiu, Wang Chunli, Feng Xin et al. Prediction of lubricating oil friction and wear performance based on grey wolf algorithm optimization of GRNN[J]. Tribology, 2023. 43(8): 947-955. doi: 10.16078/j.tribology.2022121.

[27] Xia Yanqiu, Wang Yuxing, Feng Xinet al. Optimization efficiency of swarm intelligence

search in base oil performance prediction model[J]. Tribology, 2023. 43(4): 429-438. doi: 10.16078/j.tribology.2021304.

[28] Tian Han, Zhang Linna, Li Ming, et al. Weighted SPXY method for calibration set selection for composition analysis based on near-infrared spectroscopy[J]. Infrared Physics & Technology, 2018, 95: 88–92. doi:10.1016/j.infrared.2018.10.030.

[29] Cardoso D O, Galeno T D. Online evaluation of the Kolmogorov–Smirnov test on arbitrarily large samples[J]. Journal of Computational Science, 2023, 67(2): 101959. doi:10.1016/j.jocs.2023.101959.

[30] Zeng Ganning, Ma Yuan, Du Mingming, et al. Deep convolutional neural networks for aged microplastics identification by Fourier transform infrared spectra classification[J]. Science of the Total Environment, 2024, 913(5): 169623. doi:10.1016/j.scitotenv.2023.169623.

[31] Ma Xiaohui, Chen Zhengguang, Liu Jinming. Wavelength selection method for near-infrared spectroscopy based on Max-Relevance Min-Redundancy[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2024, 310: 123933. doi:10.1016/j.saa.2024.123933.