

Grain Ration Consumption Forecasting based on Multivariate Regression Model Combined with Gliding Data Barycenter

Chunhua Zhu^{*} and Jiaojiao Wang

College of Information Science and Engineering, Henan University of Technology, Zhengzhou, 450001, China

Abstract

To address the existing and limited original data and lower prediction robustness, a new multivariate regression forecasting model combined with gliding data barycenter was proposed. In this new forecasting method, the original data was interpolated and the corresponding data barycenter was optimized. Then, the important impact factors of ration consumption were analysed and chosen for the multivariate regression model. In simulation experiments, the training data of 35 years (1981-2015) were used, and the results have shown that the proposed method can greatly improve prediction accuracy and robustness.

Keywords: grain ration consumption; interpolation; gliding data barycenter; correlation analysis; multivariate regression model

(Submitted on May 9, 2018; Revised on June 16, 2018; Accepted on July 19, 2018)

© 2018 Totem Publisher, Inc. All rights reserved.

1. Introduction

Ration consumption is an important part of grain consumption. Data analysis has shown that ration consumptions are decreasing year by year with the acceleration of urbanization level construction process, which differs between urban and rural areas. Recently, a variety of grain consumption prediction methods have been presented. Zhiqiang Li et al. proposed an EMM model [1], which focuses on the effects of socioeconomic factors and biological energy on grain consumption; Wei Jia et al. used the time extrapolation sequence to predict per capita grain consumption [2], Qiyu Luo et al. used time series to determine important parameters and analyse panel data for estimating per capita consumption equation [3], focusing on verifying the internal relations between grain consumption and time series; Yao Pan et al. adopted a new method by combining statistical analysis and quantitative analysis [4], which established a logarithmic regression model to analyse the original data and objective phenomena. Reviewing the existing methods [5], there are a series of problems in predicting ration consumption; for example, abnormal historical data has a great destructive effect on forecasting results, the small number of original data leads to large errors in prediction model training, and the independent variable is single. In actual economic issues, the change of a dependent variable is often influenced by several important impact factors; therefore, it is necessary to use two or more major impact factors as independent variables to explain the change of dependent variables, that is, multiple regression. When the multiple input independent variables are linear with the output variable, the regression analysis is expressed as one multiple linear equation [6]. This method has been used for short-term grain prediction, but this has demonstrated poorer prediction performance when there is limited training data and non-optimal impact factors. Practically, it has been proven that the historical data of ration consumption in urban and rural areas are influenced by many factors, such as urban/rural population, urbanization level, urban/rural Engel coefficient, agricultural production price index, per capita income of urban residents. These factors will change dynamically with time, along with the corresponding relational degree between these factors and the ration consumption. Therefore, pre-processing the original sample data before the prediction will be important for prediction accuracy, including increasing the amount of training data and selecting key impact factors. In this paper, a new method for pre-processing sample data is proposed, in which the original sample data is interpolated, and its gliding data barycenter of the above data are used to train the predictive model. This presented processing method can remove the abnormal points of the historical data and keep the amount of training data after the gliding data barycenter. Based on the above-processed training data, the key impact factors are selected through

^{*} Corresponding author.

E-mail address: zhuchunhua@haut.edu.cn

relational analysis, and the multivariate regression dynamic prediction models of urban and rural ration consumption are established respectively. Finally, the prediction results can be obtained using the inverse pre-processing operation.

2. Multivariate Prediction Model based on Gliding Data Barycenter

Compared to the existing multivariate prediction model, the data pre-processing unit and the group gliding barycenter calculation unit are added to the multivariate prediction model based on the interpolate gliding data barycenter. The historical data will be pre-processed first and then inputted into the multivariate linear predictive model. The prediction principle is shown in Figure 1.

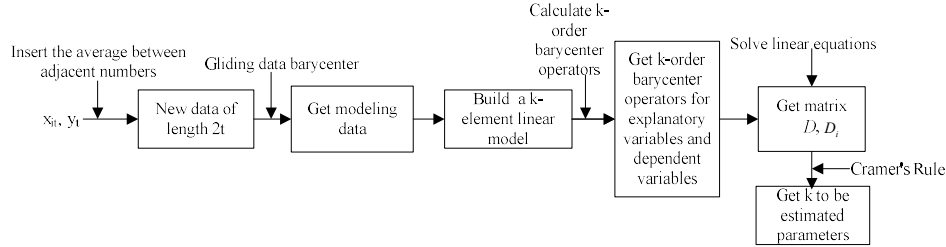


Figure 1. The prediction principle of the multivariate model based on the interpolate gliding data barycenter

2.1. Data Pre-Processing

The historical data of ration consumption spans about 35 years, from 1981 to 2015. The lower the amount of training historical data, the larger the fitting error in the training prediction model. Therefore, this paper adopts linear interpolation to increase the amount of historical data. Assuming the original data is $\{x_i\}$, $(i = 1, 2, \dots, n)$, the interpolated sequence $\{x'_i\}$, $(i = 1, 2, \dots, 2n - 1)$ can be expressed as

$$\begin{cases} x'_{2i} = \frac{x_i + x_{i+1}}{2} \\ x'_{2i-1} = x_i \\ x'_{2n-1} = x_n \end{cases} \quad i = 1, 2, \dots, n-1 \quad (1)$$

From Equation (1), the interpolated data $\{x'_i\}$ will be used as training data for the prediction model to calculate the gliding data barycenter and predict model parameters.

The gliding data barycenter method is one new data processing method that can smooth the abnormal points in the original data, thereby improving the fitting and forecasting results [2]. In this new processing method, the original data will be recursively grouped, and the first-order data barycenter of each group is obtained, which will become the new training data. In each group, the Euclidean distance from the data barycenter to each point is the shortest, which will ensure the removal of the abnormal point in the original data.

For the interpolated data $\{x'_i\}$, $(i = 1, 2, \dots, 2n - 1)$, when the length of the group is equal to 2, the recursive group data is $\{x''_i\}$, which can be expressed as

$$x''_i = \frac{x'_i + x'_{i+1}}{2}, \quad i = 1, 2, \dots, 2n - 2 \quad (2)$$

If the length of the group is equal to 3, the recursive group data can be expressed as

$$x''_i = \frac{x'_i + x'_{i+1} + x'_{i+2}}{3}, \quad i = 1, 2, \dots, 2n - 3 \quad (3)$$

Practically, the length of the group can be decided through stability analysis of the training data.

2.2. *K-order Barycenter Operators*

From Figure 1, the k -order barycenter operators are used to compute the prediction parameters of the multivariate model. Assuming the coordinates of n original data points are (x_i, y_i) , $(i = 1, 2, \dots, n)$, its barycenter can be expressed as $(x^{(n)}, y^{(n)})$ [7]

$$\begin{aligned} x^{(n)} &= \frac{x^{(n)} + (n-1)x^{(n-1)}}{n} = \frac{1}{n} \sum_{i=1}^n x_i \\ y^{(n)} &= \frac{y^{(n)} + (n-1)y^{(n-1)}}{n} = \frac{1}{n} \sum_{i=1}^n y_i \end{aligned} \quad (4)$$

Here, the barycenter of the n data points is unique (in [7], Theorem 1). Let n be equal to 1, 2, 3, \dots in Equation (4), then the corresponding data set (only expressing abscissa, below is the same) can be written as

$$x_1, \frac{x_1 + x_2}{2}, \frac{x_1 + x_2 + x_3}{3}, \dots, \frac{x_1 + x_2 + \dots + x_n}{n} \quad (5)$$

Its k -order barycenter operator is

$$\frac{\sum_{i=1}^n {}^{(k)}x_i}{(k-1)! \sum_{i=1}^n (n-i+1)(n-i+2) \cdots (n-i+k-1)} \quad (6)$$

Here,

$$\sum_{i=1}^n {}^{(k)}x_i = \frac{1}{(k-1)!} \sum_{i=1}^n (n-i+1)(n-i+2) \cdots (n-i+k-1)x_i, \quad k \geq 2$$

The order k of the barycenter operator is equal to the number of the prediction model parameters.

2.3. *Model Parameters Calculation*

The linear prediction model can be expressed as

$$y_i = X_i' \beta + e_i, \quad i = 1, 2, \dots, n \quad (7)$$

Here, $\beta = (\beta_0, \beta_1, \dots, \beta_{p-1})$, $X_i' = (1, x_{i1}, \dots, x_{i(p-1)})$, $(i = 1, 2, \dots, n)$, e_i is a random error term. p is the number of the independent variables in Equation (7), and n is the number of the historical data. Unlike the common least-square method, the parameters of the multivariate prediction model based on gliding data barycenter will be computed using k -order barycenter operators.

For Equation (7), the k -order barycenter operator on both sides is

$$\sum_{i=1}^{p-1} {}^{(k)}y_i = \beta_0 \sum_{i=1}^{p-1} {}^{(k)}1 + \beta_1 \sum_{i=1}^{p-1} {}^{(k)}x_{i1} + \beta_2 \sum_{i=1}^{p-1} {}^{(k)}x_{i2} + \dots + \beta_{p-1} \sum_{i=1}^{p-1} {}^{(k)}x_{i(p-1)} + e_i \sum_{i=1}^{p-1} {}^{(k)}1 \quad (8)$$

Let $k = 1, 2, \dots, n$ orderly in Equation (8). We can obtain one linear equation set. The corresponding parameters can be solved by Cramer's Rule

$$\hat{\beta}_0 = \frac{D_1}{D}, \hat{\beta}_1 = \frac{D_2}{D}, \dots, \hat{\beta}_{p-1} = \frac{D_p}{D} \quad (9)$$

Here,

$$D = \begin{vmatrix} \sum_{i=1}^{p-1} {}^{(1)}1 & \sum_{i=1}^{p-1} {}^{(1)}x_{i1} & \dots & \sum_{i=1}^{p-1} {}^{(1)}x_{i(p-1)} \\ \sum_{i=1}^{p-1} {}^{(2)}1 & \sum_{i=1}^{p-1} {}^{(2)}x_{i1} & \dots & \sum_{i=1}^{p-1} {}^{(2)}x_{i(p-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{p-1} {}^{(n)}1 & \sum_{i=1}^{p-1} {}^{(n)}x_{i1} & \dots & \sum_{i=1}^{p-1} {}^{(n)}x_{i(p-1)} \end{vmatrix} \quad (10)$$

Replacing the j^{th} column of D with $\sum_{i=1}^{p-1} {}^{(1)}y_i, \sum_{i=1}^{p-1} {}^{(2)}y_i, \dots, \sum_{i=1}^{p-1} {}^{(n)}y_i$, that is, the left term in Equation (8), we can get the coefficients D_j , ($j = 1, 2, \dots, p$). Furthermore, by substituting Equation (9) into Equation (7), the optimal predictive model parameters can be obtained after iterative iteration.

2.4. Selecting the Key Impacting Factors

Grey relational degree [8-9] is used to analyse the dependencies between one main behaviour factor and relevant behaviour factors. The higher correlation degree indicates that these factors affect the main behavioural factor to a greater degree; therefore, tracing the changes of the key factors and building the corresponding multivariate prediction model will increase the prediction accuracy of the main behavioural factor.

The predicted main behaviour factor is used as a reference series, and the relevant behaviour factors are comparative series. Assuming the reference series is $x_0(k) = \{x_0(1), x_0(2), \dots, x_0(n)\}$, the comparative series are

$$\begin{aligned} x_1(k) &= \{x_1(1), x_1(2), \dots, x_1(n)\} \\ x_2(k) &= \{x_2(1), x_2(2), \dots, x_2(n)\} \\ &\dots\dots\dots \\ x_m(k) &= \{x_m(1), x_m(2), \dots, x_m(n)\} \end{aligned}$$

For the above m factors, each has a different measurement unit in dimension and order of magnitude. Therefore, before calculating the correlation degree, it is usually necessary to do dimensionless processing of the original data [10], which includes initializing, averaging, normalizing, etc. This paper uses the normalization method, and the dimensionless reference series and comparative series are $y_0(k)$ and $y_i(k)$, ($i = 1, 2, \dots, m; k = 1, 2, \dots, n$) respectively. The i^{th} absolute difference sequence is $\Delta_{0i}(k)$:

$$\begin{aligned} \Delta_{01}(k) &= |y_0(k) - y_1(k)| = \{\Delta_1(1), \Delta_1(2), \dots, \Delta_1(n)\} \\ \Delta_{02}(k) &= |y_0(k) - y_2(k)| = \{\Delta_2(1), \Delta_2(2), \dots, \Delta_2(n)\} \\ &\dots\dots\dots \\ \Delta_{0m}(k) &= |y_0(k) - y_m(k)| = \{\Delta_m(1), \Delta_m(2), \dots, \Delta_m(n)\} \end{aligned}$$

Finding the maximum absolute difference Δ_{\max} and the minimum absolute difference Δ_{\min} from the absolute difference sequence, the i^{th} gray correlation coefficient is

$$L_{0i}(k) = (\rho\Delta_{\max} + \Delta_{\min}) / (\Delta_{0i}(k) + \rho\Delta_{\max}) \quad (11)$$

Where ρ is the resolution coefficient, which is generally between 0.5 to 1; the smaller the parameter ρ , the stronger the distinguishing ability for the difference among the correlation coefficients, usually taking 0.5.

Averaging the correlation coefficient, the relational degree is obtained as

$$R_{0i}(k) = \frac{1}{n} \sum_{k=1}^n L_{0i}(k) = \frac{1}{n} \{L_{0i}(1) + L_{0i}(2) + \cdots + L_{0i}(n)\} \quad (12)$$

Using Equation (12), the key factors can be chosen. In this paper, rural/urban ration consumption is the main behaviour factor, and the relevant behaviour factors are the population size, the level of urbanization, the Engel coefficients, the agricultural production price index [11], and so on.

3. Simulation Analysis

The data of China's grain consumption and its impacting factors comes from the "Statistical Yearbook of China" (1978-2015) [12]. To verify the predictive performance of the proposed multivariate regression model in this paper, the prediction results are compared to the common multivariate regression prediction. In simulation, defining the year of 1981 as $t = 1$, and thus 2015 as $t = 35$ orderly, urban and rural ration consumption are respectively expressed as $y_0(t)$ and $y_1(t)$; the four factors are respectively defined as $x_1(t)$, $x_2(t)$, $x_3(t)$, and $x_4(t)$. According to the prediction principle in Figure 1, the forecasting process of rural ration consumption is shown in the following:

(1) When $t=1 \sim n$, calculate the degree of correlation and correlation order respectively between $y_1(t)$ and $x_1(t)$, $x_2(t)$, $x_3(t)$, and $x_4(t)$. Two of the most relevant factors are selected as the key impacting factors, expressed as $x'_1(t)$ and $x'_2(t)$.

(2) The rural consumption of ration $y_1(t)$ and its key impact factors $x'_1(t)$ and $x'_2(t)$ were interpolated. The interpolation formula is shown in formula (1). The data obtained are respectively expressed as $y_{1c}(t')$, and its key impact factors are expressed as $x'_{1c}(t')$ and $x'_{2c}(t')$, $t'=1 \sim 2n-1$.

(3) Setting the group length as 2 and according to Equation (2), after interpolating the pre-processing data $y_{1c}(t')$, $x'_{1c}(t')$, and $x'_{2c}(t')$, the data in the group gliding barycenter is expressed as $y_{1cz}(t'')$, $x'_{1cz}(t'')$, and $x'_{2cz}(t'')$, $t''=1 \sim 2n-2$.

(4) $y_{1cz}(t'')$, $x'_{1cz}(t'')$, and $x'_{2cz}(t'')$ as training data will be inputted into the multivariate regression prediction model; then, model predictive parameters are computed. The predictive equation is

$$y'_{1cz}(t'') = \alpha_0 + \alpha_1 x'_{1cz}(t'') + \alpha_2 x'_{2cz}(t'')$$

Here, $y'_{1cz}(t'')$ is the prediction result of rural ration consumption, and its key impact factors are $x'_{1cz}(t'')$ and $x'_{2cz}(t'')$.

(5) Through the inverse pre-processing operation on the prediction result $y'_{1cz}(t'')$ using (3), obtain the actual prediction value of the rural ration consumption $y'_1(t)$.

Similar to the forecasting process of rural ration consumption, the actual urban ration consumption can be obtained as $y'_0(t)$.

3.1. Selection of Key Impact Factors

We selected the urban/rural ration consumption, urban/rural population, urbanization level, urban/rural Engel coefficients, and agriculture production price index from 1981 to 2015 in the "Statistical Yearbook of China" [12] as the original sample data. Using Equations (11) and (12), the correlation coefficients of rural/urban ration consumption are shown in Figure 2, and the corresponding relational degree is shown in Table 1.

From Table 1, according to the relational degree and correlation order of urban/rural ration consumption, the key impacting factors of urban ration consumption are urban population and urbanization level, and the key factors affecting the rural ration consumption are the rural Engel coefficient and rural population.

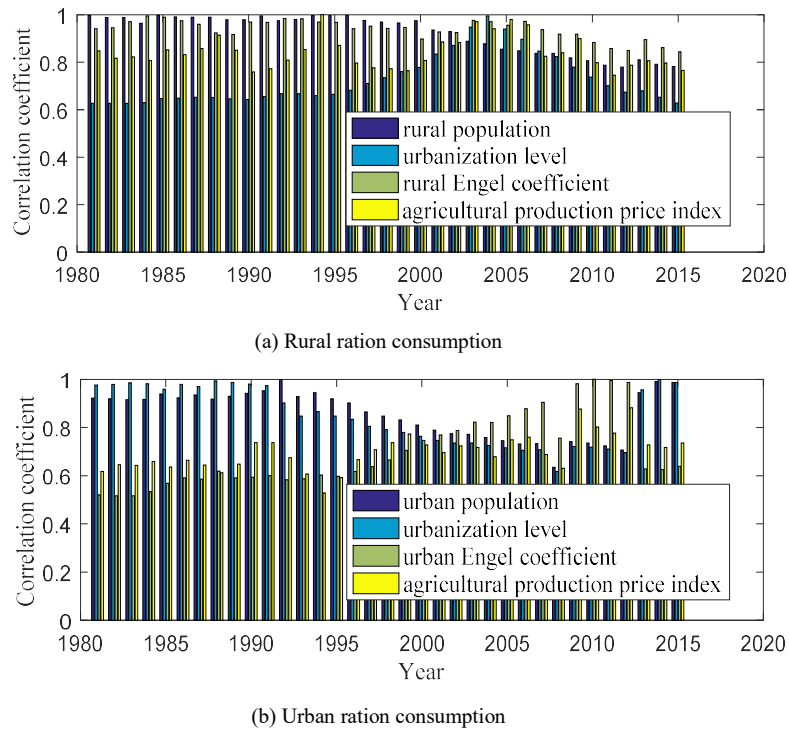


Figure 2. (a) Correlation coefficients of rural ration consumption and its impact factors; (b) Correlation coefficients of urban ration consumption and its impact factors

Table 1. The relational degree and correlation order between the urban/rural ration consumption and its impact factors in 1981-2015

Impact factor	Urban		Rural	
	Relational degree	Correlation order	Relational degree	Correlation order
(Urban/Rural) population	0.8580	1	0.92220	2
Urbanization level	0.8481	2	0.72540	4
(Urban/Rural) Engel coefficient	0.6978	3	0.93810	1
Produce price index of agricultural products	0.6977	4	0.84204	3

3.2. Prediction Performance Simulation

The historical data of rural ration consumption, urban ration consumption, urbanization level, agricultural production price index, rural population and rural Engel coefficient from 1981 to 2015 are divided into three sample intervals: from 1981 to 1995, from 1981 to 2005, and from 1981 to 2015. They were respectively inputted into the multivariate linear regression (MLR) model, the MLR model based on gliding data barycenter (GDB-based MLR), and the interpolated GDB-based MLR (IGDB-based MLR) model for model training. The predicted value was compared to the actual value to calculate the prediction error respectively, as shown in Figure 3.

The fitting test between the prediction data and the original data is carried out, and the average error and the Theil coefficient are adopted. The Theil coefficient is a constant between 0 and 1; the closer to zero, the better the fitting effect. The test results are shown in Table 2.

It can be seen from Table 2, for the original data of the same sample interval, that the fitting error of the IGDB-based MLR model is the smallest. This is because the gliding barycenter of historical data is used to train the prediction model, which can smooth the training data. In addition, the training data is increased by interpolation pre-processing; therefore, the fitting accuracy can be doubled compared to the GDB-based MLR model.

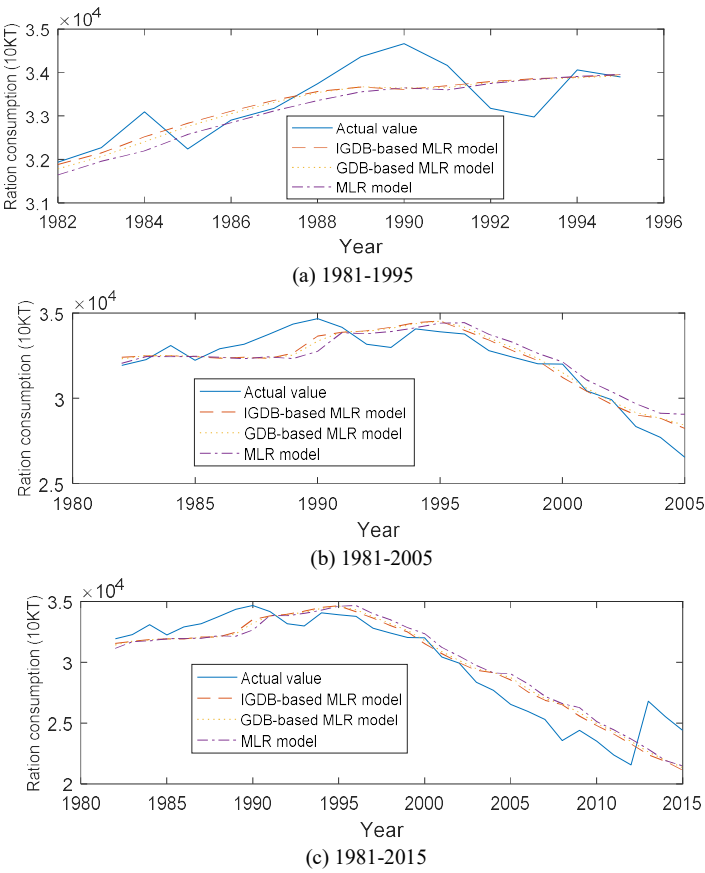


Figure 3. (a) Comparison of consumption forecasting models in 1981-1995; (b) Comparison of consumption forecasting models in 1981-2005; (c) Comparison of consumption forecasting models in 1981-2015

Table 2. The fitting test between the prediction data and the original data

Year Range Model	1981-1995		1981-2005		1981-2015	
	Theil coefficient	Average error	Theil coefficient	Average error	Theil coefficient	Average error
MLR model	0.0084	0.56%	0.0161	0.77%	0.0290	0.99%
GDB-based MLR model	0.0080	0.27%	0.0133	0.42%	0.0275	0.60%
IGDB-based MLR model	0.0078	0.15%	0.0127	0.27%	0.0270	0.31%

4. Conclusions

This paper proposed a new prediction method for ration consumption. The use of interpolation and gliding data barycenter pre-processing can ensure the amount of training data and remove singular points, thereby significantly improving the prediction accuracy of the multivariate regression model. The simulation results in different sample intervals have proven the above conclusions. Therefore, the proposed prediction method is adaptable for less training data, and it can also be used in other areas of data prediction.

Acknowledgements

This research was financially supported by the National Science Foundation of China (61741107), the National Food Industry Commonweal Special Scientific Research Projects (201413001), and the Henan Provincial Department of Science and Technology Project (172102210230).

References

1. Z. Q. Li, J. Z. Wu, and D. J. Wang , “Change Analysis and Demand Forecast of Grain Consumption in China,” *Food and Nutrition in China*, Vol. 18, No. 3, pp. 38-42, 2012
2. W. Jia and F. Qin, “China’s Grain Demand Forecast,” *Food and Nutrition in China*, Vol. 19, No. 1, pp. 40-44, 2013
3. Q. Y. Luo, J. Mi, and M. J. Gao, “Research on Forecasting for Long-term Grain Consumption Demands in China,” *China’s Agricultural Resources and Zoning*, Vol. 35, No. 5, pp. 1-6, 2014
4. Y. Pan and L. Z. Liu, “Analysis and Prediction of Food Consumption of Rural Residents in China,” *Population and Rconomy*, No. 3, pp. 1-8, 2005
5. J. Mi, Q. Y. Luo, and M. J. Gao, “Review on the Method of Forecasting Grain Demand,” *China Agricultural Resources and Regional Planning*, Vol. 34, No. 3, pp. 28-33, June 2013
6. Y. S. Zhou, Y. H. Xiao, and R. S. Huang, “Prediction of Grain Yield in Guangxi based on Multivariate Linear Regression,” *Journal of Southern Agriculture*, Vol. 42, No. 9, pp. 1165-1167, 2011
7. J. L. Zhang, “Research on Gliding Data Barycenter Forecasting Method and its Application,” *Mathematical Statistics and Management*, Vol. 29, No. 6, pp. 1036-1041, November 2010
8. H. Y. Chen, J. B. Zhao, and C. L. Liu, “Properties of Combination Forecasting Model Based on Gray Incidence,” *Journal of Southeast University (Natural Science Edition)*, Vol. 34, No. 1, pp. 23-27, January 2004
9. Q. F. Li, G. L. Kang, and X. F. Li, “Factors Influencing Grain Production in Henan Province based on Gray Correlation,” *Asia Agricultural Reaearch*, Vol. 1, No. 5, pp. 23-27, May 2009
10. X. M. Xu, W. Erika, and M. B. Angela, “Management of Raspberry and Strawberry Grey Mould in Open Field and Under Protection,” *Agronomy for Sustainable Development*, Vol. 32, pp. 531-543, 2012
11. S. M. Wang, “The Main Factors Affecting Grain Consumption,” *Contemporary Economy*, No. 22, pp. 40-41, 2015
12. National Bureau of Statistics, “Statistical Yearbook 2015 of China,” *China Statistics Press*, 2016

Chunhua Zhu borns in 1976, doctor, master’s tutor. Her research interests include broadband wireless communication and grain yield prediction.

Jiaojiao Wang is a master student at Henan University of Technology. Her research interests include signal and information processing.