Relation Model of Soil Nutrients and Tobacco Quality Based on Principal Component and Regression Analysis

Houli Hu, Jiande Wu*, Xiaodong Wang
Faculty of Information Engineering and Automation
Kunming University of Science and Technology
Kunming 650500, China
xiaoli01030@126.com

Kuake Huang
Qujing Branch, Yunnan Tobacco Company
Kunming 655000, China

***Abstract*—In order to study the relationship of tobacco soil nutrients and chemical indicators, principal component analysis and stepwise regression analysis were used to establish the relation model between soil nutrients and tobacco quality. In the first, discussing the correlation of soil nutrients and the major tobacco chemical composition; Then , using principal component analysis to analyze the impact of factors of the leaf chemical indicators , eliminated the intrinsic link and repeat information between properties, and extracting the principal component which strong influence on the chemical composition of tobacco; And finally stepwise regression analysis was used to established the equation between soil nutrients and tobacco chemical components.**

Keywords- principal component analysis; stepwise regression analysis; soil nutrients；Tobacco quality

I. INTRODUCTION

Soil is the base of planting tobacco and the nutrient sources for tobacco, it directly impact on the growth and nutrition levels of tobacco, thereby affecting the yield and quality of tobacco, the suitable soil conditions is an important foundation for the production of quality tobacco[1]. Therefore, the study of relationship between nutrients in the soil and leaf quality and clear of the quantitative level between them, it can provide a scientific basis to improve tobacco soil environment, enhance quality leaf and foster high-quality tobacco.

It is a complex multi-factor multivariable problem to soil nutrient effects on plant nutrition [2]. On such issues, the popular ways are simple regression analysis and correlation analysis in the past [3-5], these methods only examine the relationships between a number of variables and a certain variable, they reflect neither the internal relationship between a whole set of variables with another whole set of variables, nor able to clearly explain relationships between the Internal variable groups or external groups [6]. On analysis of soil nutrient and chemical components of tobacco, we often need to create a relationship model between variables affecting chemical indicators and chemical composition variables. General multivariate linear regression model in a high degree of correlation between the independent variables, and it has a serious multi commonality, thus it will affect the model fitting goodness. Principal component analysis (PCA) is a statistical method which is the study of the multiple indicators converted to fewer composite indicators, and it is the result of linear transformation and discards some of the information and high-dimensional variable space to reduce the dimension, the relevant process variables compressed to a few independent variables. It also reflected the original information of multiple variables [7]. The method of combining Principal component analysis and stepwise regression analysis not only can reduce the number of variables, but also solve the problem of multiple common.

This article is a correlation analysis between soil nutrient and chemical composition of the tobacco sample in Yunnan seven tobacco growing areas, it aimed at in-depth understanding of how the soil nutrients to affect the chemical composition. By using the method of combining Principal component analysis and stepwise regression analysis, it not only reduce the number of variables, solve the problem of multiple common, but also establish a relational model between soil nutrient and chemical composition of tobacco, which provide a scientific basis for planting the high-quality tobacco.

II. SOURCES AND DETERMINATION OF SAMPLE

A. Sources of samples

Samples are taken from the tobacco growing areas of Yunnan Luliang, Qilin District, Shizong, Fuyuan, Malong, Zhanyi, Huize, where the tobacco leaves grows on three different types of soil and its corresponding root zone soil. Flue-cured tobacco samples selected local cultivars, using fixed-grade, fixed leaf sampling method, samples were collected 59 from the middle of tobacco leaves.

B. Maintaining the Integrity of the Specifications

(1) Determination of soil sample determination of soil samples including PH value, organic matter, total N, total P, and total K, Hydrolytic N, Available P, Available K of eight indicators. Determination of soil nutrients specifically refer to the literature [8].

(2) Determination of tobacco sample determination of tobacco samples including the total sugar, reducing sugar, total N, nicotine, potassium, chloride, starch of seven indicators. Determination of Chemical composition indicators of tobacco refer to the literature [9].

C. *The data processing*

a statistical method is defined as by SPSS17.0 statistical software, using principal component analysis and stepwise regression analysis as a statistical method.

III. RESULTS AND ANALYSIS

A. *Test data descriptive statistics and related analysis*

The selected eight indicators of soil nutrients and seven tobaccos leave chemical composition of the raw data for statistical to analysis the results described in Table 1.

TABLE I. TOBACCO RAW DATA FOR MODELING DESCRIPTIVE STATISTICS

Index	*Minimm*	*Maxima*	*Mean*	*Standard deviation*	*Coefficient of variation（%）*	*Skewness*	*Kurtosi*
PH	4.39	7.58	6.203	0.750	12.091	-0.176	-0.577
Organic matter /%	1.38	6.07	3.613	0.972	26.903	-0.227	-0.228
Total N (mg/kg)	0.064	0.351	0.176	0.047	26.705	0.848	2.707
Total P (mg/kg)	0.036	0.254	0.106	0.041	38.679	0.951	1.550
Total K (mg/kg)	0.160	2.658	0.798	0.510	63.910	1.338	2.103
Hydrolytic N (mg/kg)	57.58	205.10	126.084	31.667	25.116	0.041	-0.277
Available P(mg/kg)	7.58	84.45	28.543	17.661	61.875	1.253	1.176
Available K(mg/kg)	38.10	350.13	151.269	72.288	47.788	0.576	-0.170
Total sugar /%	19.33	39.98	30.421	4.278	14.063	-0.344	-0.113
Deoxidize sugar /%	11.64	30.53	22.221	4.151	18.681	0.063	-0.266
N/%	0.94	3.15	1.946	0.378	19.424	0.240	0.982
Nicotine /%	1.12	3.66	2.156	0.614	28.479	0.304	-0.840
K/%	1.09	2.23	1.605	0.276	17.196	0.172	-0.734
Cl/%	0.05	1.14	0.273	0.241	88.278	1.616	2.870
Starch /%	1.00	5.73	2.960	1.074	36.284	0.419	-0.113

TABLE II. THE CORRELATION MATRIX BETWEEN SOIL NUTRIENTS AND TOBACCO CHEMICAL COMPO

	PH	*Organic matter*	*Total N*	*Total P*	*Total K*	*Hydro-lytic N*	*Avail-able p*	*Avail-able k*	*total sugar*	*deoxidize sugar*	*N*	*nicotine*	*K*	*Cl*	*starch*
PH	1.000														
Organic matter	-0.167	1.000													
Total N	0.033	0.799	1.000												
Total P	-0.029	0.374	0.372	1.000											
Total K	0.322	-0.203	0.024	-0.230	1.000										
Hydrolytic N	-0.399	0.909	0.722	0.336	-0.208	1.000									
Available P	-0.064	-0.107	-0.023	0.354	-0.021	-0.142	1.000								
Available k	-0.042	0.310	0.313	0.432	0.057	0.280	0.390	1.000							
total sugar	-0.029	-0.148	-0.103	0.218	0.057	-0.015	0.277	0.037	1.000						
deoxidize sugar	-0.050	-0.126	-0.170	0.098	0.024	0.043	0.009	-0.105	0.819	1.000					
N	-0.075	0.271	0.206	-0.032	-0.288	0.142	-0.033	0.092	-0.647	-0.690	1.000				
nicotine	0.128	0.044	0.142	0.160	0.020	-0.165	0.276	0.142	-0.440	-0.597	0.587	1.000			
K	-0.215	0.012	0.010	-0.276	0.112	0.133	-0.254	-0.117	-0.312	-0.165	0.021	-0.245	1.000		
Cl	0.080	0.112	0.155	-0.073	0.125	-0.047	0.011	0.080	-0.080	-0.141	0.216	0.148	-0.243	1.000	
starch	-0.138	0.073	0.019	0.285	-0.124	0.173	0.116	0.020	0.594	0.504	-0.502	-0.241	-0.266	-0.154	1.000

Indicated in Table 1, the variation of nutrient indicators was moderate intensity variation, in which the largest variability of, the smallest variability of total is PH, explaining the difference is that total K content maximum, and the sample is relatively uniform PH in the selection of tobacco soil samples. In the selected leaf chemical indicators, tobacco total sugar, reducing sugar, total N, K and other variation is relatively uniform, while Cl varied greatly.

Indicated in Table 2, the PH value of soil nutrients compared with nicotine showed a significant positive correlation; organic matter in soil nutrients with leaf chemical composition of total sugar, deoxidize sugar was a significantly negative correlation, and with total N, Cl was a significant positive correlation; total P, available P in soil nutrients with chemical composition of K was a significantly negative correlation, with other correlation was not significant; K, Hydrolytic N, available K in soil nutrients with chemical composition correlation is not significant. To further understand the relationship between the Yunnan tobacco leaf and tobacco soil nutrient quality, should think highly of the analysis of its principal component.

B. *Principal component extraction*

According to table 3 and table 4, we could find that the first three principal components of the cumulative contribution rate has reached 85.635%, basically reflects the original variable information. The first principal component to the total variance contribution rate of 49.121% , represents the organic matter, total N, total P, hydrolytic N; The second principal component to the total variance contribution rate of 28.609%，represents the available P, available K; The third principal component to the total variance contribution rate of 19.795%, represents the PH, total K.

TABLE III. EIGENVALUES AND CUMULATIVE VARIANCE CONTRIBUTION RATE

Main ingredient	*Eigenvalues*	*Contribution rate %*	*Cumulative contribution rate %*
1	3.130	49.121	49.121
2	1.569	28.609	68.730
3	1.344	19.795	85.635

TABLE IV. MAIN COMPONENT EIGENVECTORS

	Main ingredient		
	1	2	3
PH	-0.290	0.204	0.730
Organic matter	0.920	-0.246	0.137
Total N	0.824	-0.097	0.400
Total P	0.600	0.506	-0.107
Total K	-0.262	0.142	0.745
Hydrolytic N	0.908	-0.319	-0.018
Available P	0.089	0.838	-0.236
Available K	0.511	0.614	0.104

The principal component scores of samples of tobacco were calculated by the principal component eigenvectors and the original value of the variable，The first three principal components as a linear combination:

$$Z1 = -0.290X_1 + 0.920X_2 + 0.824X_3 + 0.600X_4 - 0.262X_5 + 0.908X_6 + 0.089X_7 + 0.511X_8$$

$$Z2 = 0.204X_1 - 0.204X_2 - 0.097X_3 + 0.506X_4 + 0.142X_5 - 0.319X_6 + 0.838X_7 + 0.614X_8$$

$$Z3 = 730X_1 + 0.137X_2 + 0.400X_3 - 0.107X_4 + 0.745X_5 - 0.018X_6 - 0.236X_7 + 0.104X_8$$

C. *Regression model and F-test*

The tobacco chemical constituents of total sugar as the dependent variable and three principal components score as independent variables were to stepwise regression analysis. In accordance with the least principle of regression equation variable to select and build the regression models， the regression results of tobacco total sugar shown in Table 4. Similarly, you can also separately establish the regression model of reducing sugar, total N, nicotine, K, Cl and starch.

TABLE V. LINEAR REGRESSION RESULTS

Constant and Variable	*Coefficient*	*Standard error*	*t*	*Sig.*
Constant	23.545	5.427	3.170	0.000
Z1	4.840	1.384	10.577	0.000
Z2	7.014	1.861	12.791	0.000
Z3	5.751	2.721	2.536	0.000

According to the results in Table 5, a regression equation is as follows:

$Y = 4.840Z1 + 7.014Z2 + 5.751Z3$, The equation correlation coefficient after correction is $R^2 = 0.505$. The equation for the establishment of the significance test, the test results of the regression equation was shown in Table 6.According to the Table 6，we can find that he regression results is significant when $F = 68.985 > F_{0.05}\left(F_{0.05} = 2.94\right)$, that it exists a significant linear statistical relationship between soil fertility elements and tobacco chemical components. As the $Sig. = 0.000 < 0.05$, the results illustrate the probability of error is zero. Therefore, the equation is to be reliable.

TABLE VI. REGRESSION EQUATION F TEST

	Squares	*df*	*Mean square*	*F*	$F_{0.05}$	*Sig.*
Regression	78.518	3	26.173	68.985	2.94	0.000
Residuals	113.355	53	2.139			
Total	191.873	56				

IV. CONCLUSION

The paper is a correlation analysis between soil nutrient and chemical composition of the tobacco sample in Yunnan seven tobacco growing areas, discussing the correlation between each index. Principal component analysis was used on the impact factor of chemical composition of flue-cured tobacco，reducing the dimension of the sample space and eliminates autocorrelation between the input variables， stepwise regression analysis was used to established the equation between soil nutrients and tobacco chemical components. Results show that the model in a certain extent, can be based on the results of the determination of soil nutrient quality of Yunnan flue-cured tobacco smoke area prediction, it plays an important role for cultivating high quality tobacco leaves.

ACKNOWLEDGMENT

Jiande Wu is corresponding author. This work is supported by the National Natural Science Founder of China (51169007), Science & Research Program of Yunnan province (2011CI017, 2011DA005, 2012CA022).

REFERENCES

[1] Jie Chen, Yuanju Tang. “Analysis of Soil Nutrients Affecting Chemical Constituents of Tobacco in Guizhou,” Crops, pp. 68–70, January 2008.

[2] Zicheng Xu, Hanqian Xiao, Xianzhang Zhao. “Statistical Methods for Evaluating the Status of Soil Fertility in Tobacco-growing Areas, ” Chinese Journal of Soil Science, vol. 5,2004, pp. 32-35.

[3] Xia Ni, Weikun Lu, Hongbo Cha. “Research Progress of the Effects of Ecological Factors on the Leaf Chemical Components of Tobacco, ” Journal of Anhui Agri.Sci. vol. 40, 2012, pp.1355-1359.

[4] Xinmiao Tang, Feng Wang, Chunmei Ji. “Study on the Relationship of Soil Nutrient and Tobacco Chemical Components in Guizhou Province, ”Journal of Henan Agricultural Science, vol. 40,2011, pp.84-88.

[5] Yongan Chen, Yagang Zhu, Jiabo Chen. “Discussion on the Relation between Ecological Factors and Some Chemical Components in Flue-cured Tobacco, ” Journal of Anhui Agri.Sci. vol.35, 2007.

[6] Song Qin, Dacui Liu, Jing Liu. “Influence of Soil Fertility on Tobacco Chemical Components and Quality, ” Chinese Journal of Soil Science, vol.38, 2007, pp. 901-905.

[7] Lei Wang. “Regression forecast model of support vector machine based on principal component analysis,” Information Techn-ology, vol.12, 2008, pp. 58-60.

[8] Yanyan Li, Zicheng Xu, Jinping Wang. “Relationship between Sugar Contents in Flue- cured Tobacco Leaves and Soil Nutrients in Tobacco- growing Areas in Hunan,”Chinese Journal of Soil Science, vol.38, 2007, pp. 911-914.

[9] Keyu Yan, Xianzhang Zhao. Tobacco Classification, Beijing:CHINA AGRICULTURE PRESS, 2003, pp. 22-23.