New approach to switching points optimization for segmented regression during mathematical model building

Valeriyi M. Kuzmin[1], Maksym Yu. Zaliskyi[1], Roman S. Odarchenko[1] and Yuliia V. Petrova[1]

[1]National Aviation University, 1 Lubomyr Huzar Ave., Kyiv, 03058, Ukraine

Abstract

Mathematical models building is widely used in different branches of human activity to describe statistical data obtained during observation of various phenomena. The main tool for this problem solution is approximation theory, especially ordinary least squares method. Basic goal during approximation is minimizing deviation between observed and estimated data. Analysis showed that providing given accuracy is possible based on usage of segmented regression models. Such models contain one or more switching points for segments connection. This paper deals with a problem of calculation of optimal values of switching point abscissa for segmented regression. Analytical expression for segmented regression was obtained using the Heaviside function. Switching point's determination is based on the usage of multidimensional optimization paraboloid. Paper presents the methodology for optimal segmented regression building. Simulation results and example of data processing proved increasing the accuracy of approximation in case of using the proposed methodology.

Keywords

mathematical model building, approximation, ordinary least squares method, segmented regression, optimization of switching point abscissa

1. Introduction

The mathematical models are used in many applications. Such models give the possibility to determine the mathematical relationship (formulas, logical dependency) for real world objects and phenomena. The one of the main motives to build mathematical models is: a) a greater understanding of researched phenomena, b) to analyze the object mathematically, c) to provide experimentation with model using simulation methods [1, 2].

The mathematical models building starts with experimental investigations and obtaining observations of some system, object or phenomenon. These operations form input data for model. According to these data, at the second stage mathematical formulations are carried out,

CS&SE@SW 2021: 4th Workshop for Young Scientists in Computer Science & Software Engineering, December 18, 2021, Kryvyi Rih, Ukraine

valeriyikuzmin@gmail.com (V. M. Kuzmin); maximus2812@ukr.net (M. Yu. Zaliskyi); odarchenko.r.s@ukr.net (R. S. Odarchenko); panijulia.p@gmail.com (Y. V. Petrova)

0000-0003-4461-9297 (V. M. Kuzmin); 0000-0002-1535-4384 (M. Yu. Zaliskyi); 0000-0002-7130-1375 (R. S. Odarchenko); 0000-0002-3768-7921 (Y. V. Petrova)

and after those computational simulations are performed. Output data of simulation are used for model validation [3].

During mathematical models building, different models can be utilized. Researcher always tries to choose the best of them [4]. To do this the following criteria can be used: simplicity of mathematical equation with the given level of error, minimum number of coefficients in the mathematical equation, minimum sum of squared deviations between the predicted and empirical values and others [5].

The main algorithmic tool that is used to obtain information from mathematical models contains methods of linear algebra, data analysis, probability theory and mathematical statistics, functional analysis and others [6]. The mathematical models based on statistical data-driven approach can be built using the techniques of the approximation theory [7]. In case of approximation, spline functions or different polynomials are often used [8].

2. Literature review and problem statement

Nowadays, regression analysis becomes popular research tool for mathematical models building [9]. It allows to develop mathematical expressions to describe the behavior of some dependent random variable [10]. Regression analysis can be used to predict the value of dependent variable based on information of its previous realization trend.

The mathematical models building based on regression analysis can be used in different branches of human activity and scientific research:

- in econometrics: to analyze economics behavior for certain country or city dependent on one or more factors [11, 12];
- in biology: to obtain regional models of biological processes [13];
- for electrical engineering: to describe realizations of electrical signals and parameters of electronic devices [14, 15];
- in reliability theory: to build the mathematical model for trends of reliability parameters and diagnostics variables [16, 17];
- in aviation system: to build the mathematical model for Unmanned Aerial Vehicle (UAV) and aircraft flight routes [18, 19], to analyze the possibilities of UAV cyber security hazards [20], to calculate the efficiency of functioning of aviation equipment [21, 22], and others;
- for radar and navigation systems: to solve the problem of efficient target detection [23] and for approximation and prediction of data trends [24, 25, 26];
- during equipment operation: to calculate the optimal maintenance periodicity [27, 28] and to estimate the efficiency of diagnostics process [29, 30];
- for control systems: to find the correlation between statistical data for inertial stabilized platforms of ground vehicles [31] and to analyze possible control actions in case of aircraft departures and arrivals delays [32].

In practice, researchers apply simple linear regression [33] and more realistic nonlinear regression [34]. Considering nonlinear regression, it should be pointed that quadratic, cubic, exponential, segmented and even logistic regressions are widely used [35, 36]. Different software to implement such models was developed [37, 38].

As there are different types of regression curves, let $f_k(x_i, \overrightarrow{a}_{m,k})$ is set of k one-dimensional functions, any of them depends on vector $\overrightarrow{a}_{m,k}$ of m parameters and gives the estimate value $\widehat{y_i}$ for initial data in for two-dimensional array (x_i, y_i) with sample size n. According to existing results [9, 10, 33, 36], regression model with one independent variable can be presented as follows

$$Y = f_k(X, \overrightarrow{a}_{m,k}) + \epsilon,$$

where Y and X are the dependent and independent variables, ϵ is an error of evaluation.

For simple linear regression model $f_1(X, \overrightarrow{a}_{m,1}) = a_{0,1} + a_{1,1}X$, where $a_{0,1}$ and $a_{0,1}$ are parameters that must be determined [9].

To increase the accuracy of model, on the one hand, researchers use segmented regression techniques with several linear or parabolic sections for approximation empirical data [33]. On the other hand, additional analysis for heteroskedasticity in observed data trend is carried out [39, 40]. Literature analysis showed that unfortunately not enough attention is paid to another way of increasing the accuracy of model that is associated with calculation of optimal switching points (breakpoints or changepoints) between regression segments. To estimate the parameters of regression (including switching points), the maximum likelihood estimator (MLE) can be used [41, 42]. Moreover, paper [42] concentrates on replacing the traditional nonsmooth model with another that transitions smoothly at the switching point. Another approach can be based on Bayesian changepoint models [43, 44]. In some publications, there are attempts to solve this problem based on: 1) statistical simulation results using sequential search [45], 2) inverted F test confidence interval estimate for large sample sizes and bootstrapped confidence intervals estimate for small sample sizes [46]. Analysis of mentioned techniques for calculation of optimal switching points showed: a) MLEs require prior information on error distribution and approximate range of switching point, b) MLEs have bias of estimate, c) in some modifications MLE is the most computationally expensive, both in setup time and in run time, d) Bayesian estimators are more robust for difficult cases, but require additional prior limitations for model parameters. Moreover, the exact mathematical equations for optimal value of switching points in literature are not considered.

The *aim of this paper* is to develop a new approach to switching points optimization in case of segmented regression usage for mathematical models building. The calculation of the optimal values of abscissas of the switching points will give the possibility to increase the approximation accuracy and the possibility to improve the predictive properties.

From mathematical point of view, such problem can be considered as follows. At the first stage, it is necessary to choose the segmented approximation function $f_k(x_i, \overrightarrow{a}_{m,k})$ in such a way to minimize standard deviation σ between real values y_i and estimates $\widehat{y_i}$

$$k = inf(s \forall j : \sigma(f_s(x_i, \overrightarrow{a}_{m,s})) \leq \sigma(f_j(x_i, \overrightarrow{a}_{m,j}))). \tag{1}$$

At the second stage, it is necessary to carry out optimization of switching points abscissas x_{sw} and to find the corresponding values

$$(x_{swopt_1}, x_{swopt_2}, ..., x_{swopt_r}) = argmin(x_{sw_1}, x_{sw_2}, ..., x_{sw_r}), \tag{2}$$

where r is quantity of switching points in case of $r + 1$ segments for regression usage.

3. Methodology

The best preferred statistical data processing algorithms can be used in the conditions of aprioristic uncertainty [47]. In this research some limitations about aprioristic information was made.

After observation of random phenomenon, the two-dimensional array (x_i, y_i) with sample size n is collected. Initial data are plotted in two-dimensional space in form of dependence. Based on visual analysis of data, researcher can identify geometrical structure of data trend and choose the appropriate approximation function. Assume that only segmented functions can be used. Such function contains two or more segment without discontinuities. The segments are connected in the switching points. The quantity r of switching points or the quantity $r+1$ of segments is determined by researcher according to the analysis of geometrical structure of plotted data.

At the first step, type of segmented regression for data approximation is chosen. In authors opinion, it is enough to use one of three types of segmented regression:

1. Segmented linear regression

$$f_1(X) = a_{0,1} + a_{1,1}X + \sum_{i=1}^{r} a_{i+1,1}(X - x_{sw_i})h(X - x_{sw_i}), \tag{3}$$

where $h(X - x_{sw_i})$ is Heaviside step function.

In case of two segments usage, functional dependence (3) contains one switching point and three unknown coefficients. Equation (3) can be presented as follows

$$f_1(X) = a_{0,1} + a_{1,1}X + a_{2,1}(X - x_{sw_1})h(X - x_{sw_1}).$$

Unknown coefficients $a_{0,1}$, $a_{1,1}$ and $a_{2,1}$ are calculated according to ordinary least squares method in such a way

$$a = W^{-1}B, a = \begin{pmatrix} a_{0,1} \\ a_{1,1} \\ a_{2,1} \end{pmatrix}, B = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} (x_i - x_{sw_1}) y_i h_1 \end{pmatrix}, h_1 = h(x_i - x_{sw_1}),$$

$$W = \begin{bmatrix} n & \sum_1^n x_i & \sum_1^n (x_i - x_{sw_1})h_1) \\ \sum_1^n x_i & \sum_1^n x_i^2 & \sum_1^n (x_i - x_{sw_1})x_i h_1 \\ \sum_1^n (x_i - x_{sw_1})h_1 & \sum_1^n (x_i - x_{sw_1})x_i h_1 & \sum_1^n (x_i - x_{sw_1})^2 h_1 \end{bmatrix}.$$

2. Segmented parabolic regression

$$f_2(X) = a_{0,2} + a_{1,2}X + a_{2,2}X^2 + \sum_{i=1}^{r} a_{i+2,1}(X - x_{sw_i})^2 h(X - x_{sw_i}). \tag{4}$$

In the case of two segments usage, functional dependence (4) contains one switching point and four unknown coefficients. Equation (4) can be presented as follows

$$f_2(X) = a_{0,2} + a_{1,2}X + a_{2,2}X^2 + a_{3,2}(X - x_{sw_1})^2 h(X - x_{sw_1}).$$

Unknown coefficients $a_{0,2}$, $a_{1,2}$, $a_{2,2}$ and $a_{3,2}$ are calculated according to ordinary least squares method in such a way

$$a = W^{-1}B, a = \begin{pmatrix} a_{0,2} \\ a_{1,2} \\ a_{2,2} \\ a_{3,2} \end{pmatrix}, B = \begin{pmatrix} \sum_1^n y_i \\ \sum_1^n x_i y_i \\ \sum_1^n x_i^2 y_i \\ \sum_1^n t_i^2 y_i h_1 \end{pmatrix}, t_i = x_i - x_{sw_1}$$

$$W = \begin{bmatrix} n & \sum_1^n x_i & \sum_1^n x_i^2 & \sum_1^n t_i^2 h_1 \\ \sum_1^n x_i & \sum_1^n x_i^2 & \sum_1^n x_i^3 & \sum_1^n t_i^2 x_i h_1 \\ \sum_1^n x_i^2 & \sum_1^n x_i^3 & \sum_1^n x_i^4 & \sum_1^n t_i^2 x_i^2 h_1 \\ \sum_1^n t_i^2 h_1 & \sum_1^n t_i^2 x_i h_1 & \sum_1^n t_i^2 x_i^2 h_1 & \sum_1^n t_i^4 h_1 \end{bmatrix}.$$

3. Segmented linear-parabolic regression

$$f_3(X) = a_{0,3} + a_{1,3}X + a_{2,3}X^2 p(X) + \sum_{i=1}^{r} a_{i+2,1}(X - x_{sw_i})^{p(X)+1} h(X - x_{sw_i}), \quad (5)$$

where $p(X)$ is sign function. This function is equal to zero, if the segment is linear, and is equal to one, if the segment is parabolic.

In the case of two segments usage with first parabolic and second linear segment, functional dependence (5) contains one switching point and three unknown coefficients. Equation (5) can be presented as follows

$$f_3(X) = a_{0,3} + a_{1,3}X + a_{2,3}X^2 - a_{2,3}(X - x_{sw_1})^2 h(X - x_{sw_1}).$$

Unknown coefficients $a_{0,3}$, $a_{1,3}$ and $a_{2,3}$ are calculated according to ordinary least squares method in such a way

$$a = W^{-1}B, a = \begin{pmatrix} a_{0,3} \\ a_{1,3} \\ a_{2,3} \end{pmatrix}, B = \begin{pmatrix} \sum_1^n y_i \\ \sum_1^n x_i y_i \\ \sum_1^n x_i^2 y_i - \sum_1^n t_i^2 y_i h_1 \end{pmatrix},$$

$$W = \begin{bmatrix} n & \sum_1^n x_i & \sum_1^n x_i^2 - \sum_1^n t_i^2 h_1 \\ \sum_1^n x_i & \sum_1^n x_i^2 & \sum_1^n x_i^3 - \sum_1^n x_i t_i^2 h_1 \\ \sum_1^n x_i^2 - \sum_1^n t_i^2 h_1 & \sum_1^n x_i^3 - \sum_1^n x_i t_i^2 h_1 & \sum_1^n x_i^4 + \sum_1^n (t_i^4 - 2x_i^2 t_i^2) h_1 \end{bmatrix}.$$

At the second step, the quantity r of switching points and the range of possible values of abscissas of switching points is selected subjectively based on visual analysis of observed data. For this approach, it is necessary to choose at least five possible values for each switching point. So matrix of vectors of possible abscissa values is generated in the following form $(\overrightarrow{x}_{sw_1}, \overrightarrow{x}_{sw_2}, ..., \overrightarrow{x}_{sw_r})$.

At the third step, regression coefficients and standard deviations σ between real values y_i and estimates $\widehat{y_i}$ for all segmented regression types are calculated. Standard deviation is determined according to the equation

$$\sigma = \sqrt{\frac{1}{n-l} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2}, \quad (6)$$

where l is a degree of freedom for selected model.

The standard deviation is calculated for all combinations of possible values of switching point abscissa. So at this step, the r-dimensional dependence of $\sigma(x_{sw_1}, x_{sw_2}, ..., x_{sw_r})$ is obtained.

At the fourth step, the obtained dependence is approximated by r-dimensional paraboloid based on ordinary least squares method. The general equation of r-dimensional paraboloid

$$z(x_{sw_1}, x_{sw_2}, ..., x_{sw_r}) = A_0 + sum_{i=1}^{r} A_i x_{sw_i}^2 + \sum_{i=1}^{r} B_i x_{sw_i} + \sum_{i<j} C_{i,j} x_{sw_i} x_{sw_j}, \quad (7)$$

where A_i, B_i, $C_{i,j}$ are unknown coefficients need to be estimated, the sum is calculated only for $i < j$.

To simplify the calculation, it can be assumed that $C_{i,j} = 0$ and equation (7) will take a form

$$z(x_{sw_1}, x_{sw_2}, ..., x_{sw_r}) = A_0 + \sum_{i=1}^{r} A_i x_{sw_i}^2 + \sum_{i=1}^{r} B_i x_{sw_i}. \quad (8)$$

In this case unknown coefficients can be found according to the following equation

$$a = W^{-1}B, a = \begin{pmatrix} A_0 \\ A_1 \\ B_1 \\ ... \\ A_r \\ B_r \end{pmatrix}, B = \begin{pmatrix} \sum_1^v ... \sum_1^v z_{i_1,i_2,...,i_r} \\ \sum_1^v ... \sum_1^v x_{sw_1 i_1}^2 z_{i_1,i_2,...,i_r} \\ \sum_1^v ... \sum_1^v x_{sw_1 i_1} z_{i_1,i_2,...,i_r} \\ ... \\ \sum_1^v ... \sum_1^v x_{sw_1 i_r}^2 z_{i_1,i_2,...,i_r} \\ \sum_1^v ... \sum_1^v x_{sw_1 i_r} z_{i_1,i_2,...,i_r} \end{pmatrix}, g = v^{r-1}$$

$$W = \begin{bmatrix} v^r & g\sum_1^v x_{sw_1 i_1}^2 & g\sum_1^v x_{sw_1 i_1} & ... & g\sum_1^v x_{sw_r i_1} \\ g\sum_1^v x_{sw_1 i_1}^2 & g\sum_1^v x_{sw_1 i_1}^4 & g\sum_1^v x_{sw_1 i_1}^3 & ... & g\sum_1^v x_{sw_1 i_1}^2 x_{sw_r i_1} \\ g\sum_1^v x_{sw_1 i_1} & g\sum_1^v x_{sw_1 i_1}^3 & g\sum_1^v x_{sw_1 i_1}^2 & ... & g\sum_1^v x_{sw_1 i_1} x_{sw_r i_1} \\ ... & ... & ... & ... & ... \\ g\sum_1^v x_{sw_r i_1} & g\sum_1^v x_{sw_1 i_1}^2 x_{sw_r i_1} & g\sum_1^v x_{sw_1 i_1} x_{sw_r i_1} & ... & g\sum_1^v x_{sw_1 i_r}^2 \end{bmatrix}.$$

where v is quantity of chosen points in the range of possible values of abscissas of switching points.

At the fifth step, the minimum of r-dimensional paraboloid is calculated to provide the criterion (2). For this purpose, the theory of optimization is used [48]. To find the minimum, it is necessary to solve the system of equations

$$\begin{cases} \frac{\partial z(x_{sw_1}, x_{sw_2}, ..., x_{sw_r})}{\partial x_{sw_1}} = 0, \\ \frac{\partial z(x_{sw_1}, x_{sw_2}, ..., x_{sw_r})}{\partial x_{sw_2}} = 0, \\ ... \\ \frac{\partial z(x_{sw_1}, x_{sw_2}, ..., x_{sw_r})}{\partial x_{sw_r}} = 0. \end{cases} \quad (9)$$

In the case of r-dimensional paraboloid (7) usage, the system of equations (9) turns to the system of r linear equations that can be solved by one of known method. In case of simplified

paraboloid (8) usage, the simple solution can be obtained in the following form

$$x_{sw_i opt} = \frac{-B_i}{2A_i}. \tag{10}$$

At the sixth step, coefficients of segmented regression (3), (4) or (5) are recalculated, and resulting model is obtained.

4. Simulation results and numerical example

Consider the problem of analysis of proposed methodology implementation based on the results of statistical simulation.

The statistical simulation starts with obtaining initial data set with two switching points. The data set contains deterministic and random components. The deterministic component can be presented as follows

$$f_1(X) = a_{0,1} + a_{1,1}X + a_{2,1}(X - x_{sw_1})h(X - x_{sw_1}) + a_{3,1}(X - x_{sw_2})h(X - x_{sw_2}).$$

This dependence is converted into discrete form at the range [1; 100] with sampling interval $\delta = 1$ and sample size $n = 100$. The initial parameters of deterministic model can be different, but in this research, authors used the following initial numerical values: $a_{0,1} = 500$, $a_{1,1} = 10$, $a_{2,1} = -25$, $a_{3,1} = 20$, $x_{sw_1} = 20$ and $x_{sw_2} = 50$.

Random component is generated at each sample point as additive Gaussian noise with zero expected value and standard deviation $\sigma = 30$. The number of procedures reiteration is 1000.

The example of one of data sets is given in table 1. The data in the table 1 present the values of dependent variable Y that was measured at points X separated by sampling interval δ.

The graphical presentation of three examples of initial data set is shown in figure 1.

Visual analysis of data (figure 1) gives possibility to conclude that most convenient regression type for these data approximation is segmented linear regression with two switching points. Let $r = 5$. The range of possible values of abscissas of switching points is

$$x_{sw_1} = (10, 15, 20, 25, 30),$$

$$x_{sw_2} = (40, 45, 50, 55, 60).$$

In this case it is necessary to calculate estimates of regression coefficients $a_{0,1}$, $a_{1,1}$, $a_{2,1}$, $a_{3,1}$ for all combinations of possible values of abscissas of switching points. After that, standard deviation (6) is determined for each option. The results of standard deviation calculation are given in table 2.

Data from table 2 are approximated by two-dimensional paraboloid based on ordinary least squares methods. For paraboloid types (7) and (8) following equations were obtained

$$z(x_{sw_1}, x_{sw_2}) = 364.893 - 6.635x_{sw_1} - 10.012x_{sw_2} + 0.111x_{sw_1}^2 + 0.09x_{sw_2}^2 + 0.047x_{sw_1}x_{sw_2},$$

$$z(x_{sw_1}, x_{sw_2}) = 317.416 - 4.261x_{sw_1} - 9.062x_{sw_2} + 0.111x_{sw_1}^2 + 0.09x_{sw_2}^2,$$

The obtained paraboloids are shown in figure 2 and figure 3, respectively.

Table 1
Example of initial data set.

X	Y	X	Y	X	Y	X	Y	X	Y
1	478.051	21	708.727	41	430.555	61	361.496	81	391.604
2	531.887	22	716.929	42	397.554	62	357.442	82	410.622
3	488.646	23	698.735	43	440.372	63	281.227	83	370.187
4	532.988	24	662.582	44	324.692	64	362.172	84	460.596
5	437.424	25	554.083	45	372.758	65	336.362	85	345.848
6	576.916	26	663.423	46	343.182	66	341.036	86	356.448
7	558.703	27	621.014	47	304.289	67	288.759	87	408.922
8	525.774	28	692.666	48	380.215	68	401.393	88	459.006
9	561.106	29	522.092	49	252.287	69	321.402	89	340.568
10	598.737	30	659.452	50	333.319	70	290.943	90	443.487
11	631.717	31	398.557	51	307.979	71	436.479	91	541.709
12	658.255	32	520.615	52	270.906	72	333.381	92	436.921
13	647.998	33	472.390	53	290.251	73	373.471	93	462.618
14	607.476	34	463.161	54	265.407	74	343.770	94	532.297
15	648.630	35	442.640	55	240.342	75	354.348	95	484.741
16	691.087	36	443.975	56	269.936	76	402.171	96	451.064
17	638.839	37	482.674	57	338.144	77	377.978	97	505.605
18	687.825	38	433.265	58	284.574	78	303.512	98	439.356
19	689.012	39	405.900	59	351.267	79	339.748	99	450.629
20	653.723	40	444.444	60	243.165	80	312.829	100	485.727

Table 2
Standard deviations.

Abscissas	$x_{sw_1} = 10$	$x_{sw_1} = 15$	$x_{sw_1} = 20$	$x_{sw_1} = 25$	$x_{sw_1} = 30$
$x_{sw_2} = 40$	72.179	62.257	56.561	58.777	66.45
$x_{sw_2} = 45$	63.561	53.526	49.227	53.585	62.128
$x_{sw_2} = 50$	57.41	48.362	46.246	52.425	61.318
$x_{sw_2} = 55$	56.026	49.361	49.677	56.532	64.713
$x_{sw_2} = 60$	59.484	55.562	57.516	63.941	70.661

In the case of paraboloid (7) usage, it is necessary to solve system of equations (9) that takes a form

$$\begin{cases} \frac{\partial z(x_{sw_1}, x_{sw_2})}{\partial x_{sw_1}} = 0, \\ \frac{\partial z(x_{sw_1}, x_{sw_2})}{\partial x_{sw_2}} = 0. \end{cases}$$

After derivatives calculation this system of equations turns to system of linear equations

$$\begin{cases} -6.635 + 0.222 x_{sw_1 opt} + 0.047 x_{sw_2 opt} = 0, \\ -10.012 + 0.047 x_{sw_1 opt} + 0.18 x_{sw_2 opt} = 0. \end{cases}$$

The solution of this system is

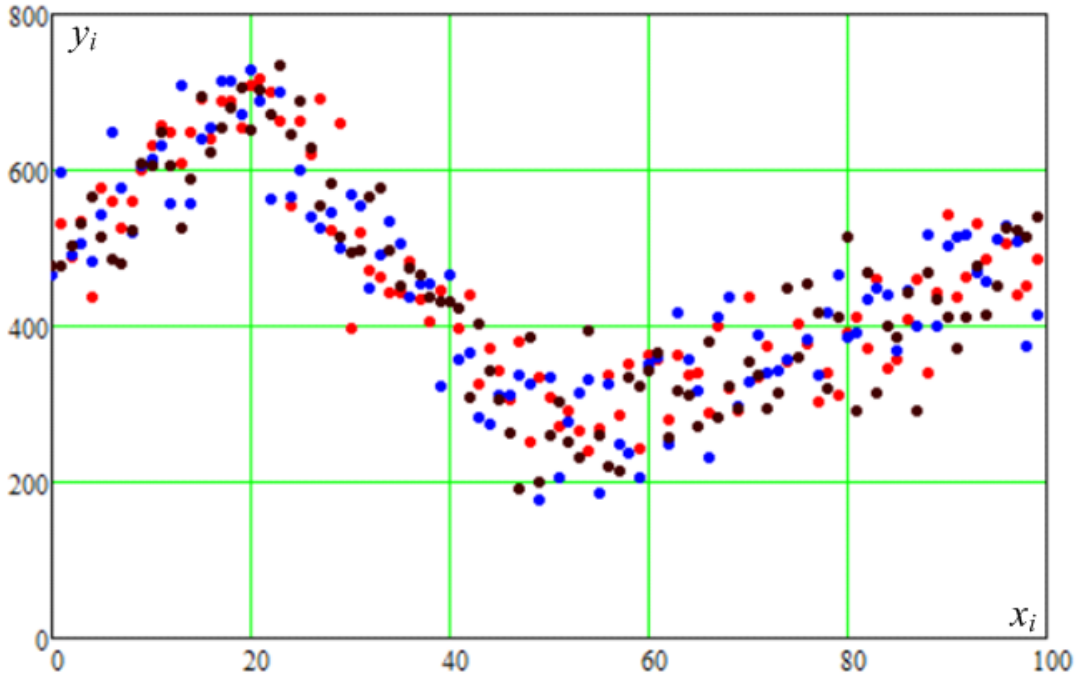

Figure 1: The initial data sets (three realizations).

$$x_{sw_1opt} = 18.941,$$

$$x_{sw_2opt} = 50.812.$$

In the case of paraboloid (8) usage, the optimal values of abscissas of switching points are calculated according to equation (10). The results of calculation

$$x^{/}_{sw_1opt} = 19.113,$$

$$x^{/}_{sw_2opt} = 50.532.$$

Analysis showed that for this particular case simplified paraboloid gives greater accuracy of switching point's abscissas estimates (relative error is 4.435 percent and 1.064 percent for the first and second switching points, respectively).

Resulting segmented linear regressions for both optimization options (paraboloids (7) and (8)) are

$$f_1(X) = 484.143 + 11.397X - 25.025(X - 18.941)h(X - 18.941)+$$

$$+18.021(X - 50.812)h(X - 50.812),$$

$$f_1(X) = 484.987 + 11.26X - 25.073(X - 19.113)h(X - 19.113)+$$

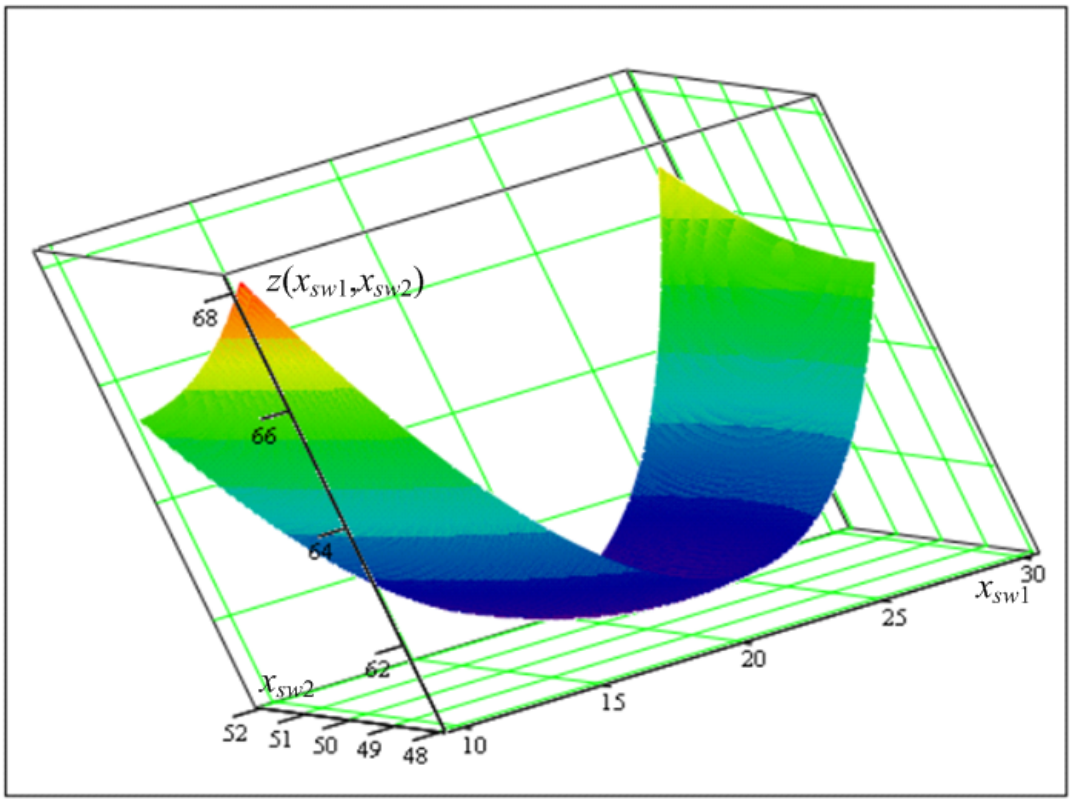

Figure 2: Obtained paraboloid (7) for data set from table 1.

$$+18.155(X - 50.532)h(X - 50.532).$$

The standard deviation for the first and second optimization options is 46.038 and 46.040, respectively. The results of approximation are shown in figure 4.

Resulting segmented linear regressions for both optimization options in figure 4 almost coincide and have approximately equal standard deviation.

Consider the statistical simulation results for 1000 reiteration procedures. Such simulation gives the possibility to build the probability density functions of estimates of switching point's abscissas. Figure 5 shows the histograms for estimate of abscissa of the first (figure 5a) and second (figure 5c) switching point for paraboloid (7), the histograms for estimate of abscissa of the first (figure 5b) and second (figure 5d) switching point for paraboloid (8). Statistical characteristics (expected value, variance, minimum and maximum) of estimates for optimal values of abscissas of switching points using paraboloids (7) and (8) are given in table 3.

Analysis showed that general paraboloid (7) in average has greater accuracy for switching points abscissas estimation. In the case of the first switching points abscissas estimation, relative error is 3.63 and 4.32 percents for paraboloid (7) and (8), respectively. In the case of second switching points abscissas estimation, relative error is 0.968 and 1.376 percents for paraboloid (7) and (8), respectively. In addition, paraboloid (7) has greater scattering of estimate.

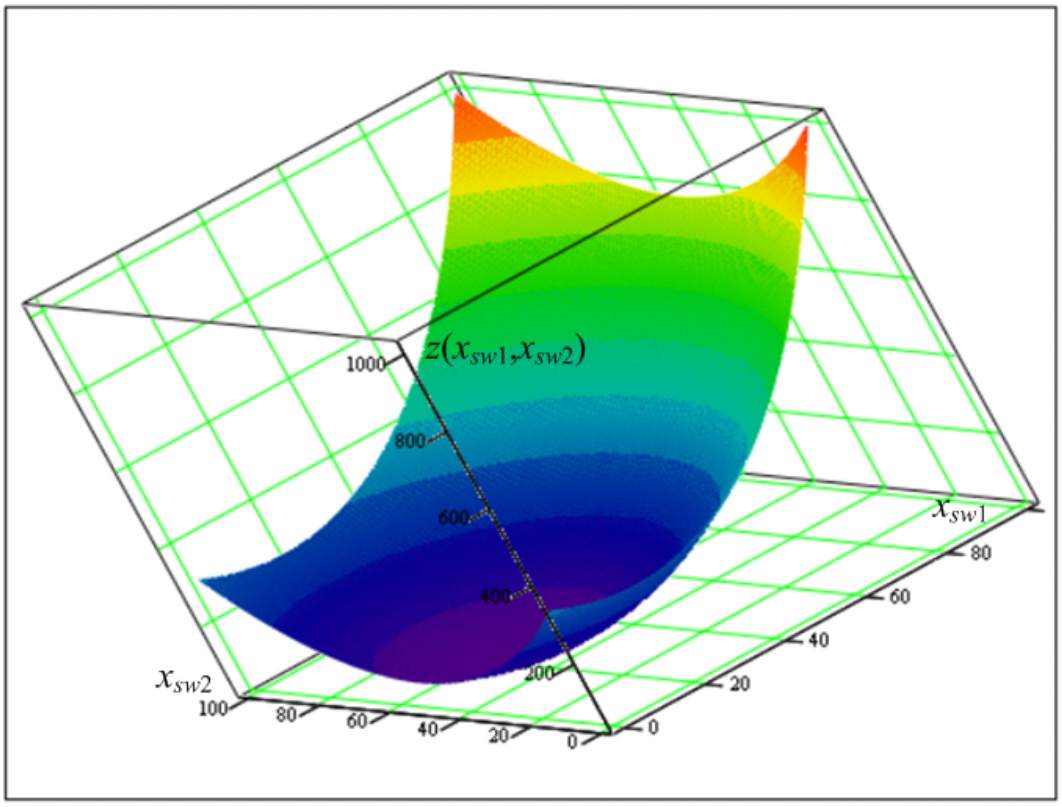

Figure 3: Obtained paraboloid (8) for data set from Table 1.

The simulation results give approximatly same efficiency of estimate and accuracy of mathematical model. So to simplify the calculation, optimizational paraboloid (8) can be used as more suitable during mathematical model building.

5. Conclusion

The paper considers new approach to switching point's optimization for segmented regression during mathematical model building. The analytical equations for segmented linear, parabolic and linear-parabolic regressions are presented based on usage of Heaviside step function. To find the optimal values of connection points between regression segments, multidimensional optimization paraboloid is used for describing the dependence of standard deviation on possible values of switching point's abscissa. The proposed methodology, in contrast to the existing ones, allows to obtain the accurate mathematical formula for calculating the abscissa of switching points. Moreover, considered methodology has property of robustness for initial distribution of errors and dataset. The analysis of proposed methodology is carried out based on statistical simulation. The implementation of methodology is explained on numerical example for generated data set. Computations prove feasibility of proposed approach. The research results can be

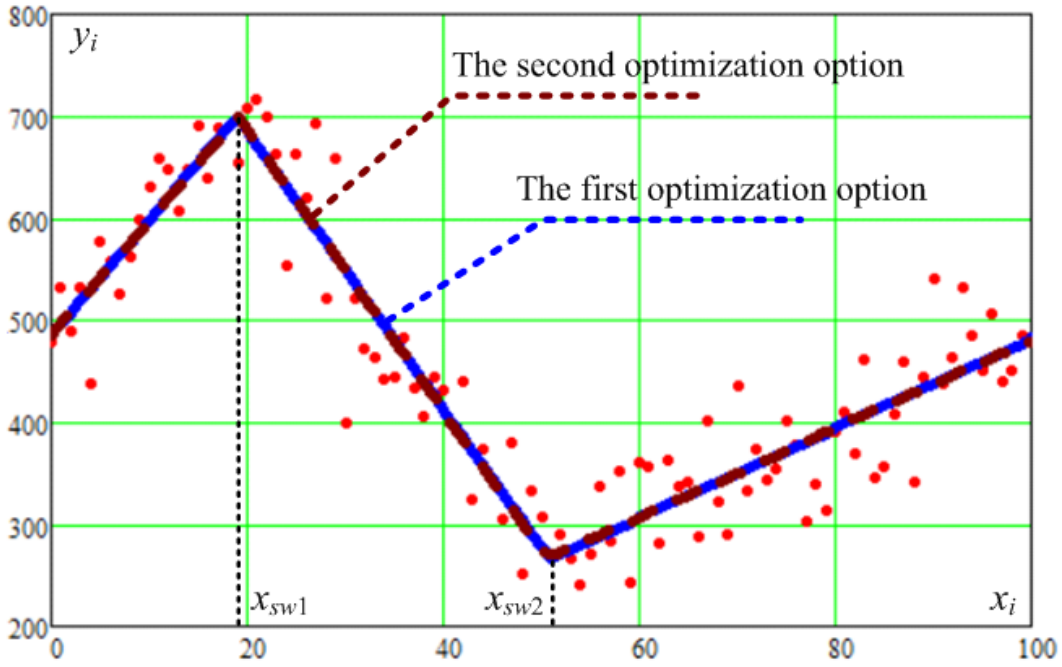

Figure 4: The initial data set and obtained optimal segmented linear regressions.

Table 3
Statistical characteristics of estimates for optimal values of abscissas of switching points using paraboloids (7) and (8)

Statistical characteristic	Paraboloid (7)	Paraboloid (8)
Expected value for x_{sw_1}	20.726	20.864
Variance for x_{sw_1}	1.427	1.317
Minimum value for x_{sw_1}	15.892	16.929
Maximum value for x_{sw_1}	25.004	25.026
Expected value for x_{sw_2}	50.484	50.688
Variance for x_{sw_2}	1.314	1.188
Minimum value for x_{sw_2}	45.978	45.914
Maximum value for x_{sw_2}	54.883	55.062

used to increase the accuracy of data approximation in mathematical model building.

Further research directions will be associated with a comparative analysis of the effeciency of the proposed methodology with other techniques for determining estimates of the abscissa of switching points (in particular, MLE and estimates based on the Bayesian approach) in the case of different limitations presence.

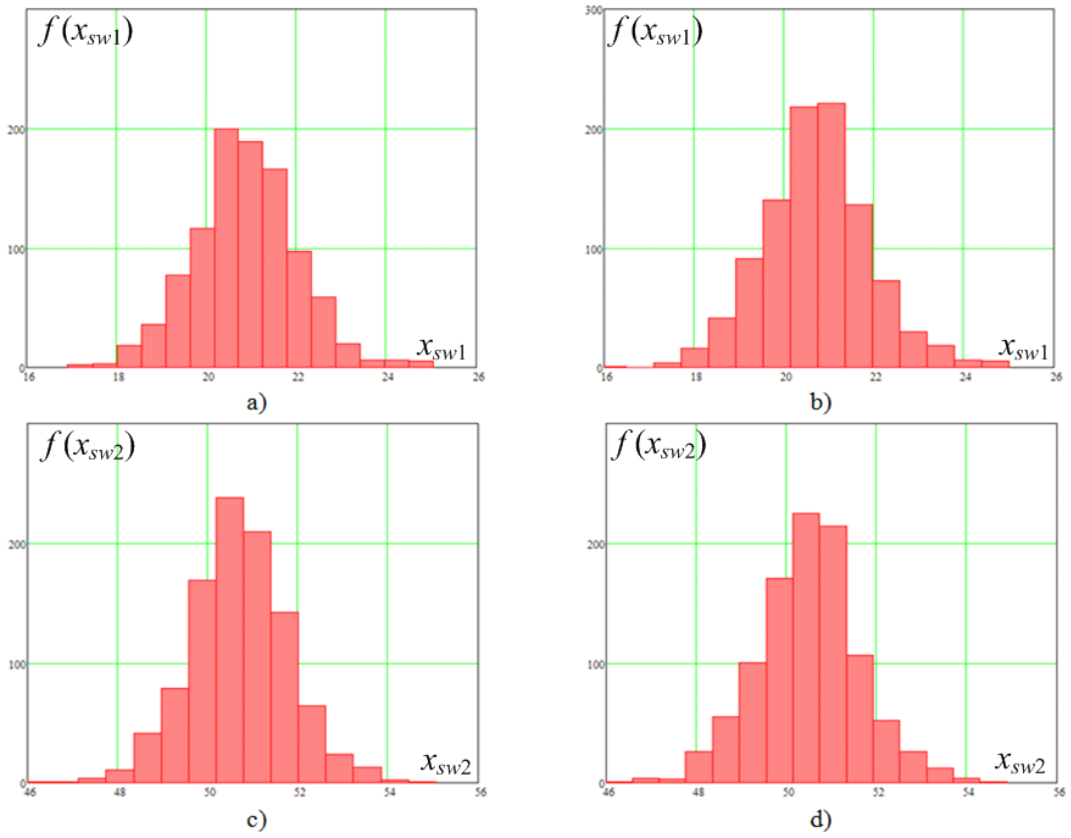

Figure 5: The histograms of estimates of switching point's abscissas.

References

[1] H. P. Williams, Model Building in Mathematical Programming, Wiley, 2013.

[2] I. V. Ostroumov, K. Marais, N. S. Kuzmenko, N. Fala, Triple probability density distribution model in the task of aviation risk assessment, Aviation 24 (2020) 57–65. doi:10.3846/aviation.2020.12544.

[3] M. Banwatth-Kuhn, S. Sindi, How and why to build a mathematical model: A case study using prion aggregation, Journal of Biological Chemistry 295 (2020) 5022–5034. doi:10.1074/jbc.REV119.009851.

[4] A. K. Mitropolsky, The Technique of Statistical Computing, Moscow, 1971.

[5] D. M. Himmelblau, Process Analysis by Statistical Methods, Wiley, 1970.

[6] A. Neumaier, Mathematical model building, in: J. Kallrath (Ed.), Modeling Languages in Mathematical Optimization. Applied Optimization, volume 88, University of Chicago Press, Boston, MA, 2004, pp. 37–43. doi:10.1007/978-1-4613-0215-5_3.

[7] M. Ezekiel, K. A. Fox, Method of Correlation and Regression Analysis. Linear and Curvilinear, John Wiley and Sons, New York, 1959.

[8] I. V. Ostroumov, N. S. Kuzmenko, Accuracy improvement of VOR/Vor navigation with

angle extrapolation by linear regression, Telecommunications and Radio Engineering 78 (2019) 1399–1412. doi:10.1615/TelecomRadEng.v78.i15.90.

[9] T. P. Ryan, Modern Regression Methods, 2 ed., John Wiley and Sons, New York, 2008.

[10] J. O. Rawlings, S. G. Pantula, D. A. Dickey, Applied Regression Analysis: A Research Tool, second ed., Springer-Verlag, New York, NY, 1998.

[11] H. Zhang, Research of the performance and influencing factors of china's listed companies based on regression model, in: Proceedings of 16th Dahe Fortune China Forum and Chinese High-educational Management Annual Academic Conference (DFHMC), 2020, pp. 176–179. doi:10.1109/DFHMC52214.2020.00041.

[12] Y. Wang, Linkages between metropolitan economy and modem logistics based on linear regression analysis, in: Proceedings of 2nd International Conference on Economic Management and Model Engineering (ICEMME), 2020, pp. 64–67. doi:10.1109/ICEMME51517.2020.00019.

[13] P. Radonja, S. Stankovic, B. Matovic, D. Drazic, Regional models for biological processes based on linear regression and neural networks, in: Proceedings of 8th Seminar on Neural Network Applications in Electrical Engineering, 2006, pp. 189–193. doi:10.1109/NEUREL.2006.341209.

[14] R. Volianskyi, O. Sadovoi, N. Volianska, O. Sinkevych, Construction of parallel piecewise-linear interval models for nonlinear dynamical objects, in: Proceedings of International Conference on Advanced Computer Information Technologies, 2019, pp. 97–100. doi:10.1109/ACITT.2019.8779945.

[15] X. Feng, Y. Zhou, T. Hua, Y. Zou, J. Xiao, Contact temperature prediction of high voltage switchgear based on multiple linear regression model, in: Proceedings of 32nd Youth Academic Annual Conference of Chinese Association of Automation (YAC), 2017, pp. 277–280. doi:10.1109/YAC.2017.7967419.

[16] O. Solomentsev, V. Kuzmin, M. Zaliskyi, O. Zuiev, Y. Kaminskyi, Statistical data processing in radio engineering devices operation system, in: Proceedings of 14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering (TCSET), 2018, pp. 757–760. doi:10.1109/TCSET.2018.8336310.

[17] M. Zaliskyi, O. Solomentsev, N. Kuzmenko, F. Yanovsky, O. Shcherbyna, O. Sushchenko, I. Ostroumov, Y. Averyanova, Sequential method of reliability parameters estimation for radio equipment, in: 2021 IEEE 12th International Conference on Electronics and Information Technologies (ELIT), 2021, pp. 37–40. doi:10.1109/ELIT53502.2021.9501099.

[18] V. P. Kharchenko, N. S. Kuzmenko, I. V. Ostroumov, Identification of unmanned aerial vehicle flight situation, in: Proceedings of 2017 IEEE 4th International Conference on Actual Problems of Unmanned Aerial Vehicles Developments (APUAVD), 2017, pp. 116–120. doi:10.1109/APUAVD.2017.8308789.

[19] O. Ivashchuk, I. Ostroumov, N. Kuzmenko, O. Sushchenko, Y. Averyanova, O. Solomentsev, M. Zaliskyi, F. Yanovsky, O. Shcherbyna, A configuration analysis of ukrainian flight routes network, in: 2021 IEEE 16th International Conference on the Experience of Designing and Application of CAD Systems (CADSM), 2021, pp. 6–10. doi:10.1109/CADSM52681.2021.9385263.

[20] Y. Averyanova, O. Sushchenko, I. Ostroumov, N. Kuzmenko, M. Zaliskyi, O. Solomentsev, B. Kuznetsov, T. Nikitina, O. Havrylenko, A. Popov, V. Volosyuk, O. Shmatko, N. Ruzhentsev,

S. Zhyla, V. Pavlikov, K. Dergachov, E. Tserne, Uas cyber security hazards analysis and approach to qualitative assessment, in: S. Shukla, A. Unal, J. Varghese Kureethara, D. K. Mishra, D. S. Han (Eds.), Data Science and Security, Springer Singapore, Singapore, 2021, pp. 258–265. doi:10.1007/978-981-16-4486-3_28.

[21] I. Ostroumov, N. Kuzmenko, O. Sushchenko, V. Pavlikov, S. Zhyla, O. Solomentsev, M. Zaliskyi, Y. Averyanova, E. Tserne, A. Popov, V. Volosyuk, N. Ruzhentsev, K. Dergachov, O. Havrylenko, B. Kuznetsov, T. Nikitina, O. Shmatko, Modelling and simulation of dme navigation global service volume, Advances in Space Research 68 (2021) 3495–3507. doi:10.1016/j.asr.2021.06.027.

[22] I. Ostroumov, N. Kuzmenko, O. Sushchenko, Y. Averyanova, O. Shcherbyna, O. Solomentsev, F. Yanovsky, M. Zaliskyi, Ukrainian navigational aids network configuration estimation, in: 2021 IEEE 16th International Conference on the Experience of Designing and Application of CAD Systems (CADSM), 2021, pp. 5–9. doi:10.1109/CADSM52681.2021.9385226.

[23] Y. Averyanova, F. Yanovsky, O. Shcherbina, I. Ostroumov, N. Kuzmenko, M. Zaliskyi, O. Solomentsev, O. Sushchenko, Polarimetric-radar drop size evaluation for wind speed estimate based on Weber criterion, in: 2021 Signal Processing Symposium (SPSympo), 2021, pp. 17–22. doi:10.1109/SPSympo51155.2020.9593349.

[24] I. V. Ostroumov, N. S. Kuzmenko, Accuracy assessment of aircraft positioning by multiple radio navigational AIDS, Telecommunications and Radio Engineering 77 (2018) 705–715. doi:10.1615/TelecomRadEng.v77.i8.40.

[25] N. S. Kuzmenko, I. V. Ostroumov, Performance analysis of positioning system by navigational AIDS in three dimensional space, in: Proceedings of IEEE 1st International Conference on System Analysis and Intelligent Computing, 2018, pp. 101–104. doi:10.1109/SAIC.2018.8516790.

[26] I. V. Ostroumov, N. S. Kuzmenko, Compatibility analysis of multi signal processing in apnt with current navigation infrastructure, Telecommunications and Radio Engineering 77 (2018) 211–223. doi:10.1615/TelecomRadEng.v77.i3.30.

[27] A. Goncharenko, A multi-optional hybrid functions entropy as a tool for transportation means repair optimal periodicity determination, Aviation 22 (2018) 60–66. doi:10.3846/aviation.2018.5930.

[28] A. V. Goncharenko, Optimal UAV maintenance periodicity obtained on the multi-optional basis, in: Proceedings of 4th International Conference on Actual Problems of Unmanned Aerial Vehicles Developments, 2017, pp. 65–68. doi:10.1109/APUAVD.2017.8308778.

[29] O. Solomentsev, M. Zaliskyi, I. Yashanov, O. Shcherbyna, O. Sushchenko, F. Yanovsky, I. Ostroumov, Y. Averyanova, N. Kuzmenko, Substantiation of probability characteristics for efficiency analysis in the process of radio equipment diagnostics, in: 2021 IEEE 3rd Ukraine Conference on Electrical and Computer Engineering (UKRCON), 2021, pp. 535–540. doi:10.1109/UKRCON53503.2021.9575603.

[30] O. Shcherbyna, M. Zaliskyi, O. Solomentsev, N. Kuzmenko, F. Yanovsky, I. Ostroumov, Y. Averyanova, O. Sushchenko, Diagnostic process efficiency analysis for block diagram of electric field parameters meter, in: 2021 IEEE 12th International Conference on Electronics and Information Technologies (ELIT), 2021, pp. 5–9. doi:10.1109/ELIT53502.2021.9501136.

[31] O. Sushchenko, F. Yanovsky, O. Solomentsev, N. Kuzmenko, Y. Averyanova, M. Zaliskyi,

I. Ostroumov, O. Shcherbyna, Design of robust control system for inertially stabilized platforms of ground vehicles, in: IEEE EUROCON 2021 - 19th International Conference on Smart Technologies, 2021, pp. 6–10. doi:10.1109/EUROCON52738.2021.9535612.
[32] I. Ostroumov, N. Kuzmenko, O. Sushchenko, M. Zaliskyi, O. Solomentsev, Y. Averyanova, S. Zhyla, V. Pavlikov, E. Tserne, V. Volosyuk, K. Dergachov, O. Havrylenko, O. Shmatko, A. Popov, N. Ruzhentsev, B. Kuznetsov, T. Nikitina, A probability estimation of aircraft departures and arrivals delays, in: O. Gervasi, B. Murgante, S. Misra, C. Garau, I. Blečić, D. Taniar, B. O. Apduhan, A. M. A. Rocha, E. Tarantino, C. M. Torre (Eds.), Computational Science and Its Applications – ICCSA 2021, Springer International Publishing, Cham, 2021, pp. 363–377. doi:10.1007/978-3-030-86960-1_26.
[33] S. Weisberg, Applied Linear Regression, John Wiley and Sons, New York, 2005.
[34] G. A. F. Seber, C. J. Wild, Nonlinear Regression, John Wiley and Sons, New York, 2003.
[35] A. Atkinson, M. Riani, Robust Diagnostic Regression Analysis, Springer, 2000.
[36] D. G. Kleinbaum, M. Klein, Logistic Regression, Springer-Verlag, New York, 2002.
[37] S. Huet, A. Bouvier, M.-A. Poursat, E. Jolivet, Statistical Tools for Nonlinear Regression. A Practical Guide With S-PLUS and R Examples, Springer-Verlag, New York, 2004.
[38] A. Zeileis, F. Leisch, K. Hornik, C. Kleiber, An R package for testing for structural change in linear regression models, Journal of Statistical Software 7 (2002) 1–38. doi:10.18637/jss.v007.i02.
[39] R. L. Kaufman, Heteroskedasticity in Regression: Detection and Correction, SAGE Publications, 2013.
[40] M. Zaliskyi, O. Solomentsev, O. Shcherbyna, I. Ostroumov, O. Sushchenko, Y. Averyanova, N. Kuzmenko, O. Shmatko, N. Ruzhentsev, A. Popov, S. Zhyla, V. Volosyuk, O. Havrylenko, V. Pavlikov, K. Dergachov, E. Tserne, T. Nikitina, B. Kuznetsov, Heteroskedasticity analysis during operational data processing of radio electronic systems, in: S. Shukla, A. Unal, J. Varghese Kureethara, D. K. Mishra, D. S. Han (Eds.), Data Science and Security, Springer Singapore, Singapore, 2021, pp. 168–175. doi:10.1007/978-981-16-4486-3_18.
[41] A. Buteikis, Practical Econometrics and Data Science, Vilnius University, Vilnius, 2020. URL: http://web.vu.lt/mif/a.buteikis/wp-content/uploads/PE_Book/index.html.
[42] A. Tishler, I. Zang, A new maximum likelihood algorithm for piecewise regression, Journal of the American Statistical Association 76 (1981) 980–987. doi:10.1080/01621459.1981.10477752.
[43] B. P. Carlin, A. E. Gefland, A. F. M. Smith, Hierarchical Bayesian analysis of changepoint problems, Applied Statistics 41 (1992) 389–405. doi:10.2307/2347570.
[44] P. E. Ferreira, A Bayesian analysis of a switching regression model: Known number of regimes, Journal of the American Statistical Association 70 (1975) 370–374. doi:10.1080/01621459.1975.10479875.
[45] V. Shutko, L. Tereshchenko, M. Shutko, I. Silantieva, O. Kolganova, Application of spline-fourier transform for radar signal processing, in: Proceedings of IEEE 15th International Conference on the Experience of Designing and Application of CAD Systems (CADSM), 2019, pp. 110–113. doi:10.1109/CADSM.2019.8779279.
[46] J. D. Toms, M. L. Lesperance, Piecewise regression: A tool for identifying ecological thresholds, Ecology 84 (2003) 2034–2041. doi:10.1890/02-0472.
[47] I. Prokopenko, I. Omelchuk, M. Maloyed, Synthesis of signal detection algorithms under

conditions of aprioristic uncertainty, in: Proceedings of IEEE Ukrainian Microwave Week, 2020, pp. 418–423. doi:10.1109/UkrMW49653.2020.9252687.

[48] G. V. Reklaitis, A. Ravindran, K. M. Ragsdell, Engineering Optimization. Methods and Applications, John Wiley and Sons, New York, 1983.