Modeling potato yield statistics by using information technology for sustainable development

A P Darmanyan and S I Bogdanov
Volgograd State Agrarian University, 26 University Avenue, 400002, Volgograd, Russia
E-mail: adarma@inbox.ru

Abstract. The potato yield statistics analysis and modeling in the Russian Federation according to the Federal State Statistics Service data for the period 2000-2020 was carried out. Based on the analysis of the autocorrelation function (Acf) as a mathematical model for modeling potato yield, the choice of a linear autoregressive model of the first order is justified. By using the multiple regression and correlation analysis methods, the parameters of the model were found, their statistical significance was proved, and the potato yield calculations proved the found model correspondence to real data for the period of 2002-2020.

1. Introduction
The successful application of digital technologies in agriculture is possible only on the basis of the processes and phenomena mathematical models that are formed by using modern IT technologies and software products. An important place among such technologies is occupied by statistical methods, which make it possible to create processes adequate mathematical models, which help to perform scientifically, based forecasting for future periods of time.

However, the use of statistical methods in scientific research, in our opinion, currently in the Russian Federation does not meet the needs of the digital economy. Thus, the study carried out by the Institute for Statistical Studies and Economics of Knowledge of the National Research University Higher School of Economics in 2019 showed that "over a third of Russian scientists do not own the tools of statistical analysis" [1]. Similar results were obtained by the authors earlier [2], when, based on the analysis of scientific publications in agriculture, it was revealed that out of 52 statistical methods used for the analysis of experimental data and mathematical modeling, the scientific studies’ authors most often use only descriptive statistics (19%), regression and correlation analysis (2.8%), methods of planning experiments (1.4%). Many other effective statistical methods are not used at all [3, 4].

To analyze the statistics of agricultural production yields, time series analysis methods are often used with the identification of linear trends, as, for example, in the research of grain crops productivity [5]. However, such methods, as a rule, lead to large approximation errors, which do not allow forecasting for future periods of time with the accuracy acceptable for practice.

At the same time, such effective statistical models as autoregressive models of various orders are known, which help to simulate time series with low volatility, i.e. small changes in the series values [6]. Autoregressive models are successfully used in econometrics [7], but they are not used at all in agricultural productivity statistics studies, including potato productivity statistics studies.
The aim of this study was the potato yield statistics mathematical modeling in the Russian Federation by using the autoregressive model (ARM).

2. Materials and methods
To perform mathematical modeling, the Federal State Statistics Service statistics on potato yield in the Russian Federation for 1990-2020 was used [8], which is shown in Figure 1 for the period of 2002-2020.

As seen in Figure 1, potato yield is a volatile process with a clear positive trend. To simulate such processes, the use of mathematical models with a linear trend allows only to reveal the model of the trend itself, but, as noted above, forecasting by using such a model for future periods of time leads to large errors.

If we do not consider the years with obvious natural anomalies (for example, 2012, point 9 in Figure 1), one of the agricultural crops yield obvious feature is that the value of this indicator in the current year is influenced by the same crop yield in previous years. Therefore, at the research first stage, it is necessary to analyze the presence (or absence) of the relationship between the yield value of the current year and the yield values in previous years (time lags). As it is known [6], this can be done based on the first autocorrelation function (Acf) calculation and analysis.

For the potato productivity statistics in the Russian Federation for the period of 1990-2020 in the MS Excel environment, the authors calculated the Acf, which is shown in the form of the correlogram in Figure 2. As can be seen from Figure 2, the correlation coefficients values on the lags 1 and 2 (1 lag = 1 year) significantly exceed the level of "white noise" shown in Figure 2 by the dashed line. This indicates a strong relationship between current potato yields and potato yields over the previous two years. Therefore, for modeling the statistics of such statistics, it is advisable to choose mathematical models that take into account this dependence. Since autoregressive models of various orders have such properties, this type of model was chosen for potato yield statistics mathematical modeling in the Russian Federation.
3. Results and discussion

A linear AR model was chosen in the following form [6] to simulate the potato yield:

\[ Y_t = a_0 + \sum_{i=1}^{n} a_i \cdot Y_{t-i}, \]  

where \( t \) is the number of the year; \( n \) is the order of the model; \( a_0 \) and \( a_i \) are model parameters.

According to the conclusions made earlier as a result of calculating the autocorrelation function (Acf), the second order or \( n = 2 \) was chosen for the APM model (1). To find the model parameters \( a_0 \), \( a_1 \) and \( a_2 \), multiple regression analysis was performed further. As the result, the multiple correlation coefficient value turned out to be equal to \( r = 0.931 \), which indicates that the current year potatoes yield is strongly influenced by both the yield of the last year \((t - 1)\) and the yield that was two years ago \((t - 2)\). This fact confirms the conclusions that were made earlier, based on the autocorrelation function calculation (Figure 1). The values obtained as the result of multiple regression analysis are presented in Table 1 to assess the statistical significance of the ARM model parameters.

**Table 1. Calculation results of the second order ARM model parameters.**

| Parameter | Value | Standard error | t-test | p-value |
|-----------|-------|----------------|--------|---------|
| \( a_0 \) | 10.657 | 5.103          | 2.71   | 0.053   |
| \( a_1 \) | 0.686  | 0.266          | 2.58   | 0.023   |
| \( a_2 \) | 0.275  | 0.248          | 1.11   | 0.287   |

From the data in Table 1, it follows that in the two-parameter ARM model, only two coefficients \( a_0 \) and \( a_1 \) are statistically significant, which is confirmed at the 95% significance level of p-values, which are less than 0.05. At the same time, the parameter \( a_2 \) turned out to be statistically insignificant, since the p-value for it is greater than 0.05. Therefore, this parameter is excluded from the model. Thus, as the multiple regression analysis result, the two-parameter ARM model was reduced to a one-parameter model with the parameters \( a_0 \) and \( a_1 \).

Further, for the ARM model, a pairwise regression analysis was performed, which showed that the pairwise correlation coefficient is \( r = 0.924 \). This indicates that 82% of changes in the current year yield depend on the changes in the previous year yield (coefficient of determination \( r^2 = 0.82 \)), and 18% is on the influence of other factors and random influences. Since the previous year yield influence on the current yield value is quite large, and then the ARM model parameters statistical significance was assessed, which were calculated by using the paired regression analysis methods and are presented in Table 2.
Table 2. The results of calculating the first order ARM model parameters.

| Parameter | Value  | Standard error | t-test | p-value |
|-----------|--------|----------------|--------|---------|
| $a_0$     | 10.076 | 5.217          | 3.71   | 0.051   |
| $a_1$     | 0.956  | 0.106          | 9.06   | 0.000   |

From the data in Table 2, it can be seen that the parameters of the ARM model $a_0$ and $a_1$ are statistically significant, which is confirmed by the calculated p-values, which at the 95% significance level are less than 0.05. Finally, the first-order autoregressive model for potato yield statistics in the Russian Federation is as follows:

$$Y_t = 10.076 + 0.956Y_{t-1}, \text{ (centners per hectare)} \quad (2)$$

The potatoes yield calculating results according to the model (2) for the period of 2011-2019 are shown in Table 3. In Figure 1 they are shown in comparison with real values for the period of 2002-2020, which shows a good correlation. The error calculation between the calculated potato yield according to the ARM model and the real values for the period of 2002-2020 averaged 3%, which is acceptable accuracy for practice and forecasting.

Table 3. Comparison of potato yield (centners per hectare) in the Russian Federation according to the Federal State Statistics Service data [8] with potato yield calculated by the model (2).

| Years | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 |
|-------|------|------|------|------|------|------|------|------|------|
| Federal State Statistics Service | 150  | 136  | 147  | 153  | 164  | 158  | 163  | 170  | 178  |
| ARM (2) | 151  | 153  | 140  | 150  | 156  | 167  | 161  | 165  | 173  |

4. Conclusion

The potato yield statistics analysis and mathematical modeling in the Russian Federation according to the Federal State Statistics Service data for the period of 2000-2020 was carried out. Based on the calculation and analysis of the first autocorrelation function (Acf) in MS Excel as a mathematical model for modeling potato yield statistics, the choice of the second order linear ARM model was justified. By using the correlation analysis methods, the multiple correlation coefficient value was found and a close relationship between the current year potato yield and the previous two years potato yield was proved, and thus the choice of a two-parameter ARM model was substantiated. By using the multiple regression analysis methods, the values of ARM model two parameters were obtained, their statistical significance was assessed and it was shown that the second parameter of the model is statistically insignificant and therefore it is excluded from the model, and the two-parameter ARM model became a linear paired autoregressive model. By using the paired correlation and regression analysis methods, the ARM model parameters were found and their statistical significance was proved.

Thus, the potato yield statistics mathematical model in the Russian Federation for the period of 2002-2020 can be described by a first-order linear autoregressive model of the following form: $Y_t = 10.076 + 0.956 \times Y_{t-1}$ (centners per hectare). The errors in calculating potato yield according to this model for the period of 2002-2020 averaged 3%, which is acceptable accuracy for practice and forecasting.

References

[1] Volkova G L and Shmatko N A 2020 Express-Information Science, Technology, Innovation of the Institute of Statistical Research and Knowledge Economics of the National Research University "Higher School of Economics" 154 1-4
[2] Darmanyan A P and Veselova N M 2018 Scientific significance of statistical methods for analysing experimental data in the system of postgraduate agricultural education *Current Issues of Professional Education* 3(12) 33-37

[3] Rogachev A F et al 2020 A set of data on retrospective grain yield for neural network modeling *IOP Conf. Ser.: Earth Environ. Sci.* 577 012006

[4] Ryabtsev V G and Utkina T Yu 2016 Information technologies for designing automated process control systems *Control Systems, Communications and Security* 1 207-239

[5] Kruzhilin I P, Ovchinnikov A S and Melikhov V V 2014 Proc. of the Nizhnevolskzhsk Agrouniversity Complex: Science and Higher Professional Education 4(36) 8-21

[6] Ayvazyan S A 2001 *Applied Statistics. Fundamentals of Econometrics* (Moscow: Unity-Dana) vol 2 pp 234-430

[7] Darmanyan A P and Filippov M V 2013 Forecasting the cost of housing construction in Russia *Business. Education. Right. Bulletin of the Volgograd Institute of business* 2(23) 120-123

[8] Rosstat URL: https://rosstat.gov.ru/folder/11186 (accessed: 07.05.2021)