Original Article
Food Science and Technology

ISSN 0101-2061 (Print)
ISSN 1678-457X (Online)

DOI: https://doi.org/10.1590/fst.41922

Geographical origin of green tea identification using LASSO and ANOVA

Tianhong PAN[1], Ru YAN[1], Qi CHEN[2]

Abstract

To standardize the tea export market and guarantee the interest of consumers, the adulteration problem in Taiping Houkui tea should be eliminated. In this study, a screening scheme comprising chemometrics and statistical analysis was proposed to estimate the geographical origin of Taiping Houkui tea. A total of 11 metal ions in Taiping Houkui tea were detected by performing a chemometric experiment. The key variables that can be used to identify the geographical origin of Taiping Houkui tea were screened using the least absolute shrinkage and selection operator method (LASSO). The statistical significance of selected key variables was also tested by analysis of variance (ANOVA), which confirmed the effectiveness of the LASSO. The proposed strategy was verified by the experimental testing and has great potential for determining the geographical origin of green tea.

Keywords: tea quality estimation; feature extraction; least absolute shrinkage and selection operator (LASSO); analysis of variance (ANOVA).

Practical Application: It is difficult to distinguish the geographical origins of Taiping Houkui tea because the differences in shape, color, and internal composition of tea with similar grades are subtle. To develop a new method to identify the geographical origins of Taiping Houkui tea, a model-based scheme comprising chemometrics and the use of the LASSO was established to estimate the geographical origin of the green tea. The effectiveness of the proposed scheme was verified using ANOVA, thereby simplifying the process of metal ions detection.

1 Introduction

Tea is one of the three most widely consumed non-alcoholic beverages worldwide and has numerous economic, health, and cultural values (Meng et al., 2022). Tea consumption has been reported to have numerous health benefits, such as reduction of serum cholesterol (Doğan et al., 2021), prevention of low-density lipoprotein oxidation (Xu et al., 2021), and decreased risk of cardiovascular syndromes (Xia et al., 2020). Tea from specific origins has a strong influence on the purchasing decision of consumers, whose price may be high compared with the green teas from other regions.

As a kind of well-known Chinese green tea, Taiping Houkui tea mainly occupies an important position in the market. Hougang Village, Houcun Village, and Yanjia Village of Huangshan city, Anhui province is the core production area in China, where the quality of Taiping Houkui tea is significantly superior to that of other production areas. The quality and price of green tea are greatly influenced by geographical origin, which further affects the consumer's desire to purchase (Pang et al., 2022). Consequently, adulteration and inferior quality of tea were reported. Some unscrupulous traders fraudulently label their green tea from core production areas to obtain profits. This caused great concern regarding food safety and tea quality, which impacts the market price and consumer satisfaction, thus attracting increasing attention (Song et al., 2021; Zhang et al., 2021). Unfortunately, it is difficult to distinguish the geographical origins of Taiping Houkui tea because the differences in shape, color, and internal composition of Taiping Houkui tea from different origins are subtle (Huang et al., 2020; Li et al., 2019).

Various technologies were developed to identify the geographical origins of Taiping Houkui tea. Traditionally, the assessment is conducted by experienced experts in tea sensory evaluation. However, the accuracy, reproduction, and standard of Taiping Houkui tea quality of each batch cannot be guaranteed (Jin et al., 2021). Some electronic instruments have been developed and applied to handle this problem. An analytical instrument named electronic tongue (E-tongue) is a typical array of chemical sensors coupled with chemometrics for processing and characterizing green tea samples (Bhuyan et al., 2019). Sensors can generate analytically useful electric signals when interacting with green tea samples. E-tongue systems take the advantage of easy operation, low cost, simple setup, easy fabrication, etc. Nevertheless, the major drawbacks of potentiometric sensors are their temperature dependence and the adsorption of solution components that affect the membrane potential (Zhang et al., 2019). Artificial olfaction (known as E-nose currently) has been exploited as another useful tool (Xu et al., 2019). E-nose devices consist of a sensor array that is adequately sensitive to the volatile. It can mimic the mammalian sense of smell by producing a composite response unique to each odorant. The E-nose has been successfully used to identify the geographical origins and geographical origin of green tea because the aroma is an important factor, which depends upon the amount of volatile organic compounds. Compared with the conventional methods,

Received 19 Mar., 2022
Accepted 10 May, 2022
[1] *School of Electrical Engineering and Automation, Anhui University, Hefei, Anhui, China*
[2] *School of Tourism, Huangshan University, Huangshan, Anhui, China*
*Corresponding author: thpan@ahu.edu.cn

E-nose is a reliable, fast, and robust technology. However, the higher sensitivity of the sensors to moisture and water vapor results in essential noise from strong sensor outputs produced due to the presence of water vapor in the headspace of green tea samples (Hidayat et al., 2010). Currently, near-infrared (NIR) spectroscopy combined with a calibration model was applied in the geographical origins analysis of green tea (Cardoso & Poppi, 2021). NIR spectroscopy with the advantage of rapid, non-destructive, and high-efficiency determination, has gained wide acceptance (Chen et al., 2022). However, one shortcoming of NIR spectroscopy is that it requires very sensitive and properly tuned instruments. The organic chemical components can be detected using NIR spectroscopy only for those whose content is more than 0.1%. More important, specific absorption in the NIR region can only express the organic molecules of the material being analyzed. Mineral elements cannot be detected using this technology. However, the mineral elements including trace metals and rare earth elements play a critical role in the geographical origin identification of green tea (Bobková et al., 2021). The amount and ratio of mineral elements in the soil have a direct impact on the quality of Taiping Houkui tea. After being absorbed by Taiping Houkui tea, some mineral elements act as the key composition of enzymes and coenzymes affecting the formation rate and amount of organic chemical components (such as certain kinds of amino acids, caffeine, tea polyphenol, catechin), the rest exist in the form of inorganic salts, accounting for 4-6% of green tea (Wang et al., 2003). There are significant differences in the amount and ratio of mineral elements in green tea from different origins (Ye et al., 2017). Therefore, the geographical origins of green tea can be traced by the difference in mineral elements. Moreover, appropriate chemometrics is also critical in the process. In multi-variant statistical approaches, principal component analysis (PCA) and partial least square discriminant analysis (PLS-DA) were widely used to build good inferential models (Shevchuk et al., 2018). The PCA and PLS-DA can handle data with high dimensionality and collinearities by projecting the original variables onto a space defined by orthogonal components (Shao, et al., 2019). The correlation between the independent and dependent variables was not considered in the PCA, which PLS-DA overcame. Unfortunately, PLS-DA cannot screen out the key factors that affect the dependent variables.

A facile and reliable method to distinguish the geographical origin of Taiping Houkui tea is developed and the feasibility of the proposed scheme is verified in this work. A total of 120 Taiping Houkui tea samples were collected from six villages including core regions and non-core regions. The contents of 11 metal ions in the collected Taiping Houkui tea samples were measured and analyzed. The key metal ions were screened using the least absolute shrinkage and selection operator method (LASSO), and analysis of variance (ANOVA) was used to validate the selected variables. The geographical origin of Taiping Houkui tea could be determined based on linear regression using the selected metal ions.

2 Materials and methods

2.1 Materials

All Taiping Houkui tea samples were collected from Huangshan city in China. Their origins, quantity, and corresponding plucking time are shown in Table 1. Houkeng village, Hougang village, and Yanjia village are the core production areas in which the quality of Taiping Houkui tea is better than in other areas. Sanhe village, Shihekeng village, and Wangling village are the other three non-core production areas, where the Taiping Houkui tea samples are the treatment group. All samples were encoded, stored, and labeled. The concentration of metal ions in samples was attained by the microwave digestion pretreatment method combined with inductively coupled plasma mass spectrometry (ICP-MS) (Patocka et al., 2017). This technology has great sensitivity, precision, and a wide linear measurement range. The name, specification, and source of materials and reagents used in this experiment are shown in Table 2.

2.2 Metal ions detection

The procedure of metal ions detection is as follows:

1) Taiping Houkui tea samples of 0.2500 g ± 0.0010 g were weighed in the microwave digestion tanks. Two sets of parallel samples were set in this experiment;
2) HNO_3 of 5mL was added and kept for 30 min. H_2O_2 of 2 mL was added and kept for 2 min. After that, the digestion tanks were covered and put into the microwave digestion instrument for digestion. Subsequently, the tanks were taken out after cooling, and their covers were opened slowly to exhaust;
3) The inner covers were rinsed with a small amount of water, and the digestion tanks were placed in the temperature control electric heating plate to outgas at 100 °C for 30 min. The process of degassing can be replaced by degassing in an ultrasonic water bath box for 2-5 min;
4) The digestion tanks were volumed to 25 mL using ultrapure water constant and sealed hermetically while doing the blank test. The high content of K, Ca, and Mg in Taiping Houkui tea samples should be diluted in a certain proportion and then determined. The model and origin of the experimental instrument in the experiment are shown in Table 3. In this study, a total of 11 metal ions (Ca, Mg, K, Mn, Fe,

Table 1. Origins, quantity, and plucking time information of collected samples.

Code	Origins	Specifications	Quantity	Plucking Time
#1	Houkeng village	50 g	20	2020-04-18
#2	Hougang village	50 g	20	2020-04-17
#3	Yanjia village	50 g	20	2020-04-17
#4	Sanhe village	50 g	20	2020-04-16
#5	Shihekeng village	50 g	20	2020-04-16
#6	Wangwangling village	50 g	20	2020-04-16

Table 2. Names, specifications, and sources of materials and reagents.

Reagents	Specification or grade	Source or manufacturer
HNO_3	Level-UP (68%)	Suzhou Jingrui Chemical Co., Ltd
H_2O_2	Degree of purity (30%)	Guoyao Group Chemical Reagent Co., Ltd.
Taiping Houkui tea standard materials	GBW10083	China Institute of Metrology
Pb, Cd, Cr, Cu, Fe, Mn, Zn, K, Na, Ga, Mg	1000 mg/L	National Center for Analysis and Testing of Nonferrous Metals and Electronic Materials
Sc, Ge, Im, Bi (Internal standard elements)	1000 mg/L	Ditto

Table 3. Information of experimental instruments in detail.

Apparatus	Model	Country and Company
Inductively coupled plasma mass spectrometry	PE NeXlON 350D	Perkin Elmer Corporation, USA
High-pressure microwave digester	CEM MARS 5	CEM Corporation, USA
Ultra-pure water system	Millipore direct 16	Millipore Corporation, USA
Thermostatic drying chamber	M5	Binder Corporation, Germany
Temperature control electric heating plate	DB-3	Changzhou Guoyu Instrument Manufacturing Corporation, China
Ultrasonic water bath box	KQ-500DE	Kunshan Ultrasonic Instrument Corporation, China
Electronic analytical balance	Mettler-AL204-IC	Mettler Toledo Instruments Corporation, China

Na, Cu, Zn, Cr, Pb, Cd) in Taiping Houkui samples were determined using microwave digestion and ICP-ES.

2.3 Key variables selection using LASSO

The pretreated metal ions of Taiping Houkui tea (i.e., independent variables) are denoted as $X=\left[\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n\right]^T\in\Re^{n\times m}$, where $\mathbf{x}_i=\left[x_{i,1},x_{i,2},\cdots,x_{i,m}\right]^T$, and the corresponding quality labels (i.e., dependent variables) are described as $Y=\left[y_1,y_2,\cdots,y_n\right]^T\in\Re^{n\times}$. The relationship between $x_{i,1},x_{i,2},\cdots,x_{i,m}$ and y_i is calculated by Equation 1:

$$y_i=\beta_1x_{i,1}+\beta_2x_{i,2}+\cdots\beta_mx_{i,m}=\beta^T\mathbf{x}_i \tag{1}$$

where $\beta=\left[\beta_1,\beta_2,\cdots,\beta_m\right]^T$ is the coefficient vector. In ordinary least squares (OLS), the estimation of the coefficient vector $\hat{\beta}$ is obtained by minimizing the residual sum of squares, which is formulated as:

$$\hat{\beta}=\arg\min\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{m}\beta_jx_{i,j}\right)^2 \tag{2}$$

However, Equation 2 is morbid because of the correlation between input variables. In other words, some metal ions have similar effects on the geographical origin of Taiping Houkui tea. Hence, it is necessary to determine the explanatory metal ions of the identification process. The LASSO algorithm is proposed by introducing an extra penalty into Equation 3, which is shown as (Tibshirani, 1996):

$$\hat{\beta}=\arg\min J(\beta)=\arg\min\left(\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{m}\beta_jx_{i,j}\right)^2+\lambda\sum_{j=1}^{m}|\beta_j|\right) \tag{3}$$

where $\lambda\sum_{j=1}^{m}|\beta_j|$ is called the LASSO penalty, and λ is a non-negative tuning parameter.

The algorithm constrains the absolute value of the regression coefficient of the model at a certain threshold and then minimizes the sum of squares of the model residuals. Therefore, the coefficient of variables with a correlation greater than the threshold continues to shrink to zero, thereby realizing variable selection (Huang et al., 2012). The algorithm heavily relies on the parameter λ, which is the amount of shrinkage. The algorithm becomes OLS when λ is equal to zero. Conversely, $\lambda=\infty$ implies no feature is considered. The bias increases with increase in λ, and variance increases with decrease in λ. To select the optimal λ, the grid search method combined with ten-fold cross-validation (CV) was implemented in the process. The appropriate range of λ was selected with a small positive number (tends to zero) to a large number. In ten-fold CV, the entire dataset was divided into ten groups. A subset was used as the prediction set, and the remaining nine sub-datasets were used to construct the model using LASSO. The procedure was repeated ten times to eliminate the occasionality.

A coordinate descent (CD) algorithm was developed to solve LASSO by successively performing approximate minimization along with coordinate directions or coordinate hyperplanes (Friedman et al., 2010; Wright, 2015). The pseudocode of the LASSO is as follows (Qu & Richtarik, 2016): The CD method minimizes the objective function in one coordinate at a time and cycles through all coordinates until convergence. Set $\lambda_{\max}$ is a sufficiently large value that ensures that $\hat{\beta}$ is a null vector. The CD algorithm produces a solution path $\hat{\beta}$ over a grid of points λ_d, $d=1,2,\cdots D$where $\lambda_1=\lambda_{\max}$ and $\lambda_D=0$. The pseudocode of the LASSO method is presented in Algorithm 1. Finally, the corresponding independent variables with $\beta_j\neq0, j=1,2,\cdots m$ are selected, which directly affect the estimation results (Zuo et al., 2021).

Algorithm 1. The flowchart of LASSO.

input: collect X and Y.

initialization: Let $\hat{\beta}=[\beta_1,\beta_2,\cdots\beta_m]\in\mathbb{R}^{m\times1}$ be an initial vector generated randomly, l is the iterations.

1: **for** $d=1,2,\cdots D$

2: $l=0$;

3: **do** {
$l=l+1$;

4: Cyclically For $j=1,2,\cdots m$

4: Update the j_{th} component $\hat{\beta}_j$ of $\hat{\beta}$ by $\hat{\beta}_j=\begin{cases}\dfrac{(\rho_j+\frac{\lambda_d}{2})}{z_j},\rho_j<-\dfrac{\lambda_d}{2}\\ 0,-\dfrac{\lambda_d}{2}\le\rho_j\le\dfrac{\lambda_d}{2}\\ \dfrac{(\rho_j-\frac{\lambda_d}{2})}{z_j},\rho_j>-\dfrac{\lambda_d}{2}\end{cases}$, where

5: $\rho_j=\sum_{i=1}^{n}(y_i-\sum_{c\neq j}^{m}\hat{\beta}_c x_{i,c})x_{i,j}$, $z_j=\sum_{i=1}^{n}x_{i,j}^2$;

6: } **while** ($\left\|\beta^{l-1}-\beta^{l}\right\|<0.001$ || $l<500$)

7: **end for**

8: Find the optimal λ with the minimum $J(\hat{\beta})$.

Output: The optimal coefficient vector $\hat{\beta}=\{\beta_1,\beta_2,\cdots\beta_m\}$ corresponding to the optimal λ.

2.4 Statistical analysis

Analysis of variance (ANOVA) is an important method in exploratory and confirmatory data analysis (Nourbakhsh et al., 2013). The method uses a variance ratio to estimate the importance of the selected metal ions. Assuming that the number of groups is denoted by q and the number of samples in each group is w, and the mean of each group is denoted as $\mu_1,\mu_2,\cdots\mu_q$. In general, the null hypothesis is tested as (Equation 4).

$$\begin{cases}H_0:\mu_1=\mu_2=\cdots=\mu_q\\ H_1:\mu_1\neq\mu_2\neq\cdots\neq\mu_q\end{cases}\tag{4}$$

The ANOVA divides the variance of all observations (SST) into within-group variance (SSW) and between-groups (SSB) variance, that is (Equation 5),

$$SST=SSB+SSW\tag{5}$$

The SSB is calculated using Equation 6,

$$SSB=\frac{1}{q-1}\sum_{h=1}^{q}c_h\left(\overline{x}_h-\overline{x}\right)^2\tag{6}$$

where c_h is the number of samples in the h_{th} group, $\overline{x}_h$ is the mean of the h_{th} group, and $\overline{x}$ is the mean of all samples.

The SSW is calculated by Equation 7,

$$SSW=\frac{1}{(q-1)w}\sum_{h=1}^{q}(\sum_{g=1}^{w}(x_{h,g}-\overline{x}_h)^2)\tag{7}$$

where $x_{h,g}$ represents the g_{th} sample of the h_{th} group. Then, the F-statistic is constructed to test the hypothesis, which is (Equation 8):

$$F=\frac{SSB}{SSW}\tag{8}$$

ANOVA allows α(the type I error rate) to be held at a predetermined level. If $F>F_t$, the null hypothesis can be rejected, where F_t is consulted from a table of critical Fvalues. Particularly, the detected metal ions are significant to the determination of the geographical origin of Taiping Houkui tea. The algorithm was applied in this study to verify the significance of the metal ions selected by LASSO.

3 Results and discussion

3.1 Experiment scheme

As demonstrated in Figure 1, an experiment on the geographical origin identification of Taiping Houkui tea was conducted using the proposed method. In this work, Monte Carlo simulation was implemented to eliminate occasionally. Compared with the K-fold cross-check (Wong, 2015) and leave-one-out cross-check, the model index obtained using Monte Carlo cross-validation (MCCV) is close to the actual prediction ability (Du et al., 2006). In this study, proportional stratified sampling randomly was used to construct the calibration and prediction sets at a ratio of 7:3 for statistical analysis. The MCCV was iteratively performed 100 times to avoid errors caused by the unreasonable division

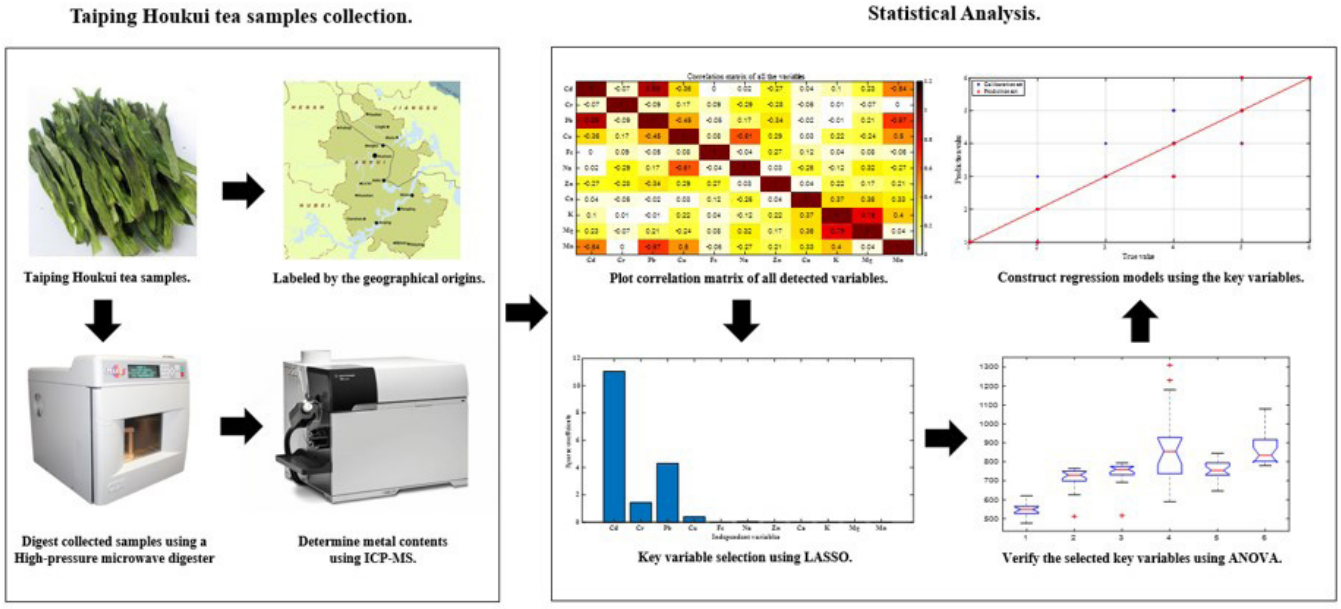

Figure 1. Experimental setup for geographical origin identification of Taiping Houkui tea.

of data sets. The models were validated using a prediction set by applying the coefficient of determination (R^2) (Cakmakyapan & Demirhan, 2017; Pejovic et al., 2018), accuracy ratio (P), and root mean square error ($RMSE$). The results of running the MCCV method 100 times are listed in order, and the median is used as the performance index (Equations 9, 10 and 11).

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{9}$$

$$P = \frac{n_c}{n} \times 100\% \tag{10}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{11}$$

where y_i and $\hat{y}_i$ are the actual and predicted values of the i^{th} test sample, respectively; $\bar{y}$ is the average value of all the test samples, and n_c and n are the number of correctly distinguished samples and the total number of samples, respectively.

3.2 Key metal ions extraction and validation

Owing to differences between the data size and units of the various components, the data of the components must be normalized to fully analyze the effect of each variable on the geographical origin of Taiping Houkui tea. Therefore, standard normal variate (SNV) was used to reduce within-class variance (Wang et al., 2020), which is given by Equation 12,

$$x_{i,j} = \frac{x_{i,j}^{org} - \mu_j}{\sigma_j} \tag{12}$$

where $x_{i,j}^{org}$ is the original data, which represents the j_{th} feature of the i_{th} sample and $\mu_j = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}^{org}$, $\sigma_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{i,j}^{org} - \mu_j\right)^2}$.

The correlation matrix of the components in the Taiping Houkui tea is shown in Table 4. There are some metal ions with large correlation coefficients. It is indicated that these metal ions play a similar role in the classification model (Zhuang et al., 2020). Therefore, it is important to screen out the key factors affecting the geographical origin of Taiping Houkui tea using LASSO. A ten-fold CV was used to select the optimal value of λ. The mean square error (MSE) corresponding to various λ is presented in Figure 2. The λ corresponding to the smallest MSE marked using the blue line was then selected as the optimal penalty coefficient, $\lambda = 0.07$ in this experiment. As a result, the regression coefficient of metal ions whose contribution is relatively small shrinks to zero in the LASSO algorithm. The critical metal ions screened by LASSO were Cd, Cr, Pb, Cu, and Na. Subsequently, ANOVA was used to verify the significance of the selected metal ions. The regression coefficients of the selected metal ions and the corresponding p-values using the F-statistic test are shown in Table 5 when the confidence level was 0.05. All p-values were less than 0.05, verifying that the key variables screened using LASSO significantly influenced the geographical origin of Taiping Houkui tea.

3.3 Results and analysis

It can be observed from Table 5 that the weight of Cd, Cr, Pb, Cu in the regression model is higher. The measurement of insoluble metal ions is more accurate and less noisy than that of soluble metals. Therefore, the selected key variables have a great significance in the practical applications. To avoid contingency and problems with data set partitioning, the MCCV method was run 100 times at random in this experiment, and the sparse

variables obtained by LASSO were used to construct the regression models. The median accuracies of the LASSO on the calibration and prediction sets were 67.9% and 61.1%, respectively, and the median R^2 was 0.882 and 0.842, respectively. The estimation results showed that LASSO achieved high performance.

To demonstrate the performance of the proposed method, PLS-DA was used for comparison. As a classical linear classification method, PLS-DA is commonly used for dimension reduction (Chen et al., 2020). It can effectively reduce the dimensions of the original variables by exporting a few principal components (PCs) from the original variables while retaining as much information as possible about the original data. The optimal number of PCs is nine which was determined using a grid search method combined with K-fold CV.

The experimental results of LASSO and PLS-DA for estimating the geographical origin of Taiping Houkui tea are shown in Table 6. The results showed that LASSO could select the crucial metal ions that affect the geographical origin of Taiping Houkui tea. Therefore, the proposed method could overcome correlations between the metal ions, resulting in superior performance. Simultaneously, the proposed method could overcome overfitting for the R^2 on the calibration set are

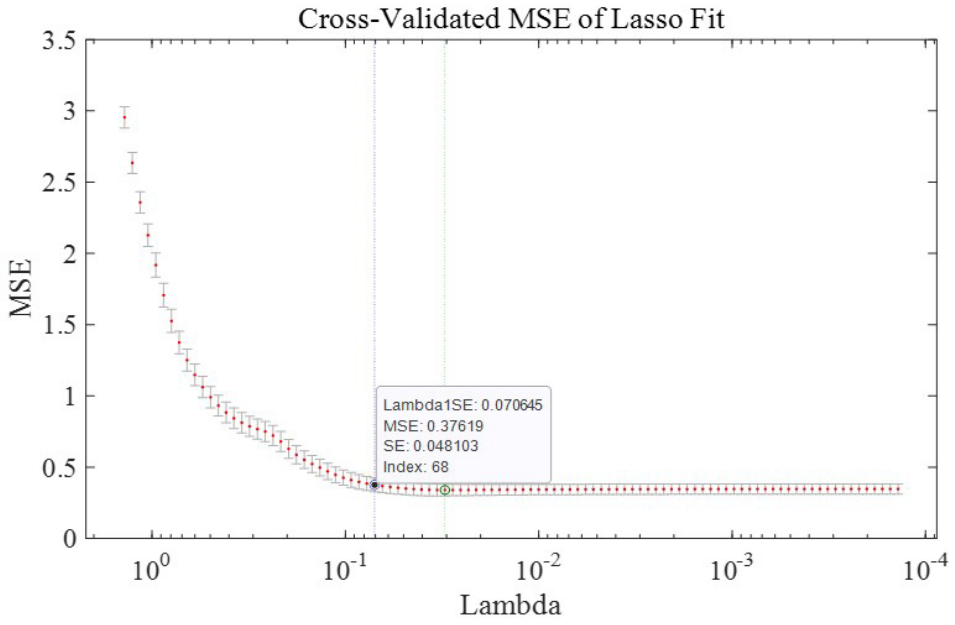

Figure 2. Mean square error corresponding to various λ in the LASSO method.

Table 4. Correlation matrix of all the variables.

Metal ions	Cd	Cr	Pb	Cu	Fe	Na	Zn	Ca	K	Mg	Mn
Cd	1.00										
Cr	-0.07	1.00									
Pb	0.89	-0.09	1.00								
Cu	-0.36	0.17	-0.45	1.00							
Fe	0.00	0.09	-0.05	0.08	1.00						
Na	0.02	-0.29	0.17	-0.61	-0.04	1.00					
Zn	-0.27	-0.28	-0.34	0.29	0.27	0.03	1.00				
Ca	0.04	-0.05	-0.02	0.03	0.12	-0.26	0.04	1.00			
K	0.10	0.01	-0.01	0.22	0.04	-0.12	0.22	0.37	1.00		
Mg	0.23	-0.07	0.21	-0.24	0.08	0.32	0.17	0.36	0.79	1.00	
Mn	-0.54	0.00	-0.67	0.50	-0.06	-0.27	0.21	0.33	0.40	0.04	1.00

Table 5. The coefficients and p-values of the metal ions selected by LASSO.

Metal ions	Cd	Cr	Pb	Cu	Na
Coefficients	-2.31	-0.83	-4.94	0.47	0.04
p-values	2.12e-42	2.39e-7	6.68e-67	2.44e-41	3.00e-60

Table 6. Comparison of 100 MCCV results.

Method	Calibration set		Median RMSE	Prediction set		
	Median R^2	Median accuracy		Median R^2	Median accuracy	Median RMSE
LASSO	0.882	0.679	0.577	0.842	0.611	0.656
PLS-DA	0.857	0.631	0.600	0.827	0.583	0.707

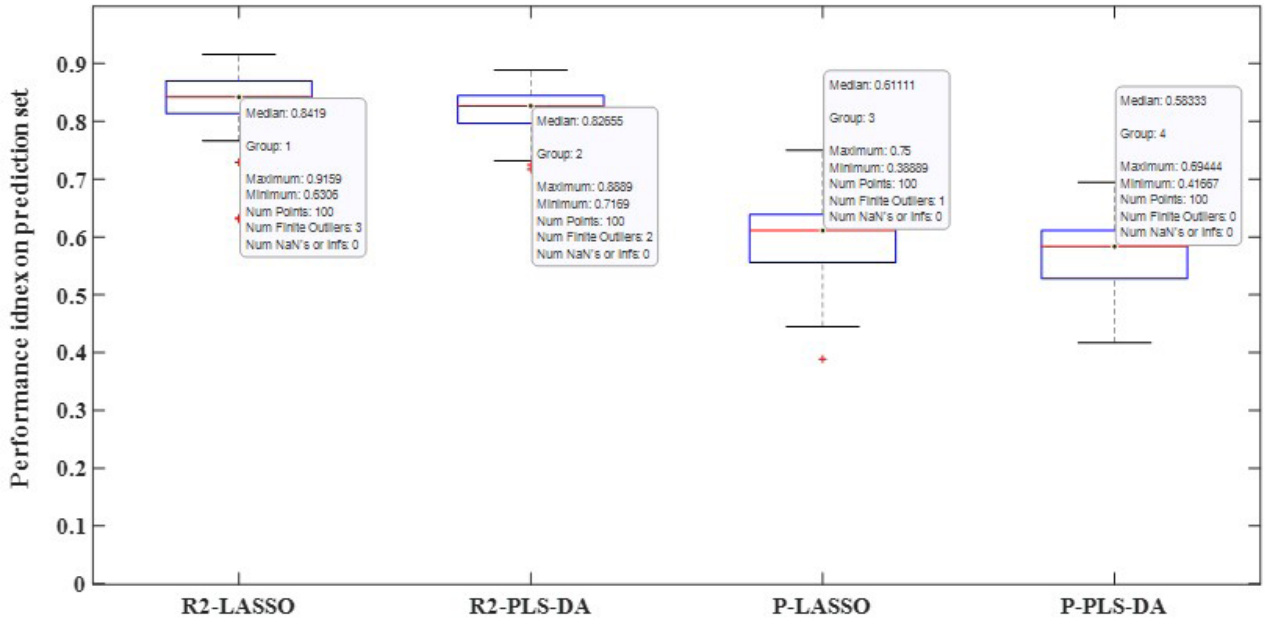

Figure 3. Estimation results of Taiping Houkui tea geographical origin for LASSO and PLS-DA.

much closer to that on the prediction set. In addition, according to the standard RMSE of the prediction set in the table and the Monte Carlo experimental distribution presented in Figure 3, the model constructed by the screened variables using LASSO was robust and achieved high accuracy in the estimation of Taiping Houkui tea quality grade discrimination analysis.

4 Conclusion

In this study, a total of 120 Taiping Houkui tea samples were collected from six villages including core regions and non-core regions. The contents of 11 metal ions in the collected Taiping Houkui tea samples were measured and analyzed using the microwave digestion pretreatment method combined with ICP-MS. An approach for simplifying the identification process based on LASSO was proposed. The screened key variables were tested by ANOVA, which verified the effectiveness of LASSO. The proposed scheme identified key metal ions which can be used to identify the geographical origins of Taiping Houkui tea. The experimental results showed that the data-driven model constructed using the key variables achieved a high and robust prediction performance. Therefore, the detection speed has been effectively improved and the cost has been reduced because of the simplification of the detection process. Therefore, it has significant value for the Taiping Houkui tea market by increasing the related economic benefits.

Acknowledgements

This work was supported by the Major Science and Technology Project of Anhui Province under Grant 202003a06020001.

References

Bhuyan, A., Tudu, B., Bandyopadhyay, R., Ghosh, A., & Kumar, S. (2019). ARMAX modeling and impedance analysis of voltammetric E-tongue for evaluation of infused tea. *IEEE Sensors Journal*, 19(11), 4098-4105. http://dx.doi.org/10.1109/JSEN.2019.2898226.

Bobková, A., Demianova, A., Belej, L., Harangozo, Ľ., Bobko, M., Jurčaga, L., Poláková, K., Božiková, M., Bilčík, M., & Árvay, J. (2021). Detection of changes in total antioxidant capacity, the content of polyphenols, caffeine, and heavy metals of teas in relation to their origin and fermentation. *Foods*, 10(8), 1821. http://dx.doi.org/10.3390/foods10081821. PMid:34441598.

Cakmakyapan, S., & Demirhan, H. (2017). A Monte Carlo-based pseudo-coefficient of determination for generalized linear models with binary outcome. *Journal of Applied Statistics*, 44(14), 2458-2482. http://dx.doi.org/10.1080/02664763.2016.1257585.

Cardoso, V. G. K., & Poppi, R. J. (2021). Non-invasive identification of commercial green tea blends using NIR spectroscopy and support vector machine. *Microchemical Journal*, 164, 106052. http://dx.doi.org/10.1016/j.microc.2021.106052.

Chen, X., Xu, Y., Meng, L., Chen, X., Yuan, L., Cai, Q., Shi, W., & Huang, G. (2020). Non-parametric partial least squares-discriminant analysis

model based on sum of ranking difference algorithm for tea grade identification using electronic tongue data. *Sensors and Actuators. B, Chemical*, 311, 127924. http://dx.doi.org/10.1016/j.snb.2020.127924.

Chen, Y., Gong, M., Nie, X., Qi, Z., Liu, X., Jin, Q., Zhang, X., & Yang, D. (2022). Characterization of botanical origin of selected popular purple *Eleutherococcus tea* grown in Yunnan province of China and quantification of Its anthocyanins using spectrophotometric method. *Food Science and Technology*, 42, e91121. http://dx.doi.org/10.1590/fst.91121.

Doğan, M., Akdoğan, M., Alizada, A., Eroğul, Ö., Sabaner, M. C., Gobeka, H. H., Gülyeşil, F. F., & Seylan, M. A. (2021). Impacts of *Camellia sinensis* fermentation end-product (black tea) on retinal microvasculature: an updated OCTA analysis. *Journal of the Science of Food and Agriculture*, 101(15), 6265-6270. http://dx.doi.org/10.1002/jsfa.11294. PMid:33934371.

Du, Y. P., Kasemsumran, S., Maruo, K., Nakagawa, T., & Ozaki, Y. (2006). Ascertainment of the number of samples in the validation set in Monte Carlo cross validation and the selection of model dimension with Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, 82(1-2), 83-89. http://dx.doi.org/10.1016/j.chemolab.2005.07.004.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1-22. http://dx.doi.org/10.18637/jss.v033.i01. PMid:20808728.

Hidayat, W., Shakaff, A. Y. M., Ahmad, M. N., & Adom, A. H. (2010). Classification of agarwood oil using an electronic nose. *Sensors*, 10(5), 4675-4685. http://dx.doi.org/10.3390/s100504675. PMid:22399899.

Huang, D., Qiu, Q., Wang, Y., Wang, Y., Lu, Y., Fan, D., & Wang, X. (2020). Rapid identification of different grades of huangshan maofeng tea using ultraviolet spectrum and color difference. *Molecules*, 25(20), 12-31. http://dx.doi.org/10.3390/molecules25204665. PMid:33066248.

Huang, J., Breheny, P., & Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science*, 27(4), 481-499. http://dx.doi.org/10.1214/12-STS392. PMid:24174707.

Jin, G., Wang, Y., Li, M., Li, T., Huang, W., Li, L., Deng, W.-W., & Ning, J. (2021). Rapid and real-time detection of black tea fermentation quality by using an inexpensive data fusion system. *Food Chemistry*, 358, 129815. http://dx.doi.org/10.1016/j.foodchem.2021.129815. PMid:33915424.

Li, Y., Sun, J., Wu, X., Lu, B., Wu, M., & Dai, C. (2019). Grade identification of tieguanyin tea using fluorescence hyperspectra and different statistical algorithms. *Journal of Food Science*, 84(8), 2234-2241. http://dx.doi.org/10.1111/1750-3841.14706. PMid:31313313.

Meng, J., Cheng, M., Zhang, K., El Hadi, M. A. M., Zhao, D., & Tao, J. (2022). Beneficial effects of *Paeonia ostii* stamen tea in extending the lifespan and inducing stress resistance on *Caenorhabditis elegans*. *Food Science and Technology*, 42, e76521. http://dx.doi.org/10.1590/fst.76521.

Nourbakhsh, M., Mashinchi, M., & Parchami, A. (2013). Analysis of variance based on fuzzy observations. *International Journal of Systems Science*, 44(4), 714-726. http://dx.doi.org/10.1080/00207721.2011.618640.

Pang, X., Chen, F., Liu, G., Zhang, Q., Ye, J., Lei, W., Jia, X., & He, H. (2022). Comparative analysis on the quality of Wuyi Rougui (*Camellia sinensis*) tea with different grades. *Food Science and Technology*, 42, e115321. http://dx.doi.org/10.1590/fst.115321.

Patocka, J., Bendakovska, L., Krejcova, A., Cernohorsky, T., Resano, M., & Belina, P. (2017). Thallium in spruce needles: a comparison of the analytical capabilities of spectrochemical methods. *Analytical Methods*, 9(4), 705-715. http://dx.doi.org/10.1039/C6AY02760A.

Pejovic, M., Nikolic, M., Heuvelink, G. B. M., Hengl, T., Kilibarda, M., & Bajat, B. (2018). Sparse regression interaction models for spatial prediction of soil properties in 3D. *Computers & Geosciences*, 118, 1-13. http://dx.doi.org/10.1016/j.cageo.2018.05.008.

Qu, Z., & Richtarik, P. (2016). Coordinate descent with arbitrary sampling I: algorithms and complexity. *Optimization Methods & Software*, 31(5), 829-857. http://dx.doi.org/10.1080/10556788.2016.1190360.

Shao, Y., Xuan, G., Hu, Z., & Wang, Y. (2019). Detection of adulterants and authenticity discrimination for coarse grain flours using NIR hyperspectral imaging. *Journal of Food Process Engineering*, 42(7), 232-251. http://dx.doi.org/10.1111/jfpe.13265.

Shevchuk, A., Jayasinghe, L., & Kuhnert, N. (2018). Differentiation of black tea infusions according to origin, processing and botanical varieties using multivariate statistical analysis of LC-MS data. *Food Research International*, 109, 387-402. http://dx.doi.org/10.1016/j.foodres.2018.03.059. PMid:29803464.

Song, Y., Wang, X., Xie, H., Li, L., Ning, J., & Zhang, Z. (2021). Quality evaluation of Keemun black tea by fusing data obtained from near-infrared reflectance spectroscopy and computer vision sensors. *Spectrochimica Acta. Part A: Molecular and Biomolecular Spectroscopy*, 252, 119522. http://dx.doi.org/10.1016/j.saa.2021.119522. PMid:33582437.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B. Methodological*, 73(1), 273-282. http://dx.doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Wang, D. F., Wang, C. H., Wei, Z. G., Qi, H. T., & Zhao, G. W. (2003). Effect of rare earth elements on peroxidase activity in tea shoots. *Journal of the Science of Food and Agriculture*, 83(11), 1109-1113. http://dx.doi.org/10.1002/jsfa.1507.

Wang, Y.-J., Li, T.-H., Li, L.-Q., Ning, J.-M., & Zhang, Z.-Z. (2020). Micro-NIR spectrometer for quality assessment of tea: comparison of local and global models. *Spectrochimica Acta. Part A: Molecular and Biomolecular Spectroscopy*, 237, 118403. http://dx.doi.org/10.1016/j.saa.2020.118403. PMid:32361319.

Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, 48(9), 2839-2846. http://dx.doi.org/10.1016/j.patcog.2015.03.009.

Wright, S. J. (2015). Coordinate descent algorithms. *Mathematical Programming*, 151(1), 3-34. http://dx.doi.org/10.1007/s10107-015-0892-3.

Xia, E., Tong, W., Hou, Y., An, Y., Chen, L., Wu, Q., Liu, Y., Yu, J., Li, F., Li, R., Li, P., Zhao, H., Ge, R., Huang, J., Mallano, A. I., Zhang, Y., Liu, S., Deng, W., Song, C., Zhang, Z., Zhao, J., Wei, S., Zhang, Z., Xia, T., Wei, C., & Wan, X. (2020). The reference genome of tea plant and resequencing of 81 diverse accessions provide insights into its genome evolution and adaptation. *Molecular Plant*, 13(7), 1013-1026. http://dx.doi.org/10.1016/j.molp.2020.04.010. PMid:32353625.

Xu, L., Li, K., Xv, L., Zhang, H., Zhang, Y., Liu, X., Xu, Y., Yin, J., Qin, D., Jin, P., & Du, Q. (2021). Preparation of scented teas by sustained-release of aroma from essential oils-casein nanocomposites. *Lebensmittel-Wissenschaft + Technologie*, 146, 21-35. http://dx.doi.org/10.1016/j.lwt.2021.111410.

Xu, M., Wang, J., & Gu, S. (2019). Rapid identification of tea quality by E-nose and computer vision combining with a synergetic data fusion strategy. *Journal of Food Engineering*, 241, 10-17. http://dx.doi.org/10.1016/j.jfoodeng.2018.07.020.

Ye, X., Jin, S., Wang, D., Zhao, F., Yu, Y., Zheng, D., & Ye, N. (2017). Identification of the origin of white tea based on mineral element content. *Food Analytical Methods*, 10(1), 191-199. http://dx.doi.org/10.1007/s12161-016-0568-5.

Zhang, J., Yang, R., Li, Y. C., & Ni, X. (2021). The role of soil mineral multi-elements in improving the geographical origin discrimination

of tea (*Camellia sinensis*). *Biological Trace Element Research*, 199(11), 4330-4341. http://dx.doi.org/10.1007/s12011-020-02527-8. PMid:33409909.

Zhang, L., Wang, X., Huang, G.-B., Liu, T., & Tan, X. (2019). Taste recognition in E-tongue using local discriminant preservation projection. *IEEE Transactions on Cybernetics*, 49(3), 947-960. http://dx.doi.org/10.1109/TCYB.2018.2789889. PMid:29994190.

Zhuang, J., Dai, X., Zhu, M., Zhang, S., Dai, Q., Jiang, X., Liu, Y., Gao, L., & Xia, T. (2020). Evaluation of astringent taste of green tea through mass spectrometry-based targeted metabolic profiling of polyphenols. *Food Chemistry*, 305, 125507. http://dx.doi.org/10.1016/j.foodchem.2019.125507. PMid:31622805.

Zuo, Y., Tan, G., Xiang, D., Chen, L., Wang, J., Zhang, S., Bai, Z., & Wu, Q. (2021). Development of a novel green tea quality roadmap and the complex sensory-associated characteristics exploration using rapid near-infrared spectroscopy technology. *Spectrochimica Acta. Part A: Molecular and Biomolecular Spectroscopy*, 258, 119847. http://dx.doi.org/10.1016/j.saa.2021.119847. PMid:33940571.