ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

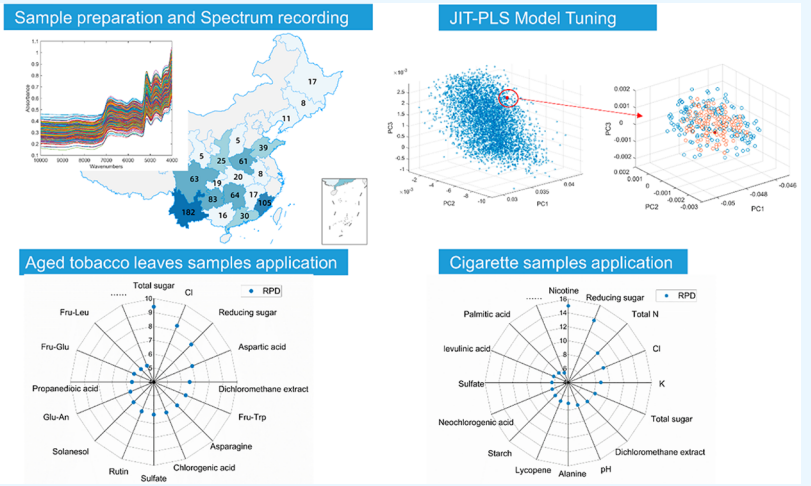
Just-in-Time Learning-Integrated Partial Least-Squares Strategy for Accurately Predicting 71 Chemical Constituents in Chinese Tobacco by Near-Infrared Spectroscopy

Youyan Liang, Le Zhao,* Junwei Guo, Hongbo Wang, Shaofeng Liu, Luoping Wang, Li Chen, Mantang Chen, Nuohan Zhang, Huimin Liu, and Cong Nie

Cite This: *ACS Omega* 2022, 7, 38650−38659

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Near-infrared spectroscopy has been widely used to characterize the chemical composition of tobacco because it is fast, economical, and nondestructive. However, few predictive models perform ideally when applied to large spectral libraries of tobacco and its various chemical indicators. In this study, the just-in-time learning-integrated partial least-squares (JIT-PLS) modeling strategy was applied for the first time to quantitatively analyze 71 chemical components in Chinese tobacco. Approximately 18000 tobacco samples from China were analyzed to find appropriately similar measurements and propose suitable and flexible similar subsets from the calibration for each test sample. In total, 879 representative aged tobacco leaf samples and 816 cigarette samples were used as external instances to evaluate the practical predicting ability of the proposed method. The most suitable similar subsets for each test sample could be selected by limiting the Euclidean distance and number of similar subsets to $0-3.0 \times 10^{-9}$ and 10−300, respectively. The majority of the JIT-PLS models performed significantly better than traditional PLS models. Specifically, using JIT-PLS instead of traditional PLS models increased the R^2 values from 0.347−0.984 to 0.763−0.996, and from 0.179−0.981 to 0.506−0.989 for the prediction of 67 and 71 components in aged tobacco leaf and cigarette samples, respectively. Good prediction ability was demonstrated for routine chemical components, polyphenolic compounds, organic acids, and other compounds, with the mean ratios of prediction to deviation (RPD_{mean}) being 7.74, 4.39, 4.05, and 5.48, respectively). The proposed methodology could simultaneously determine 67 major components in large and complicated tobacco spectral libraries with high precision and accuracy, which will assist tobacco and cigarette quality control in collecting as well as processing stages.

1. INTRODUCTION

Tobacco (*Nicotiana tabacum*) is one of the most extensively cultivated nonfood crops. It is currently cultivated in more than 125 countries[1] and has become a major economic force in several developing countries.[2] Approximately 8400 compounds have been identified in tobacco leaves and cigarette smoke, many of which contribute to the unique flavor, aroma, and physiological effects of tobacco.[3] The chemical composition of tobacco dictates its quality and flavor and is influenced by its

Received: July 1, 2022
Accepted: October 6, 2022
Published: October 20, 2022

https://doi.org/10.1021/acsomega.2c04139
ACS Omega 2022, 7, 38650−38659

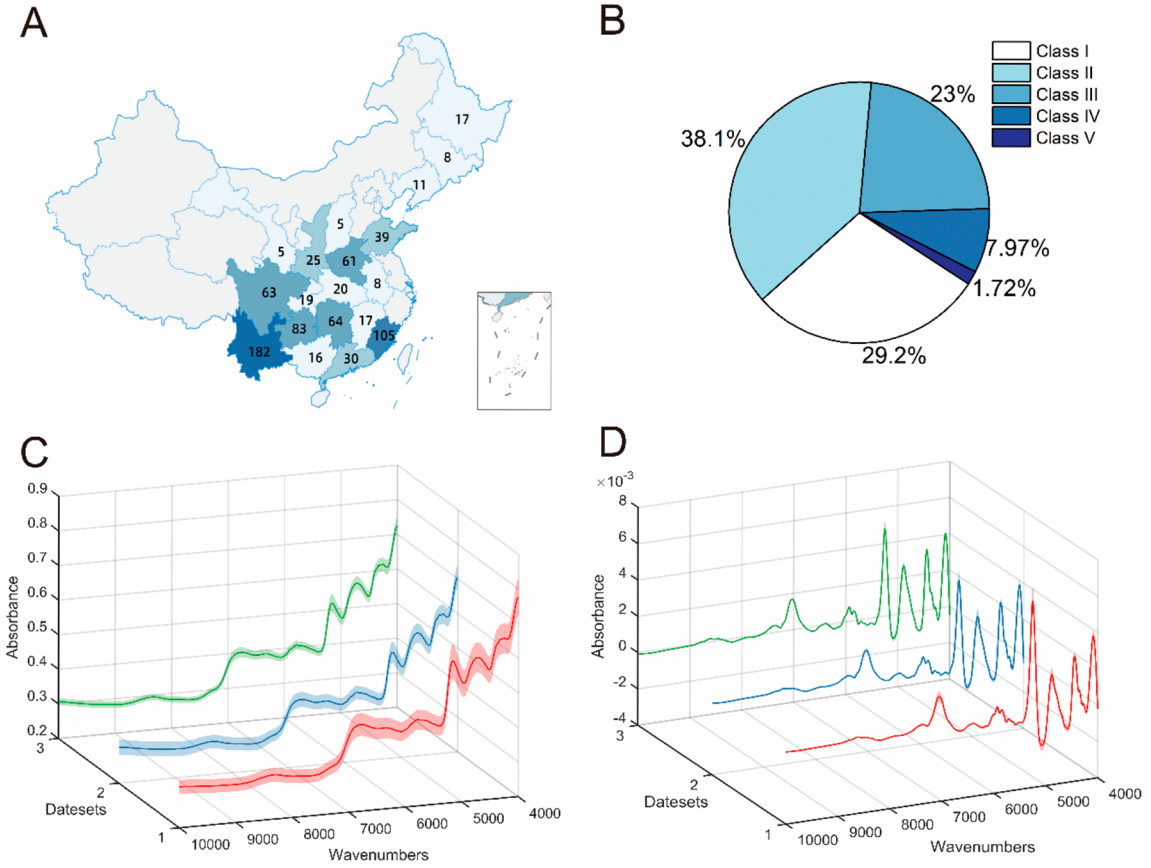


Figure 1. Sampling information: (A) Locations of dataset1 and dataset2; (B) classification of samples in dataste2; (C) average reflectance spectra and their standard deviations of original spectra in all data sets; (D) average reflectance spectra and their standard deviations of derivatived spectra in all data sets.

variety, growing conditions, and processing parameters. Therefore, qualitative and quantitative analyses of the chemical components present in tobacco leaves and cigarettes are critical for quality control.[4,5] However, traditional quantitative analysis methods, such as chromatographic analysis and continuous-flow analysis, are time-consuming and expensive and require complex sample pretreatment procedures.[6,7]

Owing to its rapidity, high efficiency, and nondestructivity, near-infrared (NIR) spectroscopy has been used widely in biological,[8,9] petrochemical,[10] pharmaceutical,[11] agricultural,[12−14] and food-processing[15,16] applications. NIR spectroscopy is also effective in the quantitative analysis of tobacco components.[4,17,18] Chemical models for components having high contents, such as total sugar (TS), reducing sugar (RS), total nitrogen (TN), nicotine (NIC), and chlorine (Cl), are relatively robust.[4,18−20] However, the performance of trace-component models may be subideal.[4,21] Furthermore, models of organic acids, amino acids, and Amadori compounds have rarely been reported, although these compounds critically influence the unique style of tobacco. The prediction of tobacco components requires the creation of a spectral library that relates recorded spectra with reference data. Such a library should be designed to represent variations in the tobacco components of interest. However, the majority of existing spectral libraries are limited to small areas, and few large-scale spectral libraries for tobacco have been developed to date.

Soares et al. quantified 22 components in over 640 samples using NIR hyperspectral imaging, of which 20 were considered satisfactory (R^2_{CV} ranged from 0.67 to 0.86).[17] Zhou et al. used 87 NIR spectra of dark sun-cured tobacco samples from six provinces in China to predict six heavy metals (R^2_{CV} ranged from 0.788 to 0.948).[22] Duan et al. collected 500 samples from Yunnan province to quantitatively analyze 27 chemical components using NIR spectroscopy; results showed remarkable correlation between predicted and measured values of the 15 indices, in which the correlation coefficients of these PLS models were all greater than 0.85.[4] Jiang et al. analyzed the nicotine composition in tobacco leaves collected from Guizhou Province in China using NIR spectroscopy analyzed by cloud computing, in which the correlation coefficients of one-dimensional fully convolutional network model reached to 0.997.[23] Owing to the immense differences between tobacco planting regions and environments, as well as differences between the parts of tobacco plants, large variances exist in the chemical composition of tobacco. Models established using tobacco samples from one particular region may not be suitable for analyzing samples from other regions.

While creating large and complex data sets, tobacco samples may be collected from throughout the country, introducing large heterogeneities in the spectral library. The relationships derived from such a library between the properties and spectra of tobacco may be complex and nonlinear. Therefore, the prediction accuracy achieved by a model calibrated from a large spectral library may deteriorate because the underlying assumption of the model—e.g., linearity for partial least-squares regression (PLSR)—may be invalid.[24] Traditional

https://doi.org/10.1021/acsomega.2c04139
ACS Omega 2022, 7, 38650−38659

global modeling methods generally require the creation of highly complex models when strong nonlinearities are present in the process and its local characteristics are difficult to capture. Furthermore, as model maintenance continues and the modeling sample size increases, a global model may not be available because it incurs considerable computational costs and time.[11] Therefore, building local models may be a reasonable solution for modeling large spectral libraries of complex samples.

Just-in-time (JIT) learning facilitates local modeling to prevent the degradation of prediction accuracy, and the output of each test sample is predicted by the similarities between a test sample and calibration samples.[25] Un-informative or unrelated calibration samples can be effectively removed, and the advantages of having a spectral library covering a large domain can be combined with the accuracy obtained by local calibration models.[26] PLSR integrated with the JIT modeling (JIT-PLS) has been successfully applied to various industrial processes. For the pharmaceutical industry, JIT-PLS has been employed to estimate the content of active pharmaceutical ingredients[27] and rapidly measure residual drug substances (specifically, ibuprofen and magnesium) without sampling. The prediction error in the real-time monitoring of active pharmaceutical ingredient concentration during blending using NIR has also been improved by implementing JIT-PLS.[28] In addition, JIT-PLS has been successfully applied to determine four clinical parameters in human serum samples (total protein, triglyceride, glucose, and urea) by Fourier transform infrared spectroscopy,[29] and local regression approaches exhibited superior performance compared to those of global approaches. However, few studies have employed local modeling methods for complex systems with large spectral libraries. No JIT-PLS model has been reported for the quantitative analysis of chemical constituents in tobacco leaves.

This study attempts to establish JIT-PLS models for multiple components in tobacco samples acquired from different locations in China. The accuracy of prediction depends critically on the selection of the most similar subsets from calibration. The similarity measurements, size of similar subsets, and distance between each test sample and similar subsets were optimized using approximately 18000 NIR spectra of Chinese tobacco leaves and their total nitrogen (TN), nicotine (NIC), total sugar (TS), and reducing sugar (RS) contents. The proposed method was applied to predict the contents of 71 chemical components in sample sets of 879 representative aged tobacco leaf samples and 816 cigarette samples.

2. MATERIALS AND METHODS

2.1. Sample Preparation. Tobacco samples were acquired from various locations having different soil characteristics and climates. They were randomly selected from 14 main planting provinces in China (Figure 1A). Dataset1 included flue-cured tobacco leaf samples and aged tobacco leaf samples purchased by a tobacco company in China from 2010 to 2014. Aged tobacco leaf samples (denominated dateset2) were collected from aged leaf warehouses of 14 cigarette companies from 2015 to 2020 and included different varieties and grades. Cigarette samples (denominated dateset3) covered all valence classes (Classes I, II, III, IV, and V) that were produced in November 2018, March 2020, and March 2021 by all cigarette companies in China (Figure 1B).

All tobacco samples were dried in a drying room at 40 °C for 1−3 days, ground to a certain granularity using a whirlwind grinding mill, and sieved through a 60-mesh sieve. The moisture content of the samples ranged between 6 and 8% and was analyzed by the oven-drying method. The contents of routine chemicals, polyphenolic compounds, organic acids, amino acids, Amadori compounds, and other constituents were tested using different analytical methods, as listed in Table S1 (Supporting Information).

2.2. Spectrum Recording and Pretreatment. NIR spectra were recorded for all tobacco samples using an Antaris II NIR spectrophotometer (Thermo Electron Co., USA). Measurements were performed in triplicate, and each measurement comprised 64 co-added scans recorded at a resolution of 8 cm^{-1} in the wavenumber range of 4000−10000 cm^{-1}.

Multiplicative scatter correction (MSC) was performed prior to modeling to eliminate the uneven distribution of sample particles and reduce the effect of particle size on the spectra. The constant difference in the spectra was eliminated by taking the first derivative—because the calculation of the derivative tended to increase the noise—and performing Savitzky−Golay convolution smoothing prior to derivative preprocessing.

2.3. Just-in-Time Learning-Integrated Partial Least-Squares Regression. A JIT-PLS model was implemented to predict the chemical contents of tobacco leaves (Y) using the spectral matrix ($\mathbf{X}$). Given a set of n reference samples (X_{cal}, Y_{cal}):

$$(X_{cal}, Y_{cal}) = \{x_{cal_j}, y_{cal_j}\}_{j=1}^{M} \quad (1)$$

(i.e., a spectral library) and a set of m samples to predict (X_{val}, Y_{val}),

$$(X_{val}, Y_{val}) = \{x_{val_i}, y_{val_i}\}_{i=1}^{N} \quad (2)$$

where X_{val} is measured and Y_{val} has to be estimated, a basic JIT-PLS algorithm can be described by the following pseudocode:

1 For each sample to predict val_i (i = 1, 2, ..., N), do
2 Compute d_i, the distance vector between x_{val_i} and X_{cal}
3 Find the most similar samples in X_{cal} as a subset of the calibration for each test sample
4 Fit a multivariate model with the subsets of calibration
5 Select the optimal model parameters for the prediction of val_i, e.g., the appropriate number of latent variables (LVs) for a PLS model or the size of subsets
6 Predict samples val_i and compute the squared error
7 Compute the model performance using the coefficient of determination (R^2) between the measured and predicted values, the root mean squared error (RMSE) of the test set, and the ratio of prediction to deviation (RPD).

2.3.1. Spectral Similarity Measurement. Spectral similarity measurements are typically used to measure the similarities between test samples and calibration samples. Considering the efficiency of calculation, Euclidean distance measurement[30] (EDM), locally weighted Euclidean distance measurement[31] (LW-EDM), spectral information divergence[32] (SID), and spectral correlation measurement[33] (SCM) were selected to compute d_i from different perspectives.

2.3.2. Local Multivariate Model. For step 4 of the pseudocode, PLSR was used for regression analysis, wherein the latent variables (LVs) were selected as new predictor variables of the response Y. PLS models were established by 10-fold cross-validation with the maximum number of latent

https://doi.org/10.1021/acsomega.2c04139
ACS Omega 2022, 7, 38650−38659

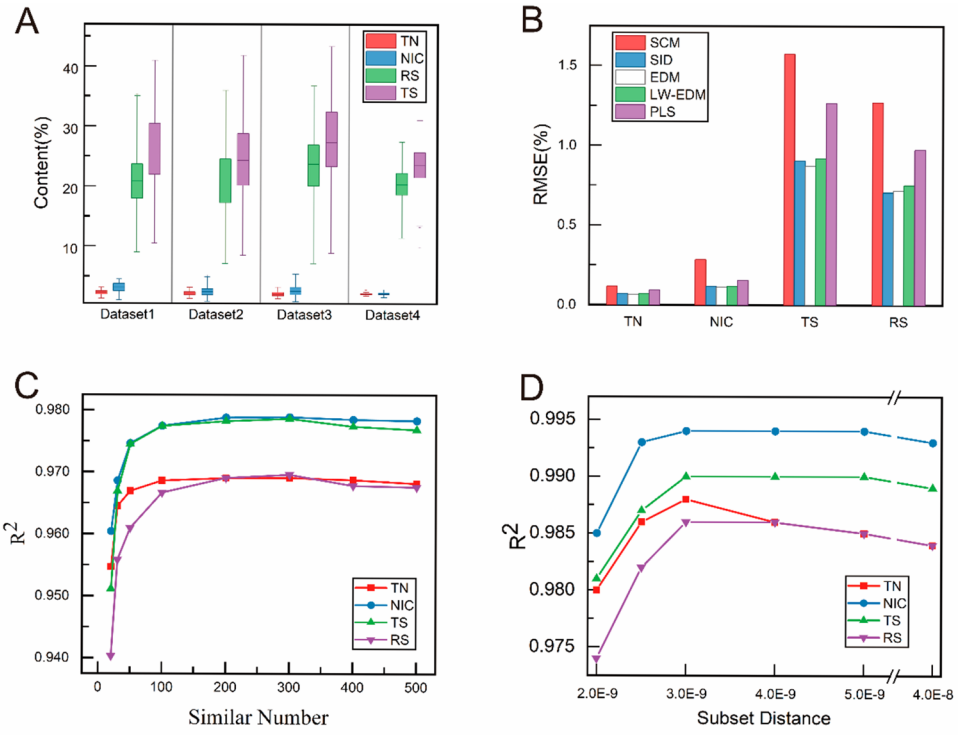


Figure 2. (A) Chemical values of TN, NIC, RS, and TS; (B) results of models of five distance measures in local modeling; (C) R^2 of all models as a function of the number of nearest subset for TN, NIC, RS, and TS; (D) R^2 of all models as a function of distance range between each test sample and nearest subsets in calibration for TN, NIC, RS, and TS.

Table 1. Compositional Characteristics of Dateset1 Used for the Calibration and Validation of the Spectral Models

	calibration set				validation set			
component	*n*	range (%)	mean (%)	SD	*n*	range (%)	mean (%)	SD
TN	17801	1.3−3.1	2.26	0.365	1000	1.32−3.12	2.26	0.373
NIC	17801	1.01−4.5	3.14	0.762	1000	1.11−4.52	3.18	0.767
TS	17801	10.4−40.9	26.2	5.94	1000	10.9−40.7	25.80	5.96
RS	17801	9.02−35.6	20.8	4.15	1000	9.11−33.4	20.55	4.103

variables limited to 10, as the ones producing a model having the smallest value within one standard error of the minimal RMSE of cross-validation, to avoid overfitting.

2.4. Model Evaluation. The following evaluation indicators were used: the coefficient of determination (R^2) between d and the predicted values, the RMSE of the test set, and the ratio of prediction to deviation (RPD). The smaller the RMSE and bias or the larger the RPD and R^2 (<1), the better the performance of the model.

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{N} (yp_i - \hat{y}p_i)^2 / N} \quad (3)$$

where yp_i is the observed value of sample i and $\tilde{y}p_i$ is the predicted value of sample i.

$$\mathrm{RPD} = \mathrm{SD}_y / \mathrm{RMSE} \quad (4)$$

where SD_y is the standard deviation of observed values.

2.5. Software. All computations were performed using MATLAB R2013b (Mathworks, USA). The programs were written in-house.

3. RESULTS AND DISCUSSION

3.1. Model Tuning. Dataset1 (included 1880 samples) was used for method optimization, and these samples were split by the Kennard−Stone method[34] (17801 calibration sets and 1000 validation sets). The averaged original and preprocessed spectra of each data set and their corresponding standard deviations are shown in Figure 1C,D. The peaks and valleys of the original and preprocessed spectra for all samples appeared in identical positions. The spectra of tobacco leaves are complex, as can be observed, showing many bands which reflect the complex composition of the tobacco leaves. The chemical compounds commonly present in tobacco leaves such as carbohydrates (e.g., reducing sugar and starch), nicotine, polyphenolic compounds, organic acids, amino acids, Amadori compounds, etc., have already specific bands assignment in the NIR spectra.[35] However, obtaining chemical information from first-derivative spectra remained challenging.

The contents of TN, NIC, RS, and TS from all data sets are summarized in Figure 2A. The contents of TN, NIC, RS, and TS in dataset1 were varied from 1.3 to 3.1% (w/w) with a standard deviation (SD) of 0.36%, from 1.01 to 4.5% (w/w) with an SD of 0.76%, from 9.02 to 35.62% (w/w) with an SD of 4.16%, and from 10.45 to 40.98% (w/w) with an SD of 5.94%, respectively. The contents of TS and RS varied

https://doi.org/10.1021/acsomega.2c04139
ACS Omega 2022, 7, 38650−38659

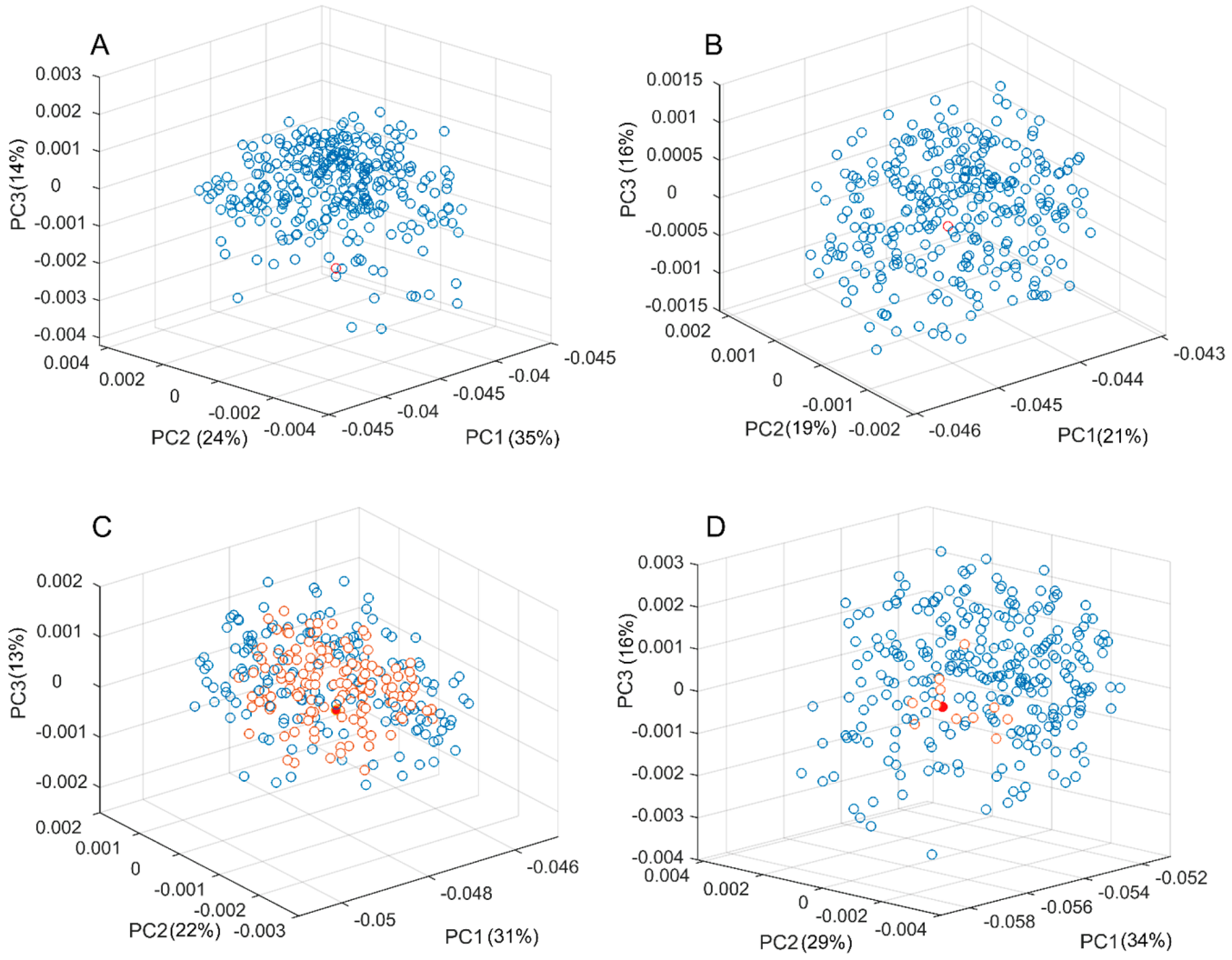


Figure 3. Relationship between the test sample and the most similar samples in calibration. Different distributional forms of the similar samples: (A) Uneven distribution of samples; (B) uniform distribution of samples; (C) similar model performance with (137 orange points) and without (300 blue points) the limit of the Euclidean distance range for the same test sample; (D) similar model performance with (12 orange points) and without (300 blue points) the limit of the Euclidean distance range for the same test sample.

significantly, whereas those of the other constituents remained stable, as presented in Table 1.

There are no prior guidelines for defining the parameters of the JIT-PLS algorithm, such as the appropriate similarity measurement, number of nearest neighbors, or distance metrics. Therefore, these parameters were optimized via a grid search approach using dataset1. The JIT-PLS algorithm was run successively over the following model parameters: similarity measurements, number of nearest subsets, and distance range of the nearest subset.

3.1.1. Optimization of Similarity Measurement in Local Modeling. The prediction performance of different similarity measures in just-in-time learning on a large tobacco NIR spectral library (specifically, SCM, SID, EDM, and LW-EDM) was evaluated. An identical number of nearest neighbors (e.g., 300) was selected from the calibration set for each validation sample to build the JIT-PLS model. Figure 2B shows the RMSE of the TN, NIC, TS, and RS models based on different similarity measurements. All JIT-PLS models except SCM performed better than global PLSR, which indicates that it is inappropriate to evaluate the similarity of the two spectra solely on the basis of the correlation coefficient. EDM exhibited the highest accuracy of measurement for TN, NIC, and TS, while SID was the most accurate method for RS. The model performance of SID and EDM for RS was similar (the difference in R^2 was lower than 0.0009). Therefore, EDM was selected for the follow-up studies.

3.1.2. Optimization of the Number of Nearest Subsets. The number of nearest subsets strongly influenced the prediction accuracy of TN, NIC, TS, and RS. The following numbers of nearest neighbors were tested: 20, 30, 50, 100, 200, 300, 400, and 500. For each indicator, as the number of nearest subsets increased, the R^2 (Figure 2C) of the model gradually increased and stabilized after the number of nearest subsets reached 100. Models having 300 predictors yielded the best results for all chemical indicators. Therefore, the number of nearest subsets was set at 300 for further studies.

3.1.3. Optimization of the Distance Range of Nearest Subset. After sorting the calibration samples in ascending order of similarity, some of the most relevant samples were

https://doi.org/10.1021/acsomega.2c04139
ACS Omega 2022, 7, 38650−38659

Table 2. Results of Models of TN, NIC, TS, and RS used PLS in Dataset1

				validation			calibration		
component	wave band (cm^{-1})	LVs	no. of outliers	no.	R^2	RMSEP (%)	no.	R^2	RMSECV (%)
TN	9005.9−8153.6 7011.9−6525.9 6298.4−4096.1	16	116	1000	0.9842	0.06	17685	0.9822	0.058
NIC	9005.9−8153.6 7011.9−6525.9 6298.4−4096.1	11	98	1000	0.9908	0.112	17703	0.9878	0.093
TS	9005.9−7791.0 7011.9−6525.9 6298.4−4096.1	13	109	1000	0.9947	0.463	17692	0.9954	0.471
RS	9005.9−7791.0 7011.9−6525.9 6298.4−4096.1	13	113	1000	0.9961	0.375	17688	0.9941	0.351

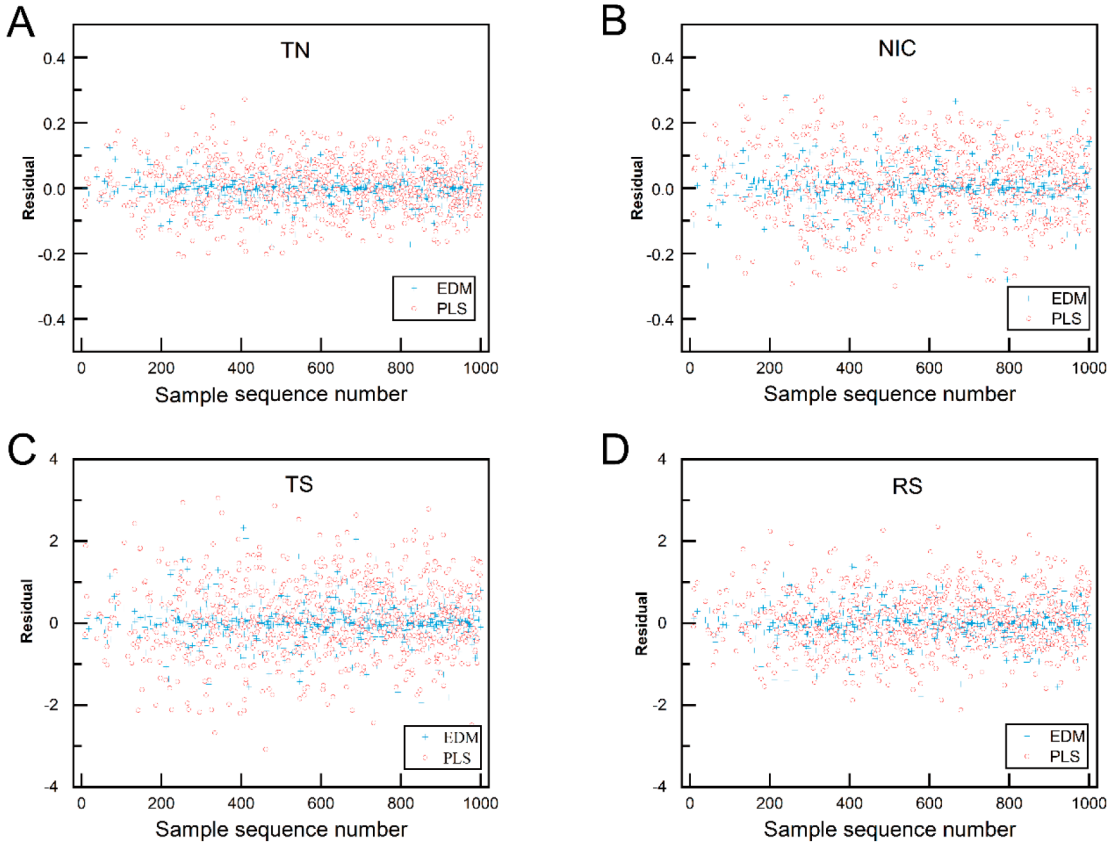


Figure 4. Residuals of the optimized EDM-PLS and traditional PLS models for TN(A), NIC(B), RS(C), and TS(D).

selected as the training samples of the JIT-PLS model. The Euclidean distance distribution of the nearest 300 similar subsets corresponding to the 1000 query sample has been counted using the histogram in Figure S1A. Because the distribution of the calibration samples was irregular, there were two types of situations in which a query sample was related to relevant samples:

(1) The query sample is closely surrounded by its relevant samples, indicating that the similarity between the query sample and the group of relevant samples is relatively high. As shown in Figure 3B, the red point is closely surrounded by the group of blue points, indicating that the local model does not need to generalize a query sample that is significantly beyond the scope of the training samples. Less information is required to construct a local model to predict the output of this query sample.

2 The query sample is distant from the group of relevant samples. The red point in Figure 3A is distant from the group of blue points, which implies that the similarity between the query sample and the relevant samples was lower than that of Figure 3B.

The Euclidean distance distribution of the four query samples and their similar subsets also show that the range of Euclidean distances varies widely for different query samples (Figure S1B), so limiting the number of similar subsets may not be appropriate in some cases. The loadings spectra of the above four query samples are placed in Figure S2; they are very

https://doi.org/10.1021/acsomega.2c04139
ACS Omega 2022, 7, 38650−38659

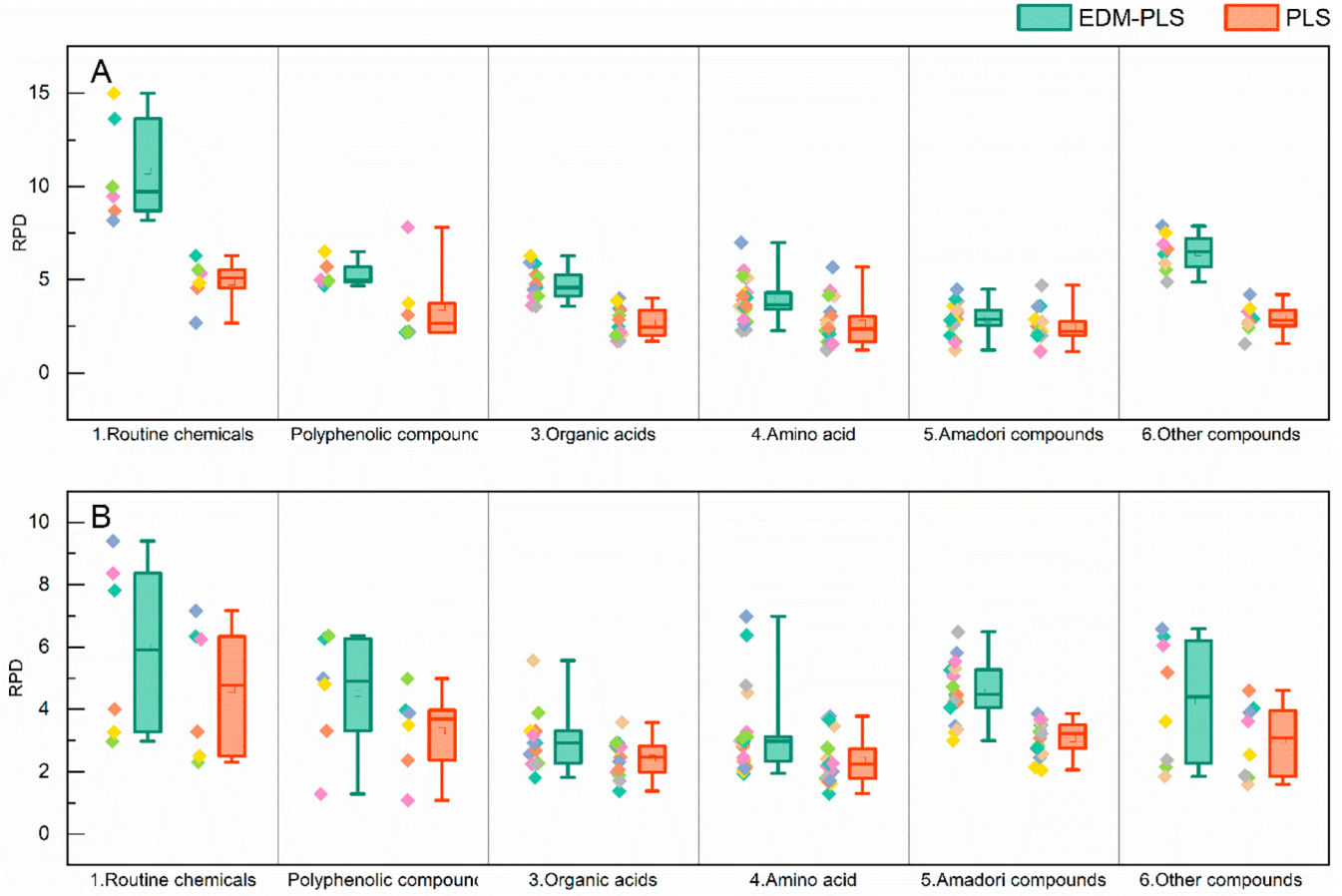


Figure 5. Box plot of the RPD of 71 component models in data set2 (A) and data set3 (B). Each chemical component is represented by a different colored data point.

similar. In addition, to having access to larger numbers of training samples, sufficient valid information is required from the most relevant samples to improve the prediction accuracy. The Euclidean distance between similar subsets and each validation sample was restricted to $(0-2.0) \times 10^{-9}$, $(0-2.5) \times 10^{-9}$, $(0-3.0) \times 10^{-9}$, $(0-4.0) \times 10^{-9}$, $(0-5.0) \times 10^{-9}$, and $(0-4.0) \times 10^{-8}$. It is inadvisible to select too few or too many similar samples for modeling. Therefore, we considered that a validation sample cannot be predicted if the number of similar subsets is lower than 10. The first 300 calibration samples were selected when the number of subsets satisfying the aforementioned conditions exceeded 300.

Among the 1000 validation samples, 735, 865, 925, 965, 982, and 1000 samples were valid when the Euclidean distance was set in the ranges of $(0-2.0) \times 10^{-9}$, $(0-2.5) \times 10^{-9}$, $(0-3.0) \times 10^{-9}$, $(0-4.0) \times 10^{-9}$, $(0-5.0) \times 10^{-9}$, and $(0-4.0) \times 10^{-8}$, respectively. Some boundary test samples were rejected because insufficient numbers of similar samples were found in the calibration set.

To evaluate the influence of a similar distance range on the same validation set, 735 validation samples satisfying the Euclidean distance range of $(0-2) \times 10^{-9}$ were selected as the validation set for all subsequent ranges. As shown in Figure 2D, modeling was most accurate when the Euclidean distance was maintained in the range of $(0-3.0) \times 10^{-9}$ for all indicators.

The model performance with and without the Euclidean distance range limit was compared for the same calibration sample. Figure 3C shows that the prediction bias decreased from 0.0118 (300 subsets, depicted as blue points) to 0.000303 (137 subsets, depicted as orange points). Figure 3D illustrates that the prediction bias decreased from 0.0315 (300 subsets, blue points) to 0.0000234 (12 subsets, orange points). Therefore, limiting the distance range to $(0-3.0) \times 10^{-9}$ and similar subsets in the range of 10−300 were critical.

The model performances of the optimized EDM-PLS and traditional PLS were compared. The optimization of regression for chemical indicators with traditional PLS is presented in Table 2. Evidently, EDM-PLS performed better than traditional PLS (Figure 4).

3.2. External Instance. Dataset2 and dataset3 were used as external instances to evaluate the practical predictive ability of the final prediction model. The average absorbance spectrum of each data set and its corresponding standard deviation (Figure 1C,D) clearly demonstrated that the spectral drift was improved by pretreatment, and the standard deviations decreased significantly, except for several absorption peaks. Two external models were constructed in this study. For each subset, 70 and 30% of the data points were used as calibration and validation sets, respectively, by the Kennard−Stone method following appropriate spectral preprocessing. Some outliers were eliminated because insufficient numbers of suitable similar samples were selected from the calibration set when the distance range was limited to $(0-3.0) \times 10^{-9}$.

Seventy-one chemical models were built using EDM-PLS and PLS methods. Routine chemical components, including NIC, RS, TS, TN, potassium, and Cl, are significant indices for

https://doi.org/10.1021/acsomega.2c04139
ACS Omega 2022, 7, 38650−38659

evaluating the quality of tobacco. However, the aforementioned indices do not reflect the taste and aroma of tobacco. Polyphenolic compounds, such as neochlorogenic acid, chlorogenic acid, cryptochlorogenic acid, scopoletin, rutin, and neophytadiene, are directly transferred from tobacco to smoke by distillation and directly influence the flavor of the smoke. Thirteen organic acids that add organoleptic characteristics to tobacco smoke, such as smoothness and waxy taste, were included. Twenty-one amino acids and 17 Amadori compounds were tested; although the amounts of the amino acid and Amadori compounds varied widely, which is significant for the Maillard reactions. Dichloromethane extracts, starch, magnesium, calcium, sulfate, phosphate, lycopene, and pH were analyzed because these factors can also influence the quality of tobacco from different perspectives.

3.2.1. Aged Tobacco Leaves Samples. The spectral distribution of the aged tobacco samples was highly similar to that of dataset1 (Figure 1C). The amounts of TN, NIC, RS, and TS in aged tobacco leaves were in the ranges of 1.29−3.14, 0.84−5.08, 7.12−35.99, and 8.49−43.83% with SDs of 0.34, 0.70, 5.05, and 6.14%, respectively. The TN and NIC contents differed slightly between dataset1 and dataset2, while the RS and TS contents increased more significantly in aged tobacco compared with dataset1.

The JIT-PLS and PLS models of 71 indicators were built using the previously optimized ranges of the model parameters: The Euclidean distance and the number of similar subsets were restricted to the ranges of $(0-3.0) \times 10^{-9}$ and 10−300, respectively. Sample sets were split according to the Kennard−Stone (KS) algorithm (659 calibration sets and 155 validation sets), from which 65 samples were removed because fewer than 10 similar calibration samples were found when establishing a predictive model.

Figure 5A demonstrates that the performance of EDM-PLS models exceeded that of PLS for the majority of the indicators, particularly for routine chemicals including NIC, RS, TS, TN, potassium, and Cl, with increased RPDs between 4.1 and 10.2 (Table S2, Supporting Information). Polyphenolic compound models also exhibited good performance, with all RPDs exceeding 4.7. The EDM-PLS models (RPD_{mean} = 4.78) performed better than the PLS models (RPD_{mean} = 2.68) for organic acids. The EDM-PLS model of amino acid and Amadori compounds was credible, except for glycine (R^2 = 0.346), cystine (R^2 = 0.624), Fru-Amb (R^2 = 0.68), and Fru-Phe (R^2 = 0.763).

3.2.2. Cigarette Samples. The spectral distribution of cigarette samples was highly similar to those of aged tobacco leaf samples, with smaller SDs (Figure 1C). The contents and SDs of TN, NIC, RS, and TS were lower than those of the aged tobacco leaf samples. A total of 201 validation sets and 612 calibration sets were created using the Kennard−Stone method after removing 3 outliers. The JIT-PLS and PLS models of 71 indicators were also built using the previously optimized model parameters, including the Euclidean distance and number of similar subsets ranges of $(0-3.0) \times 10^{-9}$ and 10−300, respectively.

All indicators exhibited superior performance when using JIT-PLS than when using PLS in our cigarette sample study (Figure 5B). The average R^2 of routine chemicals, polyphenolic compounds, organic acids, amino acids, Amadori compounds, and other compounds for the EDM-PLS models were 0.949, 0.88, 0.862, 0.867, 0.945, and 0.891, respectively (Table S3, Supporting Information). Significant linear correlations were found between the predicted and measured values, except for neophytadiene (R^2 = 0.506), succinic acid (R^2 = 0.694), and Mg (R^2 = 0.706).

In general, EDM-PLS outperformed traditional PLS in terms of prediction ability, and generated the optimal results, increasing the average RPDs from 2.89 to 4.81 and from 2.88 to 3.95 for aged tobacco leaf samples and cigarette samples, respectively. The results demonstrated accurate prediction ability for six routine chemical components (R^2 = 0.89−0.996), six polyphenolic compounds (R^2 = 0.749−0.989), 13 organic acids (R^2 = 0.773−0.99), 19 amino acids (R^2 = 0.706−0.985), 14 Amadori compounds (R^2 = 0.793−0.986), and eight other compounds (R^2 = 0.873−0.989). Predictions of glycine, cystine, Fru-Amb, and Fru-Phe were less accurate, mainly because of their extremely low levels in the samples.

4. CONCLUSIONS

This study investigated the feasibility of JIT-PLS for predicting 71 analytes of interest that influence tobacco quality, including TN, NIC, TS, RS, potassium, and Cl with high precision in aged tobacco leaf samples and cigarette samples. EDM exhibited superior performance compared to SCM, SID, and LW-EDM, and limiting the Euclidean distance range and number of similar subsets to $(0-3.0) \times 10^{-9}$ and 10−300, respectively, optimized the accuracy of finding the most suitable similar subsets for each test sample by analyzing approximately 18000 NIR spectra of Chinese tobacco samples and their TN, NIC, TS, and RS contents. EDM-PLS outperformed traditional PLS in terms of prediction ability. The experimental results proved that it is feasible to directly establish quantitative models of multiple compounds in large and complicated spectral libraries using JIT-PLS. This method can help analyze the spectra of complex plant samples from national or larger sources, using multiple chemical indicators to assist in the evaluation of plant quality on a large scale, which will improve cash crops quality in the collecting and processing stages.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c04139.

Statistics of 71 contents for dateset2 and dateset3 (Table S1); Euclidean distance statistics for similar subsets corresponding to each test sample in dateset1 (Figure S1); PCA loadings in Figure 3 corresponding spectra (Figure S2); and results of models using EDM-PLS and PLS of external instances (Tables S2 and S3) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Le Zhao − *Zhengzhou Tobacco Research Institute of CNTC, Zhengzhou, Henan 450001, China;* orcid.org/0000-0001-8750-6183; Phone: +860371 67672506; Email: 52881027@qq.com

Authors

Youyan Liang − *Zhengzhou Tobacco Research Institute of CNTC, Zhengzhou, Henan 450001, China*

https://doi.org/10.1021/acsomega.2c04139
ACS Omega 2022, 7, 38650−38659

Junwei Guo − *Zhengzhou Tobacco Research Institute of CNTC, Zhengzhou, Henan 450001, China*
Hongbo Wang − *Zhengzhou Tobacco Research Institute of CNTC, Zhengzhou, Henan 450001, China*
Shaofeng Liu − *Zhengzhou Tobacco Research Institute of CNTC, Zhengzhou, Henan 450001, China*
Luoping Wang − *Technology Center of China Tobacco Yunnan Industrial Co. Ltd., Kunming 650231, China*
Li Chen − *Zhengzhou Tobacco Research Institute of CNTC, Zhengzhou, Henan 450001, China*
Mantang Chen − *Zhengzhou Tobacco Research Institute of CNTC, Zhengzhou, Henan 450001, China*
Nuohan Zhang − *Zhengzhou Tobacco Research Institute of CNTC, Zhengzhou, Henan 450001, China*
Huimin Liu − *Zhengzhou Tobacco Research Institute of CNTC, Zhengzhou, Henan 450001, China*
Cong Nie − *Zhengzhou Tobacco Research Institute of CNTC, Zhengzhou, Henan 450001, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c04139

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank the China National Tobacco Corp. for financial support through the project "Research on construction and application of tobacco near-infrared big data" (Project No.: 110201901023(SJ-02)).

■ REFERENCES

(1) Udagawa, H.; Ichida, H.; Takeuchi, T.; Abe, T.; Takakura, Y. Highly Efficient and Comprehensive Identification of Ethyl Methanesulfonate-Induced Mutations in Nicotiana Tabacum L. by Whole-Genome and Whole-Exome Sequencing. *Front. Plant Sci.* **2021**, *12*, 671598.
(2) Krüsemann, E. J. Z.; Cremers, J. W. J. M.; Visser, W. F.; Punter, P. H.; Talhout, R. The Sensory Difference Threshold of Menthol Odor in Flavored Tobacco Determined by Combining Sensory and Chemical Analysis. *Chem. Senses* **2017**, *42* (3), 233−238.
(3) Rodgman, A.; Perfetti, T. A.*The Chemical Components of Tobacco and Tobacco Smoke*, 1st ed.; CRC Press, 2008. DOI: 10.1201/9781420078848.
(4) Duan, J.; Huang, Y.; Li, Z.; Zheng, B.; Li, Q.; Xiong, Y.; Wu, L.; Min, S. Determination of 27 Chemical Constituents in Chinese Southwest Tobacco by FT-NIR Spectroscopy. *Ind. Crops Prod.* **2012**, *40*, 21−26.
(5) Wei, K.; Bin, J.; Wang, F.; Kang, C. On-Line Monitoring of the Tobacco Leaf Composition during Flue-Curing by Near-Infrared Spectroscopy and Deep Transfer Learning. *Anal. Lett.* **2022**, *55*, 2089−2107.
(6) Bian, X.; Diwu, P.; Liu, Y.; Liu, P.; Li, Q.; Tan, X. Ensemble Calibration for the Spectral Quantitative Analysis of Complex Samples. *J. Chemom.* **2018**, *32* (11), No. e2940.
(7) Zimmer, G. F.; Santos, R. O.; Teixeira, I. D.; Schneider, R. d. C. d. S.; Helfer, G. A.; Costa, A. B. Rapid Quantification of Constituents in Tobacco by NIR Fiber-optic Probe. *J. Chemom.* **2020**, *34*, e3303.
(8) Beć, K. B.; Grabska, J.; Huck, C. W. Near-Infrared Spectroscopy in Bio-Applications. *Molecules* **2020**, *25* (12), 2948.
(9) Guo, T.; Chen, X.; Qu, W.; Yang, B.; Tian, R.; Geng, Z.; Wang, Z. Red and Near-Infrared Fluorescent Probe for Distinguishing Cysteine and Homocysteine through Single-Wavelength Excitation with Distinctly Dual Emissions. *Anal. Chem.* **2022**, *94*, 5006−5013.
(10) Yu, H.; Du, W.; Lang, Z.-Q.; Wang, K.; Long, J. A Novel Integrated Approach to Characterization of Petroleum Naphtha Properties from Near-Infrared Spectroscopy. *IEEE T Instrum. Meas.* **2021**, *70*, 1−13.
(11) Zhong, L.; Gao, L.; Li, L.; Nei, L.; Wei, Y.; Zhang, K.; Zhang, H.; Yin, W.; Xu, D.; Zang, H. Method Development and Validation of a Near-Infrared Spectroscopic Method for in-Line API Quantification during Fluidized Bed Granulation. *Spectrochim. Acta, Part A* **2022**, *274*, 121078.
(12) Jang, D.; Sohng, W.; Cha, K.; Chung, H. A Weighted Twin Support Vector Machine as a Potential Discriminant Analysis Tool and Evaluation of Its Performance for Near-Infrared Spectroscopic Discrimination of the Geographical Origins of Diverse Agricultural Products. *Talanta* **2022**, *237*, 122973.
(13) Álvarez-Mateos, P.; Alés-Álvarez, F.-J.; García-Martín, J. F. Phytoremediation of Highly Contaminated Mining Soils by Jatropha Curcas L. and Production of Catalytic Carbons from the Generated Biomass. *J. Environ. Manage.* **2019**, *231*, 886−895.
(14) García-Martín, J. F.; Badaró, A. T.; Barbin, D. F.; Álvarez-Mateos, P. Identification of Copper in Stems and Roots of Jatropha Curcas L. by Hyperspectral Imaging. *Processes* **2020**, *8* (7), 823.
(15) Wiedemair, V.; Langore, D.; Garsleitner, R.; Dillinger, K.; Huck, C. Investigations into the Performance of a Novel Pocket-Sized Near-Infrared Spectrometer for Cheese Analysis. *Molecules* **2019**, *24* (3), No. 428.
(16) Badaro, A. T.; Garcia-Martin, J. F.; Lopez-Barrera, M. d. C.; Barbin, D. F.; Alvarez-Mateos, P. Determination of Pectin Content in Orange Peels by near Infrared Hyperspectral Imaging. *Food Chem.* **2020**, *323*, 126861.
(17) Soares, F. L. F.; Marcelo, M. C. A.; Porte, L. M. F.; Pontes, O. F. S.; Kaiser, S. Inline Simultaneous Quantitation of Tobacco Chemical Composition by Infrared Hyperspectral Image Associated with Chemometrics. *Microchem. J.* **2019**, *151*, 104225.
(18) Zhang, Y.; Cong, Q.; Xie, Y.; Yang, J.; Zhao, B. Quantitative Analysis of Routine Chemical Constituents in Tobacco by Near-Infrared Spectroscopy and Support Vector Machine. *Spectrochim. Acta, Part A* **2008**, *71* (4), 1408−1413.
(19) Qin, Y.; Gong, H. NIR Models for Predicting Total Sugar in Tobacco for Samples with Different Physical States. *Infrared Phys. Technol.* **2016**, *77*, 239−243.
(20) Wu, L.; Wang, B.; Zhang, L.; Duan, R.; Gao, R.; Yin, Y.; Liu, X.; Bai, X. Determination of Routine Chemicals, Physical Indices and Macromolecular Substances in Reconstituted Tobacco using near Infrared Spectroscopy Combined with Sample Set Partitioning. *J. Near Infrared Spectrosc.* **2020**, *28* (3), 153−162.
(21) Ma, Y.; Bai, R.; Du, G.; Ma, L.; He, A.; Li, N.; Yi, X.; Cai, W.; Zhou, J.; Shao, X. Rapid Determination of Four Tobacco Specific Nitrosamines in Burley Tobacco by Near-Infrared Spectroscopy. *Anal. Methods* **2012**, *4* (5), 1371.
(22) Huang, Y.; Du, G.; Ma, Y.; Zhou, J. Predicting Heavy Metals in Dark Sun-Cured Tobacco by Near-Infrared Spectroscopy Modeling Based on the Optimized Variable Selections. *Ind. Crops Prod.* **2021**, *172*, 114003.
(23) Jiang, D.; Hu, G.; Qi, G.; Mazur, N. A Fully Convolutional Neural Network-Based Regression Approach for Effective Chemical Composition Analysis using Near-Infrared Spectroscopy in Cloud. *J. Artif. Intell. Technol.* **2021**, *1* (1), 74−82.
(24) Gupta, A.; Vasava, H. B.; Das, B. S.; Choubey, A. K. Local Modeling Approaches for Estimating Soil Properties in Selected Indian Soils Using Diffuse Reflectance Data over Visible to Near-Infrared Region. *Geoderma* **2018**, *325*, 59−71.
(25) Zhang, X.; Kano, M.; Song, Z. Optimal Weighting Distance-Based Similarity for Locally Weighted PLS Modeling. *Ind. Eng. Chem. Res.* **2020**, *59*, 11552−11558.
(26) Pérez-Marín, D.; Garrido-Varo, A.; Guerrero, J. E. Non-Linear Regression Methods in NIRS Quantitative Analysis. *Talanta* **2007**, *72* (1), 28−42.
(27) Kim, S.; Kano, M.; Nakagawa, H.; Hasebe, S. Estimation of Active Pharmaceutical Ingredients Content Using Locally Weighted Partial Least Squares and Statistical Wavelength Selection. *Int. J. Pharm. (Amsterdam, Neth.)* **2011**, *421*, 269.

(28) Nakagawa, H.; Kano, M.; Hasebe, S.; Miyano, T.; Watanabe, T.; Wakiyama, N. Verification of Model Development Technique for NIR-Based Real-Time Monitoring of Ingredient Concentration during Blending. *Int. J. Pharm. (Amsterdam, Neth.)* **2014**, *471* (1), 264−275.
(29) Perez-Guaita, D.; Kuligowski, J.; Quintás, G.; Garrigues, S.; Guardia, M. de la. Modified Locally Weighted-Partial Least Squares Regression Improving Clinical Predictions from Infrared Spectra of Human Serum Samples. *Talanta* **2013**, *107*, 368−375.
(30) Leung, H.; Huang, Y.; Cao, C. Locally Weighted Regression for Desulphurisation Intelligent Decision System Modeling. *Simul. Modell. Pract. Theory* **2004**, *12* (6), 413−423.
(31) Nakagawa, H.; Tajima, T.; Kano, M.; Kim, S.; Hasebe, S.; Suzuki, T.; Nakagami, H. Evaluation of Infrared-Reflection Absorption Spectroscopy Measurement and Locally Weighted Partial Least-Squares for Rapid Analysis of Residual Drug Substances in Cleaning Processes. *Anal. Chem.* **2012**, *84* (8), 3820−3826.
(32) Fujiwara, K.; Kano, M.; Hasebe, S.; Takinami, A. Soft-Sensor Development using Correlation-Based Just-in-Time Modeling. *AIChE J.* **2009**, *55* (7), 1754−1765.
(33) Zhang, E.; Zhang, X.; Yang, S.; Wang, S. Improving Hyperspectral Image Classification using Spectral Information Divergence. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11* (1), 249−253.
(34) Kennard, R. W.; Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **1969**, *11* (1), 137−148.
(35) Tammer, M. G. Sokrates: Infrared and Raman characteristic group frequencies: tables and charts. *Colloid Polym. Sci.* **2004**, *283*, 235.

https://doi.org/10.1021/acsomega.2c04139
ACS Omega 2022, 7, 38650−38659