UDC 519.245 (045)

Kuzmyn V.M., PhD,
Zaliskyi M.Yu., PhD,
Kozhokhina O.V., PhD

APPROXIMATION OF THE EXPONENTIAL TYPE TIME SERIES IN TERMS OF THE USA POPULATION GROWTH RATE

National Aviation University
arec@nau.edu.ua
maximus2812@ukr.net
kozhokhina@gmail.com

The article represents a new approach to the exponential polygonal mathematical model formulation at the time series approximation in terms of the USA population growth rate

Keywords: exponential polygonal (linear two-segment) regression, approximation, sum of squared deviations minimization, cumulative curve of remainders

Introduction

The main requirement to the process of mathematical models formulation is the most accurate approximation and the correct definition of the prediction.

Another requirement is to obtain the most cost-effective model (with a minimum number of statistically calculated parameters).

Linear approximation, polynomial approximation, spline approximation and etc, are the most commonly used to formulate the mathematical models. However, these methods often have a low accuracy of prediction. Prediction objectives are widely used in all branches of human activity [1].

According to [2] the prediction process can be divided into four types:

a) short-term forecasting (from several months to two years),

b) medium-term forecasting (from two to five years),

c) long-term forecasting (from five to twenty years)

d) very-long-term forecast (twenty years old or more).

The confidence interval for errors forecasting is the greatest for the very-long-term forecast.

Main part

Let us consider the real example of the USA population changes rate within 1790 – 1990, [3]. We shall use the polygonal regression, which is the exponent order quantity.

$$y(x) = \exp(a_0 + a_1 x + a_2 (x - x_0) h(x - x_0)),$$

where $h(x)$ is Heaviside function, a_0, a_1, a_2 are unknown constant coefficients and x_0 is a switch point.

The initial data is given in the table. At the same, to simplify, we take population value in millions through decades, and also the logarithms of these values.

Table. The USA population changes rate within 1790 – 1990

Year	1790	1800	1810	1820	1830	1840	1850
Value	3,9	5,3	7,2	9,6	12	17	23
Logarithm	1,361	1,668	1,974	2,262	2,485	2,833	3,135
Year	1860	1870	1880	1890	1900	1910	1920
Value	31	38	50	62	75	91	105
Logarithm	3,434	3,638	3,912	4,127	4,317	4,511	4,654
Year	1930	1940	1950	1960	1970	1980	1990
Value	122	131	151	17	203	226	249
Logarithm	4,804	4,875	5,017	5,187	5,313	5,421	5,517

As long as we are going to use the exponential approximation, the logarithms of the original values are used for the analysis.

The first step is to approximate the logarithmic data using linear regression in terms of the least square method.

A regression equation has the form

$$z(x) = \ln y(x) = a_0 + a_1 x, \qquad (1)$$

where a_0, a_1 are unknown constant coefficients.

This equation assumes the presence of the sufficient usage of linear approximation to the original data. Coefficients in the formula (1) can be found by the least square method.

At the same time, two obtained equations are solved by the matrix method to find them:

$$XA = Z,$$

where

$$X = \begin{pmatrix} 21 & \sum_{i=1}^{21} x_i \\ \sum_{i=1}^{21} x_i & \sum_{i=1}^{21} (x_i)^2 \end{pmatrix}, \; Z = \begin{pmatrix} \sum_{i=1}^{21} z_i \\ \sum_{i=1}^{21} (x_i z_i) \end{pmatrix},$$

$$A = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = X^{-1} Z =$$

$$= \frac{1}{\Delta} \begin{pmatrix} \sum_{i=1}^{21} (x_i)^2 & -\sum_{i=1}^{21} x_i \\ -\sum_{i=1}^{21} x_i & 21 \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{21} z_i \\ \sum_{i=1}^{21} (x_i z_i) \end{pmatrix},$$

where i is the serial number of the decade, x_i is a number of year (to simplify the calculations in the article, we have taken years not from 1790, but from 1800), z_i are empirical values of the population magnitude logarithm.

Furthermore, using Cramer's method, the final formula can be obtained for the approximation coefficients in form of:

$$\Delta = 21 \cdot \sum_{i=1}^{21} (x_i)^2 - \left(\sum_{i=1}^{21} x_i \right)^2,$$

$$a_0 = \frac{\sum_{i=1}^{21} (x_i)^2 \sum_{i=1}^{21} z_i - \sum_{i=1}^{21} x_i \sum_{i=1}^{21} (x_i z_i)}{21 \sum_{i=1}^{21} (x_i)^2 - \left(\sum_{i=1}^{21} x_i \right)^2},$$

$$a_1 = \frac{21 \sum_{i=1}^{21} (x_i z_i) - \sum_{i=1}^{21} x_i \sum_{i=1}^{21} z_i}{21 \sum_{i=1}^{21} (x_i)^2 - \left(\sum_{i=1}^{21} x_i \right)^2}.$$

Graphical representation of the initial data logarithm, and their linear approximations are shown in Fig. 1.

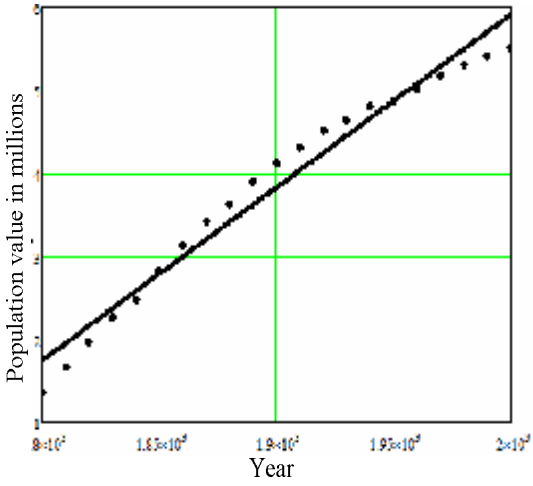

Fig.1.Graphical representation of the initial data logarithm, and their linear approximations

Visual analysis of fig. 1 shows that the tested set of data has the property of non-linearity, which generally must be checked using statistical methods.

There are several statistical methods for data analysis on non-linearity. In this article, the new method is used, proposed in [4].

To check the data on the linearity, let us detect the cumulative curve of remainders (the initial data deviations from the linear approximation).

The graphical view of the cumulative curve of remainders (CCR) is shown in Fig. 2.

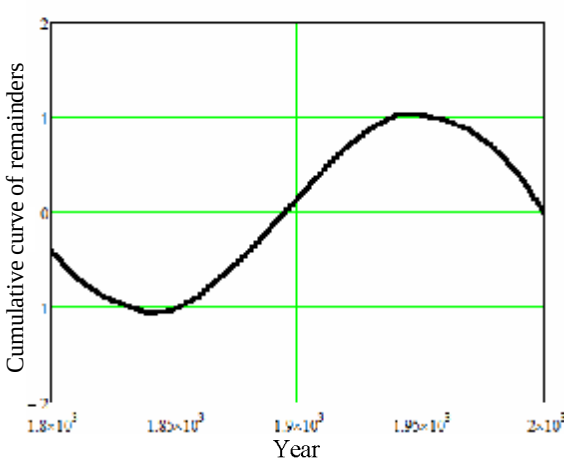

Fig. 2. Cumulative curve of remainders

The above calculations show that the resulting curve has a span of $r = 2{,}102$. To be able to apply the test method [4] it is also necessary to find the standard deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{21}(z_i)^2 - a_0\sum_{i=1}^{21}(x_i) - a_1\sum_{i=1}^{21}(x_i z_i)}{n-2}}.$$

In the equation, the parameter is $\sigma = 0{,}238$. The ratio of CCR span to the standard deviation is equal to 8,829. This suggests that, the initial data have the non-linear nature at a probability belief of more than 0,99 [5].

In addition, the resulting outlook for 2020 at such approximation is 561,645 million people, and it is overestimated.

Therefore, it is necessary to use a polygon regression for a better approximation. During plotting of polygonal regressions, the problem of finding the optimal switching point occurs.

Therefore, let us consider three approximation options at polygonal regression with different switching points of abscissas. Let abscissas switching points be: 1880, 1890 and 1900. These options should be considered for our further possibility to optimize the location of the switching point.

The equation of polygon regression has the form

$$z(x) = a_0 + a_1 x + a_2 (x - x_0) h(x - x_0),$$

where x_0 is switch point and $h(x)$ is Heaviside function.

Coefficients are found by solving a system of linear equations derived by the least square method, using a matrix method

$$XA = Z,$$

$$X = \begin{pmatrix} 21 & \sum_{i=1}^{21} x_i & \sum_{i=x_0}^{21}(x_i - x_0) \\ \sum_{i=1}^{21} x_i & \sum_{i=1}^{21}(x_i)^2 & \sum_{i=x_0}^{21} x_i(x_i - x_0) \\ \sum_{i=x_0}^{21}(x_i - x_0) & \sum_{i=x_0}^{21} x_i(x_i - x_0) & \sum_{i=x_0}^{21}(x_i - x_0)^2 \end{pmatrix},$$

$$Z = \begin{pmatrix} \sum_{i=1}^{21} z_i \\ \sum_{i=1}^{21}(x_i z_i) \\ \sum_{i=x_0}^{21}((x_i - x_0) z_i) \end{pmatrix},$$

$$A = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = X^{-1} Z.$$

The following approximation equations were obtained for the observed example

$$z(x) = 1{,}366 + 0{,}029x - 0{,}015(x - 1880)h(x - 1880),$$

$$z(x) = 1{,}397 + 0{,}028x - 0{,}015(x - 1890)h(x - 1890),$$

$$z(x) = 1{,}43 + 0{,}027x - 0{,}015(x - 1900)h(x - 1900).$$

Sums of squared deviations were found for each of the three obtained variants, and they are 0.207; 0.167; 0.209.

The data analysis shows that standard deviation data has its minimum. So, the minimum can be determined after the approximation of the obtained quadratic curve data (Fig. 3).

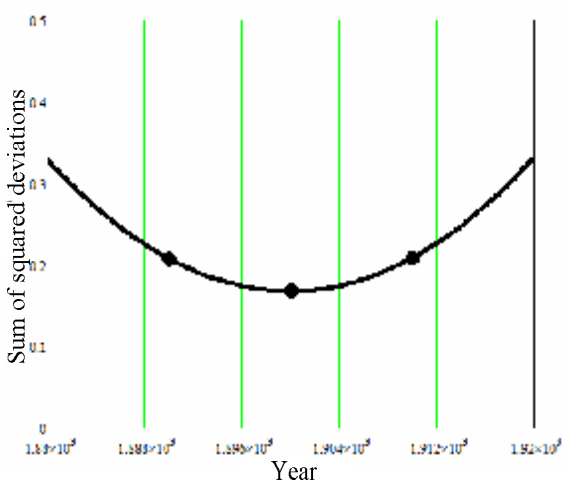

Fig.3. Sum of squared deviations approximation

The minimum is located at the point with abscissa 1889,917, which is taken as the optimum switching point abscissa.

Then the optimal equation of the polygon regression has a form

$$z(x) = 1{,}397 + 0{,}028x -$$

$$-0{,}015(x - 1889{,}917)h(x - 1889{,}917).$$

Graphical view of the polygonal approximation of the regression is shown in Fig. 4.

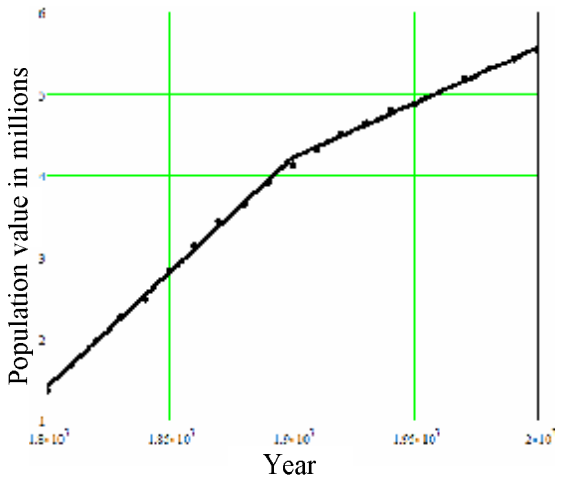

Fig.4. Graphical view of a polygonal approximation of the regression

In general case, the detecting of the switching point is an iterative procedure. The optimization is performed in the first iteration in this article. If desired to specify optimum, additional iterations may be performed.

The resulting forecast for 2020 consists of 390,753 million people at this approximation.

This result is more correct, in our point of view, than the one at the exponential-linear approximation, which reaches up to 561,645 million people. Thus, it is seen that the empirical data formal approximation with conventional linear dependence leads to large prediction errors.

Very-long-term forecast, for example in 2050, includes 585,434 million people for the obtained optimal approximation.

Conclusions

The article represents a new approach to the exponential polygonal arithmetic model formulation at the time series approximation in terms of the USA population growth rate, which made it possible to solve the very-long-term forecast problem more precise, compared with conventional linear exponential approximation. Such model formulation became possible, as a result of the new non-linearity test usage.

References

1. McClave, J.T. Statistics / J.T. McClave, F.H. Dietrich. – San Francisco: Dellen Publishing Company, 1991. – 624 p.

2. Chetyrkin E.M. Statistical methods of prediction. – M.: Statistics, 1977. – 200 p. (in Russian)

3. Blanchard P., Devaney R.L., Hall G. R. Differential Equation. Pacific Grove CA, 1997, 162 p.

4. Kuzmin V.N. New Statistical Method for Identification of Nonlinearity of Empirical Data // Computer data analysis and modeling. Proceedings of the Fifth International Conference. June, 8 – 12, 1998, Minsk, Volume 1: A-M, PP. 159 – 164.

5. Kuzmin V.N., Solomentsev O.V., Zaliskyi M.Yu. The methodical approach to testing the linearity of statistical data // Problems of Informatization and Control. – 2015. – № 4 (52). – PP. 63 – 67. (in Ukrainian)

Article to the Editor 01.09.2016