IOP Publishing Measurement Science and Technology
Meas. Sci. Technol. **31** (2020) 045501 (9pp) https://doi.org/10.1088/1361-6501/ab5f74

Comparative study of partial least squares and neural network models of near-infrared spectroscopy for aging condition assessment of insulating paper

Yuan Li[1], **Yin Zhang, Wen-Bo Zhang, Yao-Yu Xu and Guan-Jun Zhang**

State Key Laboratory of Electrical Insulation and Power Equipment, Xi'an Jiaotong University, Xi'an, People's Republic of China

E-mail: liyuan8490@xjtu.edu.cn

Received 27 September 2019, revised 29 November 2019
Accepted for publication 6 December 2019
Published 23 January 2020

Abstract

Near-infrared spectroscopy (NIRS) is a rapid and non-destructive detection method for component determination and quantitative analysis with broad applications in numerous fields. In recent years, NIRS has started to be used in the aging condition assessment of power transformers. However, the real applications of NIRS are constrained by the lack of evaluation database and accurate prediction algorithms. In this paper, we aim at comparing different NIRS modeling methods and improving diagnostic accuracy. We build the evaluation database via the preparation of 230 specimens derived from three typical types of insulating paper. Calibration models are established by linear method-partial least squares (PLS) and nonlinear method-back propagation neural network (BPNN) to map the relationship between spectra and the degree of polymerization (DP). The DP prediction results show that using full NIR spectra as the input of the PLS model does not ensure a high prediction accuracy, and it is improved by competitive adaptive reweighted sampling (CARS) that selects the optimal wavelength combinations. Prediction precisions given by BPNN and CARS-BPNN models are shown to be less satisfactory than that of CARS-PLS. We process the original spectra with principal component analysis (PCA) as the input of BPNN and the PCA-BPNN model realizes high prediction precision for three types of paper (RMSE $\leqslant 24$, $r = 0.99$). With the identification of paper type by the k-nearest neighbors (KNN) method before prediction, the KNN-PCA-BPNN model solves the problem of the low prediction precision for mixed (unknown) paper samples (RMSE $= 36$, $r = 0.98$), which facilitates future field tests as well as related applications in practice.

Keywords: insulating paper, aging condition, near infrared spectroscopy, partial least squares, back propagation neural network

(Some figures may appear in colour only in the online journal)

1. Introduction

A power transformer is one of the most expensive assets and plays a vital role in electric power transmission and distribution (T&D) systems [1]. A failure of a power transformer may lead to substantial costs and cause detrimental social impacts, e.g. power outage and complaints [2]. Oil-paper insulation inside the transformer will be gradually aged under synergistic effects such as electrical, thermal, mechanical and chemical stresses during the long-term operation, during which the probability of equipment failure gradually increases [3]. The remaining lifetime of a transformer is determined by the condition of the solid dielectrics, i.e. insulating paper wrapped on the windings, because the insulating oil in the transformers can

[1] Author to whom any correspondence should be addressed.

 © 2020 IOP Publishing Ltd Printed in the UK

be replaced or easily filtrated on site while the aging processes of the insulating paper is irreversible [4]. Among the dozens of aging-related parameters, the degree of polymerization (DP), that a factor approximately is equal to the average cellulose molecules chain length of insulating paper, is considered as the most direct and reliable chemical parameter characterizing the aging conditions of insulating paper. DP is therefore determined by the IEEE guidelines as the basis for the quantitative measurement of the aging conditions of insulating paper [5].

As the growing requirement for the reliability of power supply, the quick view of the operating condition has prevailed and the interval of planned outage for maintenance is required less than ever. The conventional measurement of the DP of insulating paper is called the viscometric method introduced by ASTM D4243 [6]. Though it provides amenable results, all tests should be conducted in a laboratory after field sampling, which is not only a destructive procedure for the insulation of the transformer [7] but also time consuming. Therefore, the non-destructive field assessment of the DP of insulating paper is in great need and might become a powerful tool for condition assessment.

Near-infrared spectroscopy (NIRS) has many advantages, like non-destructive detection and rapid processing [8]. It has been widely employed in agriculture, pharmaceutical and food industries, petrochemical engineering, etc in the past few decades. Until recent years, the power industry has gradually started to introduce the NIRS technique as an alternative to traditional laboratorial tests for the quantitative analysis of cellulose paper due to its great advantages.

Initial explorations using NIRS to assess the aging conditions of oil-paper insulation have been conducted in recent years and remarkable progress has been made for practical implementations. Ali *et al* [9] applied the chemometric method to the near-infrared (NIR) spectrum of aged kraft paper to extract the spectral characteristic parameters. Their chemometric model predicted the aging time of kraft paper with an error of 95 h for samples up to 3000 h of ageing. Santos *et al* [10] collected the diffuse-reflectance spectra between 1260 and 2500 nm from insulating paper with varying aging conditions, and established the assessment models by multiple linear regression (MLR) and partial least squares (PLS) methods, respectively. The established PLS model using the first-order derivation of spectra suggests a better prediction performance than multivariate calibration. Baird *et al* [11, 12] developed a portable fiber-optic spectroscopic system with multivariate statistical analysis to measure the DP, water and oil content of insulating paper non-destructively. Several field experiments on power transformers have been performed to verify the practicability of the system. Li [13] found that insulating oil within the paper has great influences on NIR spectra and the DP prediction accuracy of paper can be significantly improved after oil removal procedures.

Theoretically, the absorbance of NIR spectra has a linear connection with the chemical composition of material in a certain range according to Kubelka–Munk law [14]. However, for the field disassembly tests of the transformer, insulating paper wrapped on the windings is inevitably immersed with oil although oil drain-off is a standard procedure before the tests. The acquired NIR spectra comprise the vibrational information of both paper and oil, causing the nonlinear perturbation of absorption coefficient and scattering coefficient. Therefore, NIR spectra are difficult to be directly interpreted and quantified by the linear analysis methods [10, 15]. In recent years, intelligent algorithms, mostly nonlinear methods such as back propagation neural network (BPNN), have shown promising applications in the component determination of soil and seed, the quality analysis of fruit, etc [16], providing potential solutions to solve the nonlinear problems of quantitative analysis.

Up to now, however, attempts of quantitative analysis for insulating paper in a transformer by nonlinear methods are still insufficient and the comparison between linear and nonlinear methods is less unveiled. In this paper, we aim at establishing DP prediction models of NIRS to assess the aging conditions of insulating paper and comparing the different advantages of linear and nonlinear methods by revealing their operating processes. The comparative study would improve diagnostic accuracy of aging assessment through optimizing both methods.

The paper structure is organized as follows. Three typical types of insulating paper are prepared and raw spectra as well as DP tests of the insulating paper are subsequently performed to construct the paper database of varying aging conditions in section 2. We introduce the PLS method, a widely used linear method in NIR spectra quantitative analysis, to establish the aging condition calibration model of insulating paper in section 3. As a typical nonlinear method, the BPNN calibration model is established to compare the prediction performance of two modeling algorithms in section 4. We summarize the modeling results and draw the conclusion in section 5.

2. Sample preparation and experimental measurement

2.1. Collections of aging samples

The types of insulating paper used in the transformer manufacturing are diverse. We choose three typical kinds of insulating paper, i.e. kraft paper (BZZ-75), crepe paper (58HC) and thermally upgraded paper (22HCC) to characterize a broad range of paper sources employed in the 110–1000 kV transformers. The thermally accelerated aging of paper samples is conducted in the presence of oil in a vacuum-heat oven to obtain oil-immersed paper samples.

The oil has undergone a regular filtering and degassing procedures before used in the experiments. The moisture and gas dissolved in oil can be removed via the vacuum pump, which allows the vacuum degree of the cabinet under 50 Pa. The insulating paper is dried in a vacuum oven at 105 °C for 24 h. The moisture in the oil and paper are less than 10 mg kg^{-1} and 1% in mass fraction, respectively. Paper samples are then immersed with oil by the weight ratio of 1:10 and stored in a vacuum at 90 °C for another 24 h to guarantee that the paper has fully impregnated with oil [17]. Subsequently, the paper and oil are transferred into glass bottles with the same weight ratio. Thermally accelerated aging procedure are implemented at 120 °C [18, 19]. All the experiments are conducted in our laboratory under ambient control, e.g. temperature of around 20 °C and relative humidity of 65%.

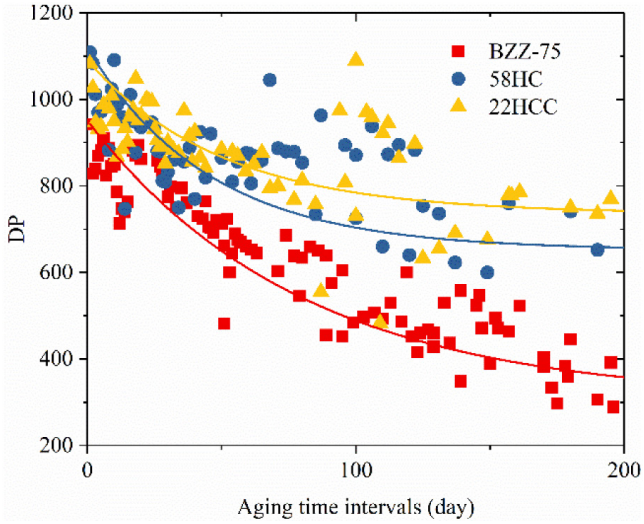

Figure 1. DP of three types of paper samples as a function of aging time. The fitted curves indicate that the DP follows a decreasing exponential function. Note that the colors of the fitted curves are the same with the colors of data points of corresponding paper types.

We obtain the oil-immersed paper samples with varying aging conditions by collecting bottles at different intervals from the oven. The sampling frequency depends on the decreasing speed of DP. A high frequency of the sample collection is necessary in the initial aging stage as the DP of paper samples decreases quickly, e.g. daily sample collection until the DP of the kraft paper falls to 700 (the kraft paper has a faster aging speed than the other two types of paper). When the DP of the paper samples decreases to 500, the intervals between two consecutive sample collections are widened to 3 days or more. In total, 230 aged samples are obtained in the experiment before the DP of the kraft paper falls to about 300, a threshold indicating that a transformer is severely aged and should be replaced at the end of its service lifetime.

The DP of the paper samples is measured by viscometry according to ASTM D4243, which consists of a series of procedures, such as degreasing, abrading, dissolving and viscosity measurement. The measurement results of DP are a function of aging time as shown in figure 1. Note that each data point of DP is averaged by two independent DP measurements of paper and the data is acceptable when the deviation is less than 2.5% of the mean value. The differences in decreasing speed of DP for three types of paper are pronounced and the thermally upgraded paper shows a strong decomposition resistance due to chemical modification techniques [20]. The DP of three types of paper decreases rapidly in the early phase and the descending rate slows down when the aging time increases, as indicated by fitted curves in figure 1. The range of DP variation is different under the influence of pulp types, manufacturing processes and the individual difference of aging samples.

2.2. Acquisition of NIR spectra

As mentioned above, the insulating paper of the disassembled transformer is immersed with oil although oil drain-off is a regular procedure in the field tests. This means that the spectral scanning and analysis of oil-immersed paper may have a greater significance in application if quantitative analysis can deal with overlaid spectral information with reliable accuracy. Therefore, in the experiment, we acquire the spectra of aged paper samples coupled with the spectra of oil and use them as the input for further modeling. Note that for the scanned oil-immersed paper in the lab, the oil contents of paper are controlled roughly under 35% by mass ratio to maintain desired homogeneity between tested samples. The oil contents in the paper are close to that of the outer layer paper in the field disassembled transformer (after oil drain-off).

A high-performance NIR spectrometer with the spectral detection range of 895–2202 nm and resolution of 5 nm is employed in the measurement. An InGaAs detector with 256 linear arrays is the main working component inside the spectrometer, and the background noise can be reduced by the imbedded auto-zero function via software and cooling the detector through a hardware set. A standard whiteboard is scanned to obtain the reference spectrum before spectral scanning of paper samples for the purposes of instrument baseline calibration and systematic error elimination. Each acquired spectrum is determined by averaging 32 times of spectral scanning at the same position of insulating paper to reduce the deviations caused by manual and ambient interferences.

The acquired NIR spectra of three types of oil-immersed paper under varying aging conditions are plotted in figure 2. Generally, the spectral differences between three types of paper are visually minor except several variations of absorbance peaks in the scope of 1700–2000 nm in wavelength. Further, each spectrum of one type of paper is too similar to identify the aging conditions just by reading the spectrum. In other words, it is impossible to have a clear connection merely by comparing the absorbance peaks of NIR spectra to the aging time or aging conditions. Consequently, developing chemometric methods and quantitative analysis are essential to extract hidden information and identify the aging states of oil-immersed paper.

3. Typical linear model: PLS method

In order to obtain the quantitative relationship between the absorbance of spectra and physicochemical properties (DP in this paper), the calibration model is usually established by a multivariate statistical analysis method with paper samples in the training set [12]. The prediction ability of the calibration model is then verified with paper samples in a testing set and statistical parameters are calculated for quantitative evaluation and comparisons.

3.1. Basic PLS procedures and modeling

PLS is a convenient and effective method for relating two data matrices by a linear multivariate model that has been widely used in the quantitative analysis of NIR spectra [21]. We develop the PLS calibration model from the training set of aging paper samples to reflect the linear relationship between the spectra (Matrix X) and aging conditions (Matrix Y).

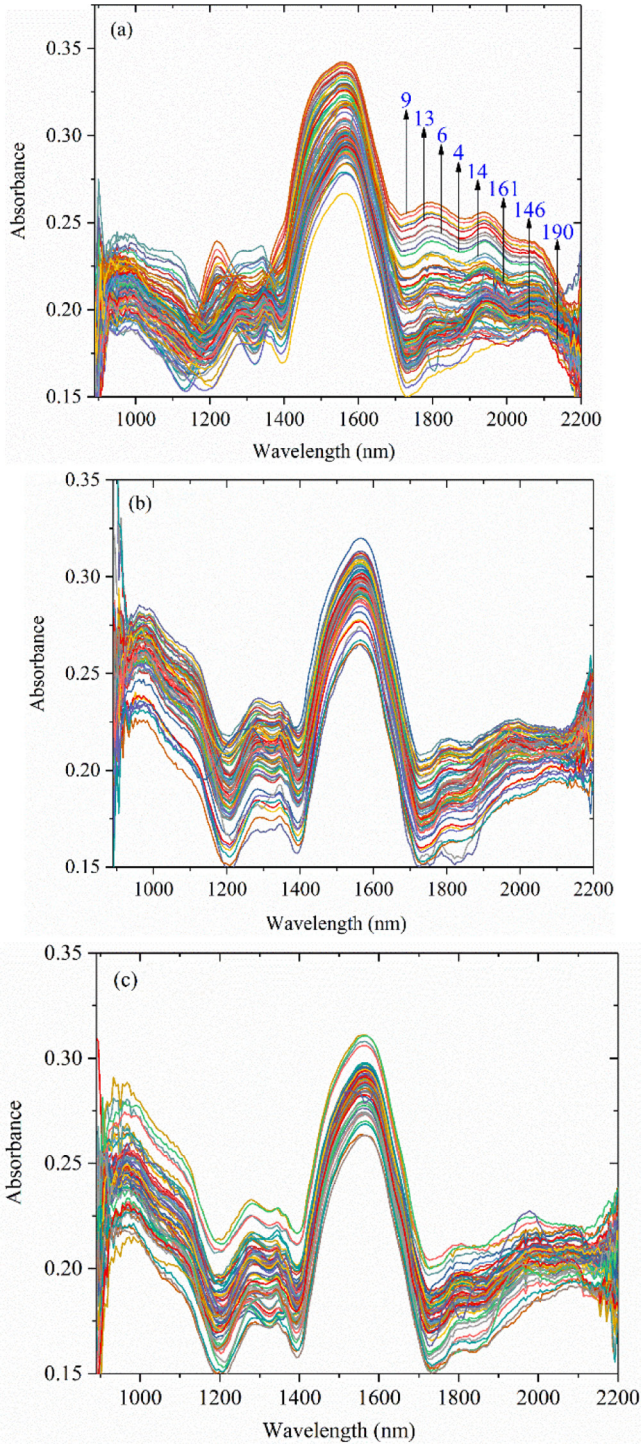

Figure 2. NIR spectra of three types of paper samples. The number of spectra in (a), (b) and (c) is 95, 67 and 68, respectively (total 230). An example of aging durations (in days) of the selected spectra are marked out in figure (a), but the correlation between aging conditions and absorbance of NIR spectra is not clear.

Matrix X consists of the original spectral data of different aging paper samples and each of the spectra owns the absorbance of 254 wavelengths. Matrix Y contains the DP of aging paper samples. Note that here Y is a matrix of $n \times 1$, therefore we actually use PLS1 method (for univariate response) in this paper. Both matrix X and Y are transformed via scaling and

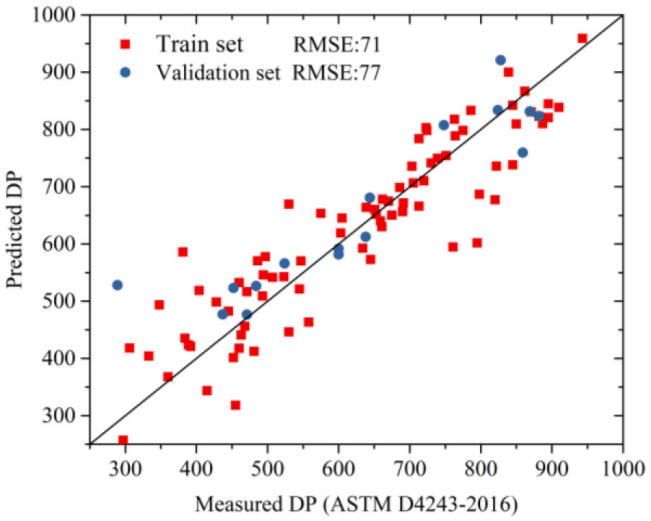

Figure 3. Prediction of PLS calibration model for BZZ-75 paper samples.

centering to increase the distinctions between samples. The derived principal components (PCs) of PLS by a cross-validation method are orthogonal to eliminate redundant information, noting that the original spectra have high collinearity. X is decomposed by equations (1). The linear relationship between X and Y is built through equation (2) in a stepwise regression manner. Therefore, the extracted PCs are primarily related to Y (i.e. DP):

$$X = TP' + E = \sum_{j=1}^{a} t_j p_j' + E \tag{1}$$

$$Y = TC' + F = XB + F \tag{2}$$

where a stands for the number of PCs, T for the score matrix, P for the loading matrix of X, and E for the matrix of X residuals. C for Y-weight matrix, F for the matrix of Y-residuals, and B is the matrix of regression coefficients of Y.

Root mean square error (RMSE) and correlation coefficient r are two independent criterions to characterize the prediction performance of calibration models [15]. RMSE is the quality standard to estimate the training network as calculated through equation (3). r is calculated to evaluate the statistical relationship between measured DP by viscometry and predicted DP by the spectral calibration model via equation (4). The optimal topology parameters are determined under the conditions of low RMSE and high r for all the selected sets at the same time:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(y_{i,act} - y_{i,predi}\right)^2}{n}} \tag{3}$$

$$r = \sqrt{1 - \frac{\sum_{i=1}^{n} \left(y_{i,act} - y_{i,predi}\right)^2}{\sum_{i=1}^{n} \left(y_{i,act} - y_{ave,act}\right)^2}} \tag{4}$$

where n stands for the number of paper samples in the selected set, e.g. the training set, validation set or testing set. $y_{i,act}$ stands for DP of ith sample measured by viscometry, $y_{ave,act}$ for the

Table 1. The prediction of different PLS calibration models.

Type	Algorithm	Number of PCs	Training set			Testing set		
			Sample Size	RMSE	r	Sample Size	RMSE	r
BZZ-75	PLS	12	79	71	0.91	16	77	0.92
	CARS-PLS	12		67	0.92		68	0.94
58HC	PLS	10	60	70	0.78	7	32	0.92
	CARS-PLS	10		74	0.75		32	0.93
22HCC	PLS	20	60	54	0.90	8	48	0.61
	CARS-PLS	20		59	0.87		21	0.93
Unclassified mixed samples	PLS	24	200	94	0.86	30	100	0.85
	CARS-PLS	24		90	0.88		62	0.89

mean DP of paper samples in the selected set, and $y_{i,predi}$ for DP of ith paper sample predicted by the corresponding calibration model.

A first attempt of the PLS method by the aforementioned modeling using the full spectrum is applied to kraft paper (BZZ-75) and the prediction performance is shown in figure 3. We can find that the prediction of the BZZ-75 paper samples has remarkable error (RMSE = 77, $r = 0.92$). A reasonable cause is that the full spectra of the aging paper samples include all the wavelengths that may contribute more collinearity, redundancies and noise than characteristic aging-relevant information to the PLS model. In other words, it indicates that not all the wavelengths in the spectrum are strongly related to the DP prediction. Table 1 lists prediction performance of all three types of insulating paper, and the prediction accuracy is unsatisfactory, hence further optimized methods are required.

3.2. Model optimization via CARS

The above attempts indicate that selecting characteristic wavelengths strongly related to the DP of paper samples is a potential solution to improve the prediction performance of the PLS model. Santos *et al* [10] employed the successive projection algorithm (SPA) to select wavelength variables. The prediction results show that the application of the modified SPA improves the root mean square error of cross-validation (RMSECV), but the RMSE obtained after external testing is slightly higher than PLS. In this paper, we use the competitive adaptive reweighted sampling (CARS) to enhance characteristic wavelength selection as it proves good performance in the wavelength selection of NIR spectra compared with some traditional methods [22].

The essence of CARS is to determine optimal wavelength combination so as to construct the PLS model with the lowest RMSECV. CARS procedures for wavelength combination selection of NIR spectra are shown in figure 4. For the first sampling run, all the absorbance of the full spectra is extracted and the PLS model is established using the extracted data. Subsequently, the RMSECV of the PLS model is calculated and the regression coefficients of the specific wavelengths are available as well. The number of selected wavelengths is characterized by the ratio of the selected wavelengths over all wavelengths, r_k. As a matter of fact, r_k is derived from the exponentially decreasing function as given by equation (5):

$$r_k = ae^{-bk} \tag{5}$$

where r_k stands for the ratio of the selected wavelengths over all wavelengths in the kth iteration ($1 \leqslant k \leqslant N$), and constants a and b are calculated according to equations(6)–(7) respectively:

$$a = \left(\frac{p}{2}\right)^{1/(N-1)} \tag{6}$$

$$b = \frac{\ln(p/2)}{N-1} \tag{7}$$

where p is the number of wavelengths for full spectrum ($p = 254$ in our experiments), and N is the default times of sampling runs.

The wavelengths (amount: $254 \times r_k$) with high regression coefficients are chosen for the next sampling run and the other wavelengths are eliminated. Eventually, the optimal wavelength combination is obtained in an iterative and competitive manner until the sampling run times reach the default number ($N = 50$ in our practices). Note that the established PLS model using optimal wavelengths has the lowest RMSECV.

Based upon the above procedures, the CARS-PLS calibration model is established and the prediction performance of different paper samples is listed in table 1. For each type of insulating paper, we select approximately 85% of its total sample size as the training set and the left samples form the testing set. The selection of samples are implemented by Kennard–Stone (K–S) method [23]. It is clear that the accuracy of the CARS-PLS model is higher than traditional PLS method in all the cases except that they share the same RMSE (32) of 58HC specimens but CARS-PLS has a higher r. It is worth mentioning that the CARS-PLS model has higher accuracy, even the mixed samples including all paper samples are taken into account, however the correlation with DP measured by viscometry is not sufficiently high for field application ($r < 0.9$).

The fluctuation of RMSE and r depends on the types of the paper samples and the calibration modeling methods. As can be seen from table 1, these two criterions fluctuate among three types of paper although the same modeling methods are applied. The fluctuation can be suppressed via optimizing algorithm framework and topology parameters to obtain reliable prediction performance, namely low RMSE and, meanwhile, high r. In this sense, the CARS-PLS method is an optimized calibration model.

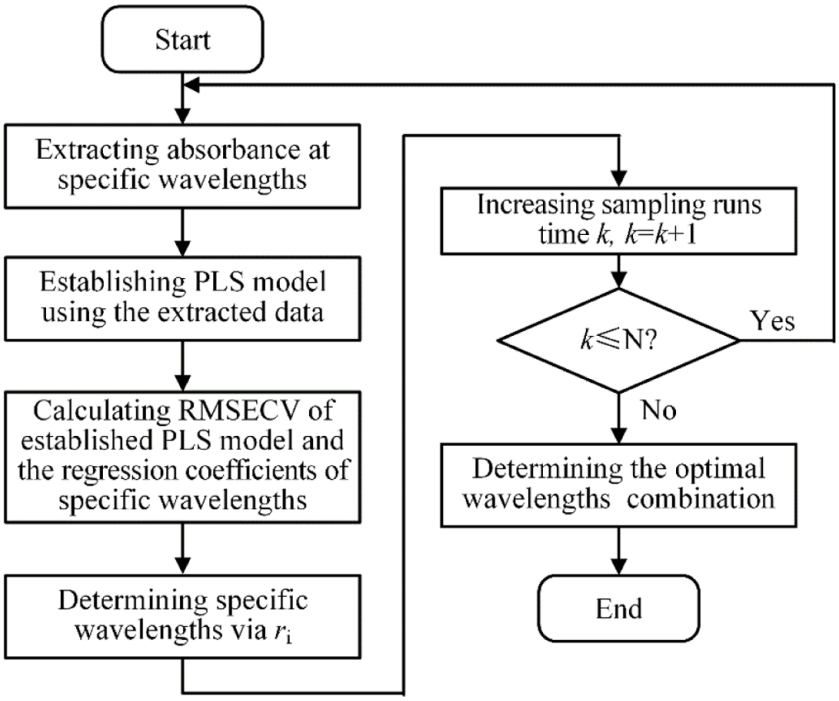

Figure 4. CARS selection procedures of optimal wavelength combination.

4. Nonlinear modeling via BPNN

4.1. Basic BPNN theory and experiments

Although the optimized linear model CARS-PLS has shown better prediction performance in DP evaluation than the traditional PLS model, it is still unable to overcome the influence of the coupled spectra in the field tests as mentioned above. Thus the nonlinear modeling method, typically as BPNN, to bridge the DP and NIR spectra is a potential solution that addresses these ambient influences in a self-adapted way.

BPNN is capable of mapping multidimensional relationships and has been one of the most studied and used algorithms for neural network learning ever since [24]. The essence of achieving accurate nonlinear mapping between NIR spectra and DP by BPNN is to look for the minimum of the error function in the weight space, which is accomplished via the gradient descent method in this paper.

The most widely used BPNN model contains three layers, i.e. a classical input-hidden-output topology, and can approximate most of the functions with sufficient freedom degrees. We use all wavelengths of full spectra as the input parameters I_j for the experiments of the BPNN model. The transferred data of the kth node in hidden layers, h_k, is calculated based on equation (8) [25, 26]:

$$h_k = s_1\left(\sum_{j=1}^{m} W^i_{1,k,j} \times I_j + b_1\right) \tag{8}$$

where s_1 is a tangent hyperbolic activation function between the input layer and the hidden layer, and $s_1(x_1) = (1 - \exp(-2x_1))/(1 + \exp(-2x_1))$ in which x_1 is an arbitrary variable. $W^i_{1,k,j}$is the weight of the jth node in the input layer connecting the kth node in the hidden layer at the ith iteration. m is the node number in the input layer and b_1 is the bias between the input and the hidden layer.

The predicted DP of insulating paper, α, is calculated by transferring the information between the hidden layer and the output layer via equation (9):

$$\alpha = s_2\left(\sum_{k=1}^{n} W^i_{2,k} \times h_k + b_2\right) \tag{9}$$

where s_2 is a linear activation function between the hidden layer and the output layer, and $s_2(x_2) = x_2$ in which x_2 is an arbitrary variable; $W^i_{2,k}$is the weight of the kth node in the hidden layer connecting the output layer at the ith iteration; i is the iteration times. n is the node number in the hidden layer and b_2 is the bias between the hidden and the output layer.

The default node number n in our experiment is 10 and the output layer exports one node representing the predicted DP by the model. Based on the above description, the BPNN model for the DP prediction of three types of insulating paper is established and the prediction performance is listed in table 2. Looking back at the PLS model in table 1, we can find that the prediction performance of BPNN is even worse than that of the PLS method (RMSE 87 versus 48 for 22HCC paper), much less prevails the CARS-PLS model. It is inferred as an essential point to improve the prediction performance that spectral information related to the aging conditions, instead of full spectra, should be extracted before being imported into the BPNN model.

4.2. CARS-BPNN model and PCA-BPNN model

The fact that the CARS-PLS model achieves good performance as shown in section 3 suggests that selecting optimal wavelength combinations and extracting principal components

Table 2. Prediction of different BPNN calibration models.

Type	Algorithm	Training set			Validation set			Testing set		
		Sample Size	RMSE	r	Sample Size	RMSE	r	Sample Size	RMSE	r
BZZ-75	BPNN	67	49	0.96	14	75	0.93	14	83	0.95
	CARS-BPNN		21	0.99		47	0.96		50	0.93
	PCA-BPNN		0	0.99		16	0.99		24	0.99
58HC	BPNN	47	18	0.98	10	91	0.74	10	74	0.85
	CARS-BPNN		68	0.80		48	0.90		64	0.81
	PCA-BPNN		0	0.99		18	0.99		18	0.99
22HCC	BPNN	48	57	0.90	10	72	0.90	10	87	0.72
	CARS-BPNN		56	0.84		69	0.81		109	0.85
	PCA-BPNN		1	0.99		18	0.99		18	0.99
Unclassified mixed samples	BPNN	160	53	0.96	35	121	0.81	35	99	0.88
	CARS-BPNN		74	0.93		82	0.88		81	0.92
	PCA-BPNN		75	0.93		98	0.90		82	0.88

(embedded in the PLS model) are two positive procedures to eliminate redundancies and noises, and thereby to improve the accuracy. We employ the characteristic wavelengths extracted by CARS in section 3 as the input of the BPNN model to verify the universality for calibration model improvement. Note that the node number of the input layer is determined by the selected wavelengths of individual paper type, therefore adjusting the weights of nodes for varying paper samples is required. The CARS-BPNN model of different types of paper samples is established and the prediction performance is shown in table 2. We conduct the sample selection by a similar method described in section 3.2, but here the ratio of the training set, validation set and testing set is 70:15:15. It shows that the prediction results of the CARS-BPNN model are still not as good as that of CARS-PLS or that of PLS, especially for 22HCC that RMSE = 109 and $r = 0.85$, although CARS-BPNN has slightly improved the prediction performance for BZZ-75 samples.

The differences in core working mechanisms between PLS and BPNN are assumed responsible for the varying prediction performance of two models. PLS is a linear modeling technique that generalizes and combines ideas from principal components analysis (PCA) and traditional least squares regression methods [27]. Instead, BPNN builds a nonlinear mapping between the spectra and DP by a given activation function. However, the unsatisfactory prediction of BPNN and CARS-BPNN may be caused by the fact that the absorbance of selected wavelengths is directly imported to BPNN without principal components selection. Therefore, we introduce the PCA method into the BPNN algorithm to strengthen spectral characteristics.

The PCs are actually the linear combinations of original spectra. Hence the dimension of the input is dramatically reduced while the PCs still maintain most information of the original spectra. Figure 5 illustrates a typical example of PCA for BZZ-75 paper samples. Note that the extracted PCs are achieved by PCA instead of PLS method. It can be seen that selecting merely four PCs can extract around 95% spectral information (quantified as the cumulative contribution rate).

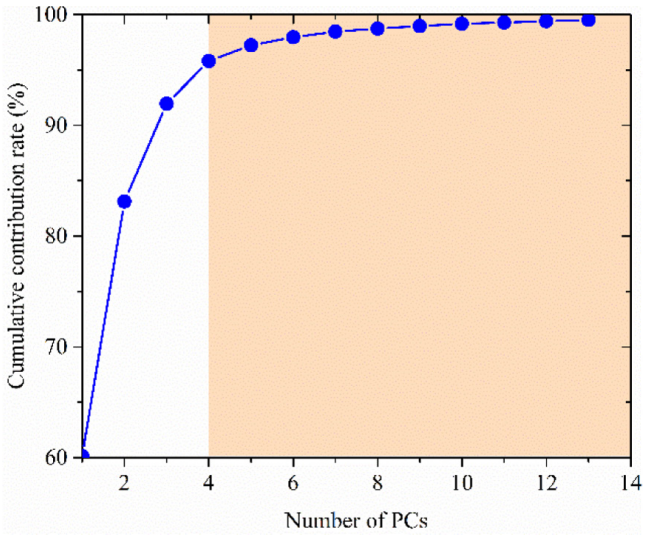

Figure 5. Illustration of PCA by cumulative contribution rate. Note that the results are derived from BZZ-75 paper samples.

Although 94 spectral PCs of BZZ-75 samples are extracted by PCA, the 13 PCs sorted by the variance contribution rate already contain 99.52% of the information of all spectral data. Therefore, we choose these 13 PCs as the input and thereby establish the PCA-BPNN model.

It is clearly seen in table 2 that the prediction performance of the PCA-BPNN model has been remarkably improved both in RMSE and r. The maximum prediction error of three types of paper is less than 24 with a high correlation coefficient ($r = 0.99$). In other words, the established PCA-BPNN model owns much greater accuracy and reliability than the above-mentioned models for all three types of paper sample. The overwhelming prediction performance of the PCA-BPNN model may be attributed to the combination of characteristic information selection by PCA and accurate nonlinear mapping between NIR spectra and DP by BPNN.

However, it is worth noting that none of the models, even PCA-BPNN, achieves a satisfactory prediction performance for the mixed samples. In fact, accurate DP prediction of

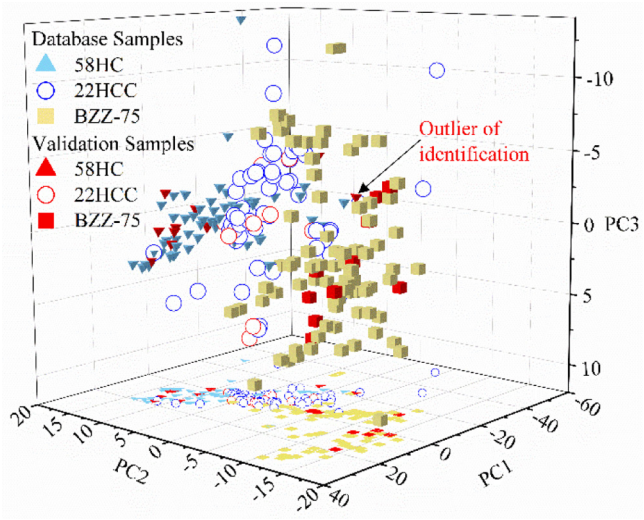

Figure 6. Paper type identification by the KNN method. The first three PCs of 230 paper samples are used for visualization, and 10 paper samples for each type paper are used to verify the identification accuracy of KNN. It is found that 29 out of 30 validation samples are correctly classified and the outlier is marked out.

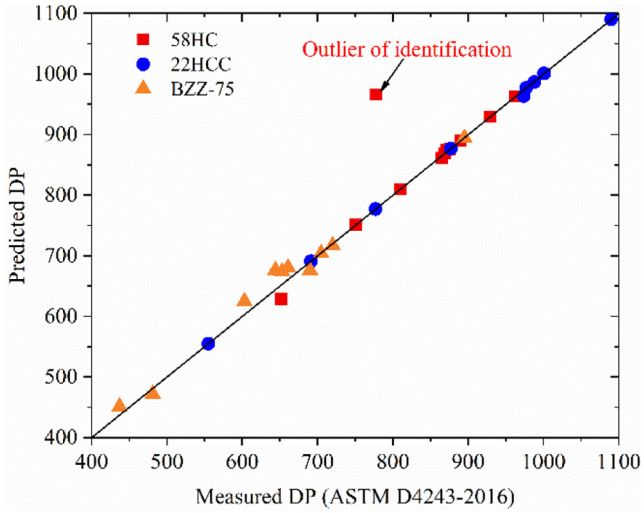

Figure 7. Prediction by the combined calibration model KNN-PCA-BPNN. The prediction of validation samples achieves great performance with only one outlier of the unmatched paper type.

unknown paper samples has greater practical significance since the insulating paper types of the power transformer in the field tests are often difficult to identify.

4.3. KNN-PCA-BPNN calibration model for mixed samples

Though we have learned that the PCA-BPNN model has great prediction performance for each single type of paper, none of the models is able to assess the DP of mixed samples precisely. It is therefore reasonable to identify the paper type before inputting the spectral data into those calibration models to further improve the prediction performance of mixed samples.

We employ the *k*-nearest neighbors (KNN) method to discriminate the type of paper samples. The essence of KNN is to calculate the Euclidean distance between the validation samples and the database samples in a specifically multidimensional space, and accordingly to determine the type of testing paper by the majority of the samples among the *k* nearest samples [28].

Among the 230 mixed paper samples, 200 samples are used to establish the principal components score database while the other 30 samples are for validation. Note that here the selected PCs are the same as the PCA-BPNN presented in section 4.2. The 13 PCs are used to calculate the Euclidean distance. The results of paper type identification by KNN are shown in figure 6, note that in the figure only the first three of 13 PCs are utilized to reveal the distribution of paper samples in a visualized way. The results show that 29 out of 30 validation samples are correctly classified (accuracy 96.7%).

Therefore, we employ KNN to classify the paper before importing the spectral data into the PCA-BPNN model. A KNN-PCA-BPNN calibration model for DP evaluation of mixed insulting paper is established. We verify the prediction performance of the KNN-PCA-BPNN model as shown in figure 7. It is found that this combined calibration model has solved the problem of low prediction precision due to mixed (unknown) paper types. Quantitatively, RMSE is 36 with *r* as high as 0.98, recall that these result sets are (100, 0.85) for PLS and (99, 0.88) for BPNN.

One out of 30 validation samples has been proven the outlier of identification, and the sample has been mistakenly classified as the 58HC class that is supposed to be a 22HCC paper type. The prediction error for the outlier is 188 DP units with the measured DP showing 778, indicating the importance of accurate paper type identification.

5. Conclusion and remarks

In this paper, we investigate different modeling methods, including linear PLS and nonlinear BPNN, using 230 paper samples of varying aging conditions to bridge the NIR spectra and the measured DP. Comparisons of prediction performance between the calibration models are studied in detail.

The linear models for DP evaluation of three types of insulating paper are established by PLS and CARS-PLS methods. It is found that using full NIR spectra as the input of the PLS model does not ensure a high prediction accuracy. Alternatively, the CARS-PLS model has a pronounced improvement in reducing the prediction by introducing the wavelength selection procedures of CARS.

The nonlinear modeling, BPNN, is then established to overcome the complex ambient influences on spectra. However, ordinary BPNN and optimized CARS-BPNN do not achieve as satisfactory a prediction precision as CARS-PLS. The improved PCA-BPNN model has shown overwhelming prediction performance with high accuracy (RMSR $\leqslant$ 24) and correlation ($r = 0.99$). Further, in order to accurately assess DP of the mixed samples, we introduce the KNN method to identify the paper type before DP prediction and build the KNN-PCA-BPNN model. The results show that the combined calibration model has solved the problem of low prediction precision for mixed (unknown) paper samples.

It is worthwhile to point out that the diversity of insulating paper should be enriched to expand the evaluation database of NIRS. Moreover, the high prediction accuracy of mixed paper samples is of great significance for on-site tests of NIRS since the paper types of the power transformer in the field is often unknown. It still deserves further investigation at a broader scale of paper in future work to enhance the prediction performance of calibration models.

Acknowledgments

We thank the editor and two anonymous referees for their careful reviews and thoughtful suggestions on our work. This work was supported in part by the National Natural Science Foundation of China (Grant No. 51607139).

ORCID iDs

Yuan Li https://orcid.org/0000-0001-5424-1764

References

[1] Stevens G C and Emsley A M 1994 Review of chemical indicators of degradation of cellulosic electrical paper insulation in oil-filled transformers *IEE Proc. Science, Meas. Technol.* **141** 324–34
[2] Chakravorti S, Dey D and Chatterjee B 2013 *Recent Trends in the Condition Monitoring of Transformers* (London: Springer) (https://doi.org/10.1007/978-1-4471-5550-8)
[3] Saha T K 2003 Review of modern diagnostic techniques for assessing insulation condition in aged transformers *IEEE Trans. Dielectr. Electr. Insul.* **10** 903–17
[4] N'cho J S, Fofana I, Hadjadj Y and Beroual A 2016 Review of physicochemical-based diagnostic techniques for assessing insulation condition in aged transformers *Energies* **9** 1–29
[5] IEEE 2011 Guide for Loading Mineral Oil-Immersed Transformers and Step-Voltage Regulators *IEEE Standard* C57.91 (https://doi.org/10.1109/IEEESTD.2012.6166928)
[6] ASTM 2016 Test Method for Measurement of Average Viscometric Degree of Polymerization of New and Aged Electrical Papers and Boards *ASTM Standard* D4243
[7] Ali M, Eley C, Emsley A M, Heywood R and Xiao X 1996 Measuring and understanding the ageing of kraft insulating paper in power transformers *IEEE Electr. Insul. Mag.* **12** 28–34
[8] Gredilla A, Fdez-Ortiz de Vallejuelo S, Elejoste N, de Diego A and Madariaga J M 2016 Non-destructive Spectroscopy combined with chemometrics as a tool for Green Chemical Analysis of environmental samples: a review *TRAC Trends Anal. Chem.* **76** 30–9
[9] Ali M, Emsley A M, Herman H and Heywood R J 2001 Spectroscopic studies of the ageing of cellulosic paper *Polymer* **42** 2893–900
[10] Dos Santos E O, Silva A M S, Fragoso W D, Pasquini C and Pimentel M F 2010 Determination of degree of polymerization of insulating paper using near infrared spectroscopy and multivariate calibration *Vib. Spectrosc.* **52** 154–7
[11] Baird P J, Herman H, Stevens G C and Jarman P N 2006 Spectroscopic measurement and analysis of water and oil in transformer insulating paper *IEEE Trans. Dielectr. Electr. Insul.* **13** 293–308
[12] Baird P J, Herman H and Stevens G C 2008 On-site analysis of transformer paper insulation using portable spectroscopy for chemometric prediction of aged condition *IEEE Trans. Dielectr. Electr. Insul.* **15** 1089–99
[13] Fu Q, Li S, Zhang L, Wang M and Qian Y 2018 Study on measuring polymerization degree of oil-impregnated paper by spectroscopy technique *Insul. Mater.* **51** 75–80 (in Chinese) (https://doi.org/10.16790/j.cnki.1009-9239.im.2018.02.014)
[14] Osborne B G 2007 Principles and practice of near infra-red (NIR) reflectance analysis *Int. J. Food Sci. & Technol.* **16** 13–9
[15] Shao Y, He Y and Mao J 2007 Quantitative analysis of bayberry juice acidity based on visible and near-infrared spectroscopy *Appl. Opt.* **46** 6391–6
[16] Magwaza L S, Opara U L, Nieuwoudt H, Cronje P J R, Saeys W and Nicolaï B 2012 NIR Spectroscopy applications for internal and external quality analysis of citrus fruit—a review *Food Bioprocess Technol.* **5** 425–44
[17] Beaudoin J and Malde J 2016 Insulation in transformers *IEEE-GMS-PES-Presentation* (http://site.ieee.org/gms-pes/files/2016/05/IEEE-GMS-PES-Presentation_Mod.pdf)
[18] Tang F, Zhang Y, Yuan B, Li Y, Zhang W and Zhang G 2019 Ageing condition assessment of oil-paper insulation using near infrared spectroscopy detection and analytical technique *J. Eng.* **2019** 3026–9
[19] Hino T and Suganuma T 1972 Rapid measuremnent of the deterioration of oil-immersed paper *IEEE Trans. Electr. Insul.* **EI-7** 122–6
[20] Prevost T A 2005 Thermally upgraded insulation in transformers *Proc. Electrical Insulation Conference and Electrical Manufacturing Expo* pp 120–5
[21] Wold S, Sjöström M and Eriksson L 2001 PLS-regression: a basic tool of chemometrics *Chemometr. Intell. Lab. Syst.* **58** 109–30
[22] Fan W, Shan Y, Li G, Lv H, Li H and Liang Y 2012 Application of competitive adaptive reweighted sampling method to determine effective wavelengths for prediction of total acid of vinegar *Food Anal. Methods* **5** 585–90
[23] Wu W, Walczak B, Massart D L, Heuerding S, Erni F, Last I R and Prebble K A 1996 Artificial neural networks in classification of NIR spectral data: design of the training set *Chemometr. Intell. Lab. Syst.* **33** 35–46
[24] Rojas R 1996 *Neural Networks* (Berlin: Springer) (https://doi.org/10.1007/978-3-642-61068-4)
[25] Li S, Li L, Milliken R and Song K 2012 Hybridization of partial least squares and neural network models for quantifying lunar surface minerals *Icarus* **221** 208–25
[26] Oğuz K and Pekin M A 2019 Predictability of fog visibility with artificial neural network for Esenboga airport *Eur. J. Sci. Technol.* **15** 542–51
[27] Jolliffe I 2011 Principal component analysis *Inter. Encyclopedia of stat. Science* 1094–6
[28] Hao Z, Berg A C, Maire M and Malik J 2006 SVM-KNN: Discriminative nearest neighbor classification for visual category recognition *2006 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR 2006)* **2** 2126–36