Comparison and Verification of Methods for Multivariate Statistical Analysis and Regression in Crop Modelling

Z.H. Jiang, J. Zhang, C.H. Yang, Y. Rao, S.W. Li
School of Information and Computer, Anhui Agricultural University, China

Abstract—**In crop modelling, factor analysis and regression type have direct influence on the accuracy of model, but the application of these methods usually depends on the experience. In this paper, the performance of some common methods of statistical analysis and regression model was compared and verified, in order to avoid the blindness in crop modelling. The monitoring data of growth environment and photosynthesis of tomato, pumpkin and cucumber was obtained by PTM-48A, for the object variable of CO_2 exchange rate, selectivity on the main environmental factors by correlation analysis and path analysis were quantify compared, the performance of four kinds of multivariate binomial regression equations was compared in respects of complexity and accuracy, then the effectiveness of modelling was verified with the selected optimized multivariate statistical analysis and regression equation. Results showed that path analysis was more comprehensive than correlation to discrimination the variables, and the pure quadratic was more suitable to crop modelling because of its simple structure and high accuracy. The conclusion of the paper has general applicability, and offers a useful reference and guide for the other crops' modelling.**

Keywords-crop model; multivariate statistical analysis; regression; comparison

I. INTRODUCTION

Crop model was one of the core technologies supporting the Precision Agriculture, and it was also the essential part of the intelligent production [1]. Crop model was to obtain accurately of growth environmental factors and physiological parameters, and to indicate quantitatively the relationship of them [2]. The typical crop models are DSSAT (Decision Support System of Agricultural Technology Transfer), APSIM (Agricultural Production Systems sIMulator), CERES (Crop and Environment Research Synthesis), and so on. Because of the difference of growth condition, location and variety, every crop model has its own adaptation and target. Based on the specific conditions, the universal model should be improved to satisfy the actual demand. The improved process not only demands the complete agriculture knowledge, but also needs high-level mathematics and computer resources. So the application of typical crop model is not widely, and the small sample size dedicated crop model for the actual demand is common in practice.

To build up a crop model, firstly the independent variables and dependent variables, which have large influence on the object variable of model, should be selected using methods of multivariate statistical analysis, using the professional data process tools like Excel, SPSS, and SAS. Then regression model or the other models should be used to build up the crop model. The practicality and accuracy of model should be verified, using the test data to predict the unknown variable. The common multivariate statistical analysis includes standard correlation analysis [3], principal components analysis [4], path analysis [5], and so on. Regression analysis is generally used in mathematical modelling. But the independent variables and form of equation are always chosen by experience. And empirical equations are often multitude and inconsistent [6].The existing small sample size crop model is well-directed but has low adaptability. Although the model has high accuracy, the multivariate statistical analysis and regression model are selected according to the experience, instead of the quantification comparison results of the methods [7]. The empirical and blind selection becomes the hindrance of application and development of model, which is also the difficulty of agricultural informatics.

In the paper, methods of multivariate statistical analysis and regression were compared, using the monitoring data of growth environment and photosynthesis of the tomato, pumpkin and cucumber. By the methods of correlation analysis and path analysis, the environmental characters impacting CO_2 exchange of the tomato and pumpkin were picked out to put the analysis methods into comparison. The regression equations were found to compare the models in respects of complexity and accuracy. The data of cucumber were used to verify the selected best multivariate statistical analysis and regression equation.

II. MATERIALS AND METHODS

A. Experimental Data

In the paper, three common outdoor vegetables, including tomato, pumpkin and cucumber, were continuously monitored for at least 24 hours, by using the PTM-48A photosynthesis monitor of B.F. Agritech Company [8][9], expecting to obtain the data of their photosynthetic physiological characteristic and growth environment. The monitoring time interval was uniform and set to 30 minutes, and the air flow was set to 0.84 LPM. The photosynthetic physiology and environmental factors online monitoring system includes five modules: the LC-4B leaf chamber, the multiple environmental factors sensors, the system console, the communication module and computer [10].

PTM-48A can obtained data of 31 variables for single monitor. The monitored data can be divided into two parts, the first part of which was four groups of data of physiological characteristic, and the rest was the ambient environmental parameters. We picked out the average value of four groups of CO2 exchange rates as the object variable, and the ambient environmental parameters as independent variables to study the influence of growth environment on the CO2 exchange rate of crop. In order to decrease the amount of experimental work and study the physiological change of the whole crop, we took the average physiological value of four leaves as the physiological data of the whole crop.

Six environmental factors which were the independent variables, were summarized as follow: Air Temperature Ta (°C), Relative Humidity RH (%RH) , Radiation Ra (micromole/m2*s), Atmospheric Pressure Pa (mbar), Vapour Pressure Deficit Vpd (kPa), Dew Point Dp (°C), defined as X1~X6. The CO2 Exchange E (micromole/m2*s)) was defined as the independent variable Y.

B. Analysis Methods

The analysis and process of experimental data was as Figure 1.

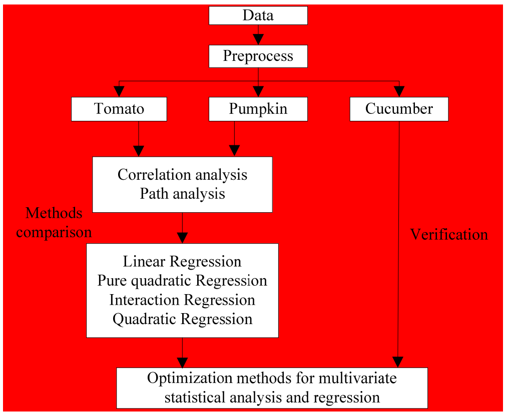

FIGURE I. FLOWCHART OF ANALYSIS AND PROCESS

At first, the monitoring data was pre-processed to convert the output table into the table which the processing software can read. Forty-eight groups of data in 24 hours were chosen from the physiological and environmental data of the tomato and pumpkin as the experimental samples. The experimental samples were used for methods comparison, and the data of cucumber was for verification. The cucumber's data were divided into two parts, the first 48 groups of which were sampled data for modelling, and the rest were test data for predicting.

The pre-processed data was multivariate analysed, using two methods, namely the correlation analysis and path analysis, for the purpose of getting the environmental factors that affect the rate of CO2 exchange. In addition, the two methods were compared. In this paper, the Person simple correlation coefficient [11] [12] was adopted to indicate the relation between variables [13].

$$R_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}} \qquad (1)$$

Here, $C_{ij} = \mathrm{cov}(x_i, x_j)$ was the covariance of x_i and x_j. The effect of environmental factors on CO2 exchange was expressed by the direct path coefficient $\rho_i (i = 1,2,...,n)$, which was calculated as follow [14]:

$$\begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_n \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} & \cdots & c_{1n} \\ c_{21} & c_{22} & c_{23} & \cdots & c_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ c_{n1} & c_{n2} & c_{n3} & \cdots & c_{nn} \end{bmatrix} \begin{bmatrix} r_{1y} \\ r_{2y} \\ \vdots \\ r_{ny} \end{bmatrix} \qquad (2)$$

n was the number of independent variables, y was the dependent variable, r was their correlation matrix, and c=r-1was the invertible matrix of r. The indirect path coefficient $b_{ij} = r_{ij}\rho_j$ was the product of the direct path coefficient and correlation coefficient, indicating the influence of independent variable xi on dependent variable y through xj.

Regression analysis was carried out, regarding the analysed environmental factors as independent variable and the CO2 exchange as a dependent factor. After determining the independent and dependent variables, the suitable model was selected. Then the coefficients were calculated by the least square method, according to the historic statistics. The multivariate binomial regression includes linear, pure quadratic, interaction and quadratic.

$$\text{Linear: } y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m \qquad (3)$$

$$\text{Purequadratic: } y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{j=1}^{m} \beta_{jj} x_j^2 \qquad (4)$$

$$\text{Quadratic: } y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \sum_{1 \le j \le k \le m} \beta_{jk} x_j x_k \qquad (6)$$

In the paper, the RMSE (Root Mean Square Error), standard error, was adopted as the standard of evaluating the curacy of the model [15]. The RMSE can be computed from the definition of it in Eq. (7).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (OBS_i - SIM_i)^2}{n}} \qquad (7)$$

In Eq. (7), n was the number of samples, OBSi is the observed value, and SIMi is the simulated value. Taking account of two aspects including accuracy and complexity comprehensively, the optimal regression model was picked out. Using the optimal method of multivariate analysis, the modeling data of cucumber was analyzed and the regression model was developed. Practicability and accuracy was verified and the results were compared with that of the other prediction models.

III. RESULTS AND DISCUSSION

A. Multivariate Analysis of Tomato and Pumpkin

Six environmental factors, including Air Temperature Ta, Relative Humidity RH, Radiation Ra, Atmospheric Pressure Pa, Vapour Pressure Deficit Vpd, and Dew Point Dp, were chosen as the variables affecting CO2 Exchange E. The chosen environmental factors were defined as X1~X6. The CO2 Exchange E was defined as the independent variable Y. In Table 1, the Person correlation coefficients between CO2 exchange and its impact factors were calculated from Eq. (1). In Table 1, the upper halves of cells were the correlation coefficients of factors in tomato, and the lower halves were those in pumpkin.

TABLE I. CORRELATION OF CO2 EXCHANGE AND ENVIRONMENTAL FACTORS IN TOMATO (T) AND PUMPKIN (P)

		X_1	X_2	X_3	X_4	X_5	X_6
X_2	T	-0.8402					
	P	-0.7560					
X_3	T	0.8031	-0.9183				
	P	0.7936	-0.9140				
X_4	T	-0.2837	0.0284	-0.0130			
	P	-0.3102	0.0792	-0.0259			
X_5	T	0.8611	-0.9962	0.9297	-0.0470		
	P	0.8500	-0.9720	0.9177	-0.1461		
X_6	T	-0.3652	0.8076	-0.7275	-0.2405	-0.7857	
	P	0.8768	-0.3487	0.4735	-0.3845	0.5023	
Y	T	0.8675**	-0.8435**	0.9234**	-0.1355	0.8513**	-0.5254**
	P	0.7820**	-0.8168**	0.9418**	0.0979	0.8188**	0.5311 **

* P<0.05; ** P<0.01.

As shown in Table 1, the results of the pumpkin were consistent with that of the tomato. There were five environmental factors greatly affecting the E, including Ta, RH, Ra, Vpd and Dp, and Pa didn't express no significant correlation with CO2 exchange. Path analysis was carried out. Estimates of direct and indirect path coefficient are presented in Table 2.

TABLE II. .PATH ANALYSIS OF CO2 EXCHANGE AND ENVIRONMENTAL FACTORS IN TOMATO (T) AND PUMPKIN (P)

X_i	Direct path coefficient to Y		Indirect path coefficient via X_j			The summation of indirect impact		The summation of impact on Y	
	T	P	j	T	P	T	P	T	P
X_1	3.0900	-4.5518	2	0.5760	2.2371	-2.2225	5.338	0.8675	0.7820
			3	0.7959	0.6069				
			4	-0.0106	-0.0628				
			5	-4.3416	-0.5358				
			6	0.7578	3.0883				
X_2	-0.6855	-2.9593	1	-2.5963	3.4410	-0.1580	2.1425	-0.8435	-0.8168
			3	-0.9100	-0.6991				
			4	0.0011	0.0160				
			5	5.0230	0.6127				
			6	-1.6757	-1.2281				
X_3	0.9910	0.7648	1	2.4817	-3.6123	-0.0675	0.1769	0.9235	0.9418
			2	0.6295	2.7049				
			4	-0.0005	-0.0053				
			5	-4.6877	-0.5784				
			6	1.5095	1.6679				
X_4	0.0372	0.2024	1	-0.8766	1.4118	-0.1727	-0.1046	-0.1355	0.0979
			2	-0.0195	-0.2342				
			3	-0.0128	-0.0198				
			5	0.2372	0.0921				
			6	0.4990	-1.3544				
X_5	-5.0421	-0.6303	1	2.6607	-3.8690	5.8935	1.4491	0.8513	0.8188
			2	0.6829	2.8765				
			3	0.9213	0.7019				
			4	-0.0018	-0.0296				

			6	1.6302	1.7694				
X_6	-2.0749	3.5224	1	-1.1286	-3.9909	1.5495	-2.9914	-0.5254	0.5311
			2	-0.5537	1.0318				
			3	-0.7210	0.3622				
			4	-0.0089	-0.0778				
			5	3.9616	-0.3166				

NOTE: X_J, WHICH WAS DIFFERENT FROM X_I, WAS A VARIABLE OF X_1~X_6.

From the information in Table 2, the results of path analysis in tomato and pumpkin, which showed that Pa had minimal effect on E, were consistent with that of correlation analysis. In addition, Pa had minimal direct effect on E in tomato, and Ra had minimal indirect effect followed by RH and Pa. While in pumpkin, Pa had both minimal direct and indirect effect on E. Based on the direct, indirect and total effect among the variables, Ta, RH, Ra, Vpd and Dp, were picked out as the independent variable affecting E.

Therefore, path analysis could not only reflect the relation between independent variables and dependent variable, but also indicate the contribution to dependent variable via other independent variables.

B. Regression Modeling

According to the results of multivariate analysis, five impact factors, containing Ta, RH, Ra, Vpd and Dp, were picked out as the independent variable. The multivariate regression equation was found using these parameters and E. The comparison of four models' performance on complexity and accuracy was shown in Table 3. Here the number of coefficients was used to measure the complexity, and the RMSE is the measurement of accuracy.

TABLE III. COMPARISON OF FOUR BINOMIAL REGRESSION MODELS

		linear	pure quadratic	interaction	quadratic
Complexity		6	11	16	21
RMSE	Tomato	0.1141	0.0704	0.0698	0.0681
	Pumpkin	0.1870	0.1304	0.1038	0.0779

From Table 3, the linear model was the simplest model, but its accuracy was the worst. Although the accuracy of full quadratic model was the best, this model had the most coefficients. The accuracy of pure quadratic, interaction and full quadratic models are similar, and the pure quadratic model had the substantially lower complexity than others. Following the growth of complexity, the error decreased, but did not decrease linearly as the complexity increased. The decrease rate of error tended to diminished. On the whole, pure quadratic is the most optimal model among the four models because of its simpler structure and higher accuracy.

C. Modeling and Verification

Since the experimental results of tomato and pumpkin showed that path analysis and pure quadratic were the optimal methods of analysis and modeling, we use the data of cucumber to verify the conclusion. The monitored data of cucumber were divided into two parts, the first 48 groups of which were sampled data for modeling, and the rest were test data. Using the modeling data of cucumber, we used path analysis to analyze the influence among the variables. The result was shown in Table 4.

From Table 4, the environmental factors except Pa had great impact on CO2 exchange. Then the environmental characters related to CO2 exchange were obtained by the method of path analysis. These environmental factors contain Ta, RH, Ra, Vpd, and Dp. On the basis of that, we found a quadratic regression equation, taking the calculated environmental factors as inputs, and CO2 exchange rate E as output. The pure quadratic regression equation was as Eq. (8).

TABLE IV. PATH ANALYSIS OF CO2 EXCHANGE AND ENVIRONMENTAL FACTORS IN CUCUMBER

		X_1		X_2		X_3		X_4		X_5		X_6	
Direct path coefficient to Y		7.3636		2.0796		0.7517		0.1609		-1.4298		-4.0549	
Indirect path coefficient	Indirect path coefficient via X_j	X_2	-1.8865	X_1	-6.6801	X_1	6.4627	X_1	2.0785	X_1	7.1332	X_1	7.1035
		X_3	0.6598	X_3	-0.5501	X_2	-1.5218	X_2	-1.1664	X_2	-1.9924	X_2	-1.5921
		X_4	0.0454	X_4	-0.0903	X_4	0.0367	X_3	0.1712	X_3	0.6281	X_3	0.6645
		X_5	-1.3851	X_5	1.3699	X_5	-1.1946	X_5	-0.5660	X_4	0.0637	X_4	0.0145
		X_6	-3.9116	X_6	3.1043	X_6	-3.5843	X_6	-0.3643	X_6	-3.5679	X_5	-1.2581
	Summation of indirect impact	-6.478		-2.8463		0.1987		0.1530		2.2647		4.9323	
Summation of impact on Y		0.8856		-0.7667		0.9504		0.3139		0.8349		0.8775	

NOTE: X_J, WHICH WAS DIFFERENT FROM X_I, WAS A VARIABLE OF X_1~X_6.

$$E = -170.1205 - 0.3611T_a + 2.703R_H + 0.0474R_a + 30.8967V_{pd} + 0.5306D_p + 0.1455T_a^2 - 0.00092R_H^2 - 0.0001R_a^2 - 17.5549V_{pd}^2 - 0.2195D_p^2 \quad (8)$$

In order to compare the pure quadratic model and the others, we also found the linear equation, interaction equation and full quadratic equation. The equations were as Eq. (9-11). And the comparison of four equations' performance on complexity and accuracy was shown in Table 5.

Linear equation:

$$E = -50.3625 + 4.0850T_a + 0.4454R_H + 0.03R_a - 13.521V_{pd} - 3.6346D_p \quad (9)$$

Interaction equation:

$$\begin{aligned} E = {} & 67.9611 + 26.0816T_a - 0.7971R_H + 0.3346R_a - 55.5630V_{pd} \\ & - 42.6381D_p - 0.2339T_aR_H + 0.0133T_aR_a - 11.8382T_aV_{pd} \\ & - 0.0424T_aD_p - 0.0018R_HR_a - 0.0884R_HV_{pd} + 0.4176R_HD_p \\ & - 0.0412R_aV_{pd} - 0.0266R_aD_p - 20.4119V_{pd}D_p \end{aligned} \quad (10)$$

Quadratic equation:

$$\begin{aligned} E = {} & (1.1244 - 0.2663T_a - 0.0061R_H + 0.0006R_a + 2.4858V_{pd} \\ & + 0.1203D_p + 0.0016T_a^2 - 0.0000T_aR_H - 0.1769T_aR_a \\ & - 0.0159T_aV_{pd} - 0.0000T_aD_p - 0.0224R_H^2 - 0.0001R_HR_a \\ & + 0.0003R_HV_{pd} + 0.0001R_HD_p + 0.1517R_a^2 + 0.0129R_aV_{pd} \\ & - 0.0001R_aD_p - 0.0000V_{pd}^2 + 0.2824V_{pd}D_p + 0.0034D_p^2) \times 10^3 \end{aligned} \quad (11)$$

TABLE V. COMPARISON OF FOUR BINOMIAL REGRESSION EQUATIONS

	linear	pure quadratic	interaction	quadratic
Complexity	6	11	16	21
RMSE	0.8080	0.6223	0.6262	0.5027

The results in Table 5 show consistency with those in Table 3. The linear equation's accuracy was the worst. And full quadratic equation was so complex that it not suitable for modeling. Besides, the errors of pure quadratic and interaction equations were similar. But the pure quadratic equation was simpler than interaction equation.

Otherwise substituting the predicted data of cucumber into the Eq. (8-11), the ordinate is CO2 exchange in terms of time. The comparison of predicted results of four regression equations was shown in Figure 2.

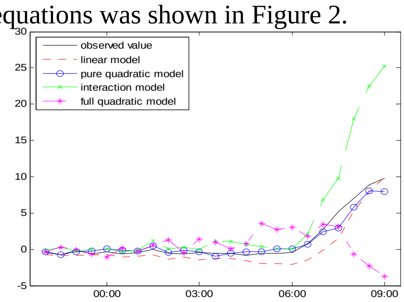

FIGURE II. MULTIVARIATE REGRESSION MODELS.

The results shown in Figure 2 indicated that the error of pure quadratic model was obviously the smallest. And the next best model was linear model. In addition, the accuracies of interaction and full quadratic models were worse, compared with the linear and pure quadratic model. The error of interaction model increased rapidly after 6 o'clock. And there was the fluctuation in the curve of full quadratic model after 3 o'clock.

IV. CONCLUSIONS

The paper took tomato, pumpkin and cucumber as examples to study the relation between CO2 exchange and its affecting environmental characters. According to the analyzed results of three plants, path analysis expresses comprehensively the inner influence between variables, because it shows not only the correlation between dependent and independent variables, but also the direct and indirect effects among variables.

In the analysis process of regression model, although linear model had the simplest structure, the error of that was the largest. So the linear model was not suitable for crop model which demands high accuracy. On the contrary, the quadratic model had the highest accuracy and the most complicated structure. For example, if the number of input variables is five, there will be twenty-one coefficients that need to be computed. From the comparison between four multivariate binomial regression models, the pure quadratic model had advantages over the others in simplicity and curacy. All in all, through the comparison of common statistical analysis and regression model, we verified the better performance of path analysis and pure quadratic model. The paper could offer reference for crop modeling with small sample size [16], by the comparison of methods of multivariate analysis and regression.

ACKNOWLEDGEMENTS

In this paper, the research was sponsored by Key Program of Nature Science Foundation of Educational Commission of Anhui Province (KJ2012A103, KJ2012A102), and Introduction of International Advanced Agricultural Science and Technology Projects (948) (2013-Z64). Besides, the authors wish to thank Prof. Jun Zhu and Gaojian Xu, for the information they provided kindly.

REFERENCES

[1] Xie Zhujie, Cao Weixing, Luo Weihong. Application and prospects of crop growth simulation models in precision agriculture and intellectualized greenhouse in Shanghai. Acta Agriculturae Shanghai, 2001, 17(2): 17-21.

[2] He Dongjian, He Yong, Li Mingzan, Hong Tiansheng, Wang Chenghong, Song Su, et al. Research progress of information science-related problems in precision agriculture. Bulletin of National natural Science Foundation of China, 2011, 1: 10-16.

[3] Nakazato T, Bogonovich M, Moyle L C. Environmental factors predict adaptive phenotypic differentiation within and between two wild Andean tomatoes. Evolution, 2008, 62(4): 774-792. doi: 10.1111/j.1558-5646.2008.00332.x.

[4] Gallo A, Moschini M, Cerioli C, F Masoero. Use of principal component analysis to classify forages and predict their calculated energy content. Animal, 2013: 7(6): 930-939. doi: 10.1017/S1751731112002467.

[5] Tiwari J K, Upadhyay D. Correlation and path-coefficient studies in tomato (Lycopersicon esculentum Mill.). Research Journal of Agricultural Sciences, 2011, 2(1): 63-68.

[6] Li Yuyu. Application of BP Neural Network for Forecasting Parameters Used in Laser Working Computer Simulation. Journal of Guangdong University of Technology, 2000, 17(3): 26-30.

[7] Luo Yi, Guo Wei. Development and problems of crop models. Transactions of the CSAE, 2008, 24(5): 307-312.

[8] Phyto-Sensor Group. http://www.phyto-sensor.com/

[9] BF Agritech - IT Systems. http://bf-ag.co.il/ITSystems.asp

[10] Jiang Zhaohui, Wang Chunsheng, Zhang Jing, Yue Yi, Li Shaowen. Online Monitoring and Analysis of Plant Photosynthetic Physiology and Environmental Factors. Applied Mechanics and Materials,2013,241:7580.doi:10.4028/www.scientific.net/AMM.241-244. 75.

[11] Aminrad Z, Zakariya S Z B S, Hadi A S, Sakari M. Relationship Between Awareness, Knowledge and Attitudes Towards Environmental Education Among Secondary School Students in Malaysia. World Applied Sciences Journal, 2013, 22(9): 1326-1333.

[12] Yao Yongzhe，Li Jianming，Zhang Rong, Sun Sanjie, Chen Kaili. Greenhouse tomato transpiration and its affecting factors: Correlation analysis and model simulation. Chinese Journal of Applied Ecology, 2012, 23(7): 1869-1874.

[13] Moghadam F M, Ahmadi A, Keynia F. A New Iris Detection Method based on Cascaded Neural Network. Journal of Computer Sciences and Applications, 2013, 1(5): 80-84.

[14] Wang Xiaosen, Liu Zugui, Liu Hao, Yang Shenjiao, Zhang Xiaopei, Meng Zhaojiang. Path Analysis and Numerical Simulation of MDS of Tomato Stem Diameter. Transactions of the Chinese Society for Agricultural Machinery, 2012, 43(008): 187-192. doi: 10.6041/j.issn. 1000-1298. 2012. 08. 034.

[15] Jami M S, Husain I A F, Kabashi N A, Abdullah N. Multiple Inputs Artificial Neural Network Model For The Prediction Of Wastewater Treatment Plant Performance. Australian Journal of Basic and Applied Sciences, 2012, 6(1): 62-69.

[16] Zhang J, Jiang Z, Wang C, et al. Modeling and prediction of CO2 exchange response to environment for small sample size in cucumber [J]. Computers and Electronics in Agriculture, 2014, 108: 39-45.