DOI 10.1007/s10559-020-00314-w
Cybernetics and Systems Analysis, Vol. 56, No. 6, November, 2020

STATISTICAL ANALYSIS OF THE DYNAMICS OF CORONAVIRUS CASES USING STEPWISE SWITCHING REGRESSION

P. S. Knopov[1] **and A. S. Korkhin**[2]

UDC 519.233.5

Abstract. *The dynamics of coronavirus cases is proposed to be modeled using switching regression whose switching points are unknown. The stepwise process of constructing the regression in time is described. The dynamics of the number of coronavirus cases in Ukraine is analyzed.*

Keywords: *regression, switching points, regression parameters, stepwise estimation, infection, coronavirus.*

INTRODUCTION

Switching regression is a set of regression models arranged sequentially in time, which can be both not related or related to each other. Regressions are divided one from another by switching points, which are often unknown. This case is a subject of the study. Noteworthy are the studies by P. Perron with co-authors (see, for example, [1–3]). They propose to use the Bellman and Roth algorithm [4] for estimation of switching points by the dynamic programming method. Developments by Perron and his co-authors were used to solve economic problems. The authors of the present paper also propose a number of results in generating switching regressions. The studies [5, 6] describe the methods of estimating switching points based on the given sampling, which allow taking into account the constraints imposed on these switching points and regression parameters that follow from the a priori information about the process being modeled. Such constraints cannot be taken into account when a dynamic programming scheme is used.

In some applications (for example, economy, public health services), the concept of a fixed observation interval, which is used in [1–3, 5, 6], is not always acceptable in view of continuous data renewal. As an example, we will consider the coronavirus infection process, which is of current concern.

In the paper, we propose to analyze the infection dynamics based on the switching regression model and create the model by steps in time. The observation interval, whose length is fixed or increases in time, is divided into a sequence of rather short overlapping intervals I_j, $j = 1, 2, \dots$, which a priori contain a small number of switching points, for example, no more than two. This considerably simplifies the problem of their estimation.

1. METHODOLOGY OF CREATING THE MODEL OF THE TIME SERIES OF CORONAVIRUS CASES

Figure 1 (whose data are taken from [7]) shows a time series of the daily number of coronavirus cases (NCC) in Ukraine since April 12, 2020, when NCC level had an evident growth tendency. As is seen from the figure, the rate of NCC variation is not constant: it varies in time not only in the value but also in the sign. Therefore, it is expedient to

[1] V. M. Glushkov Institute of Cybernetics, National Academy of Sciences of Ukraine, Kyiv, Ukraine, *knopov1@yahoo.com*. [2] Pridneprovskaya Academy of Construction and Architecture, Dnipro, Ukraine, *a.s.korkhin@gmail.com*. Translated from Kibernetika i Sistemnyi Analiz, No. 6, November–December, 2020, pp. 96–106. Original article submitted July 17, 2020.

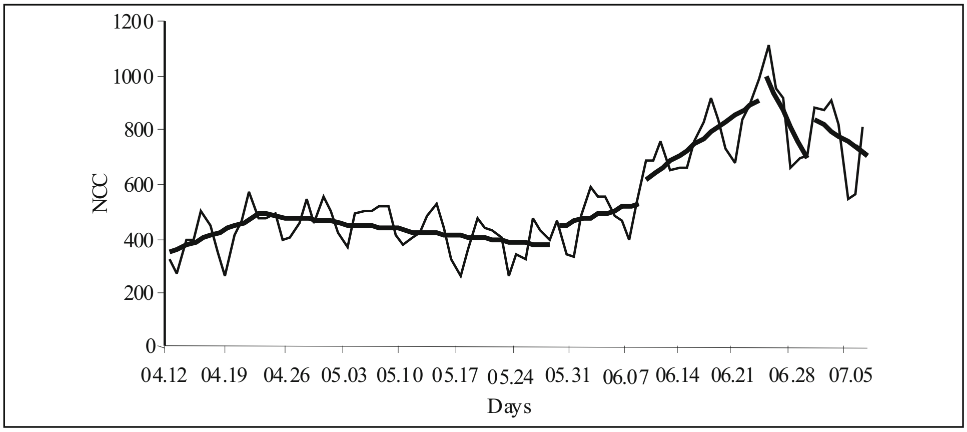

Fig. 1. NCC time series and the line of switching regression approximating it (bold line).

describe NCC variation by a linear switching regression. An analysis of the NCC dynamics has shown that regular oscillations with one week phase are typical for it. This fact should be taken into account when creating a switching regression.

The NCC time series consists of the trend TR, regular oscillations K, and a random component E. The trend is a sequence of straight lines separated one from another by switching points, which not necessarily should be linked up at these points. Regular oscillations are apparently related to labor activity. Random oscillations are caused by a set of minor factors such as coronavirus testing errors and occurrence of random centers of infection.

It is natural to assume the additive structure of NCC time series defined by the model

$$y = TR + K + E\,, \tag{1}$$

where y is the number of cases.

Let us determine the values on the right-hand side of (1). First, let us eliminate the trend from y by using moving average with the averaging interval equal to seven (the number of days in a week):

$$\bar{y}_t = \frac{\sum\limits_{i=t-3}^{t+3} y_i}{7}, \quad t = 4, 5, \dots, \tag{2}$$

where y_i is the NCC at the tth day ($t = 1$ corresponds to 04.12.2020).

According to (1), the difference $u_t = y_i - \bar{y}_t$, $t = 4, 5, \dots$, is the sum $K + E$ of regular oscillations and of a random component. The value of u_t depends on the day of the week to which the number of the day t corresponds. It can be presented by the sum $u_t = K^0_{i(t)} + E_t$, where $i(t) = i$ is the number of the day of the week corresponding to the number of observation (days); for definiteness, we assume that $i = 1$ corresponds to Monday; $K^0_{i(t)}$ is the true unknown value of the regular deviation of NCC from the trend for the ith day of the week; E_t is the value of the random component of the original time series.

To find the estimate K^0_i, let us average all the values u_t corresponding to the ith day of the week:

$$\hat{K}_i = \sum_{t \in \Omega_i} u_t \,/\, |\Omega_i|, \quad i = 1, \dots, 7, \tag{3}$$

where Ω_i is the set of indices of days in the NCC time series to which day i of the week corresponds; $|\Omega_i|$ is the number of elements in Ω_i. According to (2), six observations are lost when $\hat{K}_i$ is calculated using (3): three observations in the beginning and three at the end of the observation interval.

TABLE 1

Cases	**Results of Calculation of $\hat{K}_i$ Corresponding to Days of the Week**						
	Monday $i = 1$	Tuesday $i = 2$	Wednesday $i = 3$	Thursday $i = 4$	Friday $i = 5$	Saturday $i = 6$	Sunday $i = 7$
1	−73.799	34.774	26.348	116.922	46.496	−67.93	−5.373
2	−22.783	15.791	105.365	12.365	13.365	27.365	−125.738
3	−63.635	−8.635	75.365	−9.635	85.365	37.365	−72.635
4	−98.635	22.365	42.365	39.365	50.365	57.365	−46.635
5	−89.635	−62.635	0	0	0	0	−46.635
$\hat{K}_i^-$	−69.698	0.332	62.361	39.754	48.897	13.541	−59.404
$\hat{K}_i$	−74.81	−4.78	57.249	34.642	43.786	8.429	−64.515

From (3), we get $\hat{K}_i = K_i^0 + \sum_{t \in \Omega_i} E_t / |\Omega_i|$. If E_t, $t = 1, 2, \ldots$, is a sequence of mutually independent equally distributed random variables and $E\{E_t\} = 0$, then as $t \to \infty$ (since $|\Omega_i| \to \infty$) the second term in the formula for $\hat{K}_i$ converges to zero in mean square. Then $\hat{K}_i$ is a consistent estimate of K_i^0, $i = 1, \ldots, 7$.

Due to symmetry of the regular oscillations with respect to the trend, which can be assumed according to Fig. 1, we get $\sum_{i=1}^{7} K_i^0 = 0$. Estimates of regular oscillations for a finite number of observations, defined by (3), will not satisfy this condition. Therefore, it is necessary to modify them to obtain

$$\sum_{i=1}^{7} \hat{K}_i = 0. \tag{4}$$

Table 1 shows the results of calculation of $\hat{K}_i$ for the time interval I_1 = [04.12.20, 05.12.20] with 31 observations.

In the sixth row of Table 1, $\hat{K}_i^-$ are calculated by the formula (3), and in the last row they are modified according to the condition (4). Note that $\sum_{i=1}^{7} \hat{K}_i^- = 35.784$, which is a big number.

The graph of $\hat{K}_i$, $i = 1, \ldots, 7$, is not smooth. However, as the number of observations grows, it varies. For example, for NCC on the interval from 04.12.20 till 07.10.20, containing 90 observations out of which 84 observations are used for calculations by (3), $\sum_{i=1}^{7} \hat{K}_i^- = -\,2.143$, and the graph of these values becomes nearly smooth (Fig. 2). As we can see, the maximum oscillation amplitude is observed in the middle of the week; at the end and in the beginning of the week we have a minimum: a negative number which determines the least number of cases per week; therefore, Fig. 2 reflects the situation where household cases are accumulated with cases related to workplace and public transportation.

After separating regular oscillations, we delete them from the original NCC time series. As a result, according to (1) we get $z = y - K = TR + E$. The levels of time series z are defined by the formula $z_t = y_t - \hat{K}_{i(t)}$.

The technique presented here is similar to the procedure of extracting seasonal fluctuations from a time series described by the multiplicative model (all the components of the series are multiplied) [8, Ch. 15].

Let us describe the obtained sum of the trend and random component on each time interval I_j, $j = 1, 2, \ldots$, on which the estimation is carried out, by a switching regression consisting of $k+1$ linear regressions:

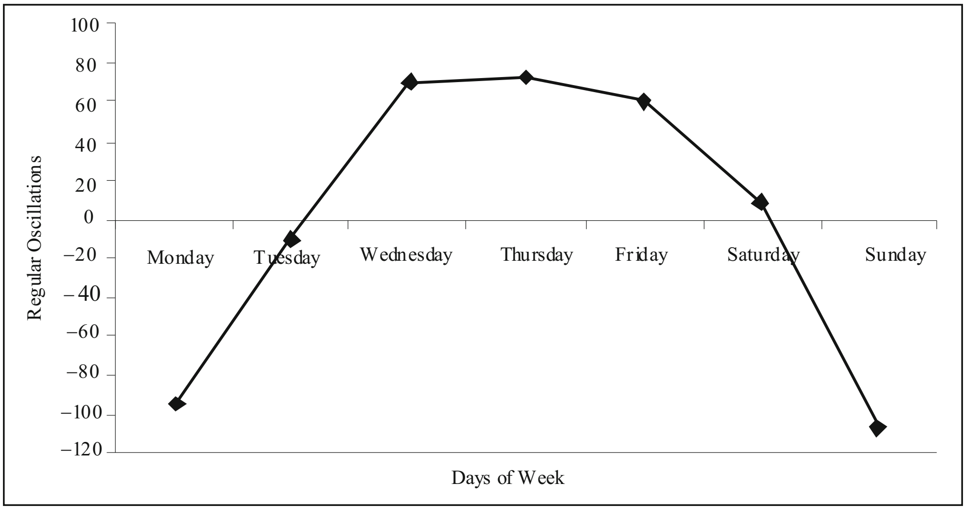

Fig. 2. The graph of regular NCC oscillations based on 84 observations.

$$\left.\begin{array}{l} z_t = \alpha^0_{01} + \alpha^0_{11} t + \varepsilon_{t1},\ t = \tau_{0j}, \tau_{0j}+1, \ldots, t^0_1, \\ z_t = \alpha^0_{02} + \alpha^0_{12} t + \varepsilon_{t2},\ t = t^0_1 + 1, \ldots, t^0_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ z_t = \alpha^0_{0k} + \alpha^0_{1k} t + \varepsilon_{tk},\ t = t^0_{k-1} + 1, \ldots, t^0_k, \\ z_t = \alpha^0_{0,k+1} + \alpha^0_{1,k+1} t + \varepsilon_{t,k+1},\ t = t^0_k + 1, \ldots, \tau_{1j}. \end{array}\right\} \tag{5}$$

Here, $k = k(j)$ is the number of switching points; α^0_{0i} and α^0_{1i}, $i = 1, \ldots, k+1$, are unknown regression parameters; t^0_i, $i = 1, \ldots, k$, are unknown switching points; τ_{0j} and τ_{1j} are respectively the beginning and the end of the interval I_j; and ε_{ti} are random components of regressions that characterize the influence of minor factors on the NCC. The superscript 0 of the above-mentioned quantities specifies that the value of the regression parameter or of the switching point is true. Hereinafter, to simplify the notation, we omit the subscript j of the interval I_j for k, parameters, and switching points of the regression.

Since random components in (5) are subsequences of the sequence E_t, $t = 1, 2, \ldots$, according to its properties presented above they correspond to standard assumptions of regression analysis: do not correlate with each other, have zero expectations, and identical variances σ^2. Let us present the aforesaid as assumptions.

Assumption 1. Random components of regressions (5) have the following characteristics:

expectation

$$E\{\varepsilon_{ti}\} = 0,\ t \in \theta^0_i = \{t^0_{i-1} + 1, \ldots, t^0_i\},\ i = 1, \ldots, k+1; \tag{6}$$

variance

$$E\{\varepsilon^2_{ti}\} = \sigma^2,\ t \in \theta^0_i = \{t^0_{i-1} + 1, \ldots, t^0_i\},\ i = 1, \ldots, k+1; \tag{7}$$

random components of one regression are uncorrelated:

$$E\{\varepsilon_{\tau_1 i}\varepsilon_{\tau_2 i}\} = 0,\ \tau_1, \tau_2 \in \theta^0_i = \{t^0_{i-1} + 1, \ldots, t^0_i\},\ i = 1, \ldots, k+1; \tag{8}$$

random components of different regressions are also uncorrelated:

$$\begin{gathered} E\{\varepsilon_{\tau_1 i_1}\varepsilon_{\tau_2 i_2}\} = 0,\ \tau_1 \in \theta^0_{i_1} = \{t^0_{i_1-1} + 1, \ldots, t^0_{i_1}\}, \\ \tau_2 \in \theta^0_{i_2} = \{t^0_{i_2-1} + 1, \ldots, t^0_{i_2}\},\ i_1,\ i_2 = 1, \ldots, k+1. \end{gathered} \tag{9}$$

Here, θ^0_i is the time interval on which the regression parameters α^0_{0i} and α^0_{1i} are constant; for the jth interval $t^0_0 = \tau_{0j}$ and $t^0_{k+1} = \tau_{1j}$.

Assumption 2. Random components of the regressions (5) are normally distributed.

As is seen from Fig. 1, a substantial difference between NCC in two adjacent days is possible. Therefore, intervals of the straight lines $\alpha_{0i}^{0} + \alpha_{1i}^{0} t$, $i = 1, \dots, k+1$, should not necessarily form a continuous piecewise linear function t, i.e., this function can be discontinuous at switching points.

Properties of the random components of the regression (6)–(9) stipulate the problem of estimation of switching points and parameters of switching regression on each estimation interval I_j, $j = 1, 2, \dots$:

$$S = \sum_{i=1}^{k+1} \sum_{t=t_{i-1}+1}^{t_i} (z_t - \alpha_{0i} - \alpha_{1i} t)^2 \to \min, \tag{10}$$

$$\tau_{0j} \le t_i \le \tau_{1j},\ t_i - t_{i-1} \ge 2,\ i = 1, \dots, k+1,\ t_0 = \tau_{0j},\ t_{k+1} = \tau_{1j}. \tag{11}$$

Minimization in the problem (10), (11) is carried out with respect to the parameters α_{0i} and α_{1i}, $i = 1, \dots, k+1$, continuous quantities and integer switching points t_i, $i = 1, \dots, k$. Since these quantities vary, their superscript 0 is omitted. To variable switching points in (10), (11) there correspond intervals θ_i of constancy of parameters with variable ends:

$$\theta_i = \{t_{i-1} + 1, \dots, t_i\},\ i = 1, \dots, k+1. \tag{12}$$

The constraint (11) establishes the minimum number of observations: two for estimation of two parameters of each of the $(k+1)$th straight line on the intervals θ_i, $i = 1, \dots, k+1$, of constancy of the switching regression parameters.

The problem (10), (11) can be considered as a generalization of the problem of parameter estimation of a constrained nonlinear regression [9, Ch. 1, 10–13]. Since it contains unknown integers, it was solved by a special method [5]. Below, we present the process and results of the solution for overlapping estimation intervals I_1, I_2, I_3: $I_1 \cap I_2 \ne \varnothing$, $I_1 \cap I_3 = \varnothing$, $I_2 \cap I_3 \ne \varnothing$. The values of $\hat{K}_i$, $i = 1, \dots, 7$, were found from the NCC at the end of the respective interval. For example, when constructing a switching regression on the interval I_1, these values were calculated based on observations over this interval: from 04.12.20 till 05.12.20. When constructing regression on the interval I_2, we used observations beginning with 04.12.20 till the end of this interval 06.21.20. Such approach has allowed, in particular, simulating step-by-step inflow of NCC data, which actually took place. Note that other approaches, based on discrete optimization methods, can also be proposed for solution of the considered problem [14, 15].

2. CONSTRUCTING THE SWITCHING REGRESSION ON THE INTERVAL I_1

According to Fig. 3, on the interval I_1 = [04.12.20, 05.12.20], containing 31 observations, there are no more than one switching point at which NCC can vary (decrease or increase). Adding one switching point, where velocity probably varies, we obtain $k = 2$ in the estimation problem (10), (11). Its solution is estimates of switching points: $\hat{t}_1 = 4$ and $\hat{t}_2 = 19$. Estimate of the length of the first interval of constancy of regression parameters θ_1^0 equals four. It is too small, which allows us to advance the null hypothesis H_0: $\alpha_{01}^0 = \alpha_{02}^0$; $\alpha_{11}^0 = \alpha_{12}^0$ (the first and second regressions in (5) coincide). Assuming that the second switching point is fixed, to test the formulated hypothesis on the basis of Assumptions 1 and 2, we will use the criterion from [16]. According to it, we find S, the squared sum of deviations of the linear regression with n parameters of observations on the time interval of length T; S_1 and S_2 are the sums of squared deviations of two other linear regressions: the first and the second ones with n parameters each of the same observations on the time intervals of length, respectively, T_1 and T_2, and $T_1 + T_2 = T$. Let us calculate the statistics

$$F^* = \frac{S - (S_1 + S_2)}{(S_1 + S_2)} \frac{T - 2n}{n}. \tag{13}$$

Hypothesis H_0 is rejected if $F^* > F_p(n, T-2n)$, where $F_p(n, T-2n)$ is the $100p\%$ -point of F-distribution with the degrees of freedom n and $T - 2n$. In this case, $T = 19$, $T_1 = 4$, $T_2 = 15$, $n = 2$, $S = 34936.83$, and $S_1 + S_2 = 31107.86$.

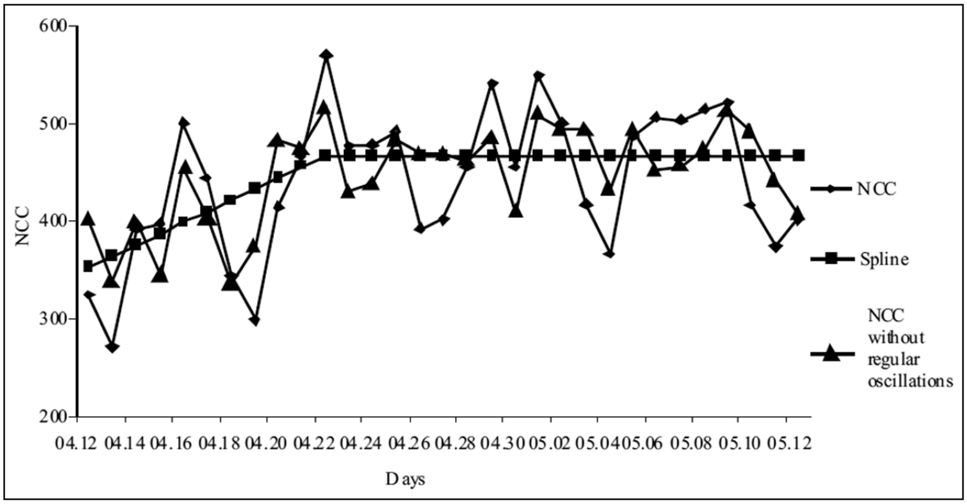

Fig. 3. Creating a switching regression on the interval I_1.

We get $F^{*}=0.925$, $F_{0.05}(2,15)=3.682$. Thus, at the 5%-level, the hypothesis about coincidence of the first and second regressions is not rejected. Therefore, the hypothesis that there is one switching point on the interval [1, 19] is not rejected. Let us determine it by solving the problem of estimating (10), (11) for $k=1$, $j=1$, $\tau_{01}=1$, and $\tau_{11}=31$. The date 04.22.20 corresponds to the obtained estimate of the unique switching point $\hat{t}_1=11<19$. The sum of squared residues in (10) makes $S=$ 42251.76, and at point $t=\hat{t}_1$ the straight lines $\hat{\alpha}_{01}+\hat{\alpha}_{11}t$ and $\hat{\alpha}_{02}+\hat{\alpha}_{12}t$, where $\hat{\alpha}_{kl}$ is the estimate α^{0}_{kl}, $k=0,1$; $l=1,2$, are pairwise close. Therefore, the constraint was added to the problem (10), (11):

$$\alpha_{01}+\alpha_{11}t_1=\alpha_{02}+\alpha_{12}t_1. \tag{14}$$

As a result, we obtained a switching regression with the regression line being a linear spline (see Fig. 3). The sum of squared deviations (10) for this case $S=42559.85$ slightly increased after adding the constraint to the estimation problem. This testifies that above-mentioned discrepancy of two straight lines in $\hat{t}_1$ was due to random factors.

Approximate accuracy analysis of the estimates of parameters of the obtained spline was based on the fact that the obtained switching point coincides with the true one. Such assumption is not a severe constraint for the problem under study since time, rather than some complex function, is an independent variable. Then according to [8, Ch. 15] we will consider the obtained two regressions as one, whose variable parameters are combined by Eq. (14), where $t_1=\hat{t}_1$ is fixed. Then the covariance matrix of regression parameter estimates is defined by the expression

$$\mathbf{V}=\sigma^2(\mathbf{X}'\mathbf{X})^{-1}[\mathbf{J}_{2(k+1)}-\mathbf{G}'\mathbf{S}\mathbf{G}(\mathbf{X}'\mathbf{X})^{-1}], \tag{15}$$

where the prime denotes transposition, $\mathbf{J}_{2(k+1)}$ is a unit matrix of order $2(k+1)$, and $\mathbf{S}=[\mathbf{G}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{G}']$.

In the case under study, $\mathbf{X}=\operatorname{diag}(\mathbf{X}_1,\mathbf{X}_2)$ and $\mathbf{G}=[1\ \hat{t}_1\ -1\ -\hat{t}_1]$, and

$$\mathbf{X}_1=\begin{bmatrix}1 & 1\\ 1 & 2\\ \vdots & \vdots\\ 1 & \hat{t}_1\end{bmatrix},\ \mathbf{X}_2=\begin{bmatrix}1 & \hat{t}_1+1\\ 1 & \hat{t}_1+2\\ \vdots & \vdots\\ 1 & \tau_{11}\end{bmatrix},\ \tau_{11}=31. \tag{16}$$

Generally, the quantities in (15) for the jth estimation interval have the form

$$\mathbf{X} = \operatorname{diag}\,(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{k+1}),\ \ \mathbf{X}_1 = \begin{bmatrix} 1 & \tau_{0j} \\ 1 & \tau_{0j}+1 \\ \vdots & \vdots \\ 1 & \hat{t}_1 \end{bmatrix},\ \ \mathbf{X}_{k+1} = \begin{bmatrix} 1 & \hat{t}_k+1 \\ 1 & \hat{t}_k+2 \\ \vdots & \vdots \\ 1 & \tau_{1j} \end{bmatrix},$$

$$\mathbf{X}_l = \begin{bmatrix} 1 & \hat{t}_{l-1}+1 \\ 1 & \hat{t}_{l-1}+2 \\ \vdots & \vdots \\ 1 & \hat{t}_l \end{bmatrix},\ \ l = 2, \dots, k;\ \ \mathbf{G} = \begin{bmatrix} 1 & \hat{t}_1 & -1 & -\hat{t}_1 & 0 & 0 & \mathbf{O}_{1m} \\ 0 & 0 & 1 & \hat{t}_2 & -1 & -\hat{t}_2 & \mathbf{O}_{1m} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & \hat{t}_k & -1 & -\hat{t}_k \end{bmatrix}, \qquad (17)$$

where $\mathbf{O}_{1m}$ is the row with m columns, $m = 2(k+1) - kn$. For $m = 0$, $\mathbf{O}_{10}$ means that the row is absent.

If straight lines do not mate at some points, then the corresponding rows of the matrix $\mathbf{G}$ are deleted. In case of zero matrix $\mathbf{G}$, which corresponds to the absence of constraints on regression parameters, from (15) we get $\mathbf{V} = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$.

The significance of the estimates of regression parameters for the fixed switching points was determined on the basis of (15) and Assumptions 1 and 2 by well-known methods. Noncorrelatedness analysis of random components of the regression was carried out by the Darbin–Watson criterion, and analysis of their normality by the criterion based on asymmetry and kurtosis coefficients [17, Sec. 3.2], for its application see [18, Sec. 9.3].

For the case $k = 1$, according to the formula (15), estimate of the covariance matrix $\hat{\mathbf{V}}$ was found by replacing σ^2 with its estimate $s^2 = 1464$.

Estimates of spline parameters turned out to be significant at the 5%-level, except for the estimate $\hat{\alpha}_{12} = -1.374$. Therefore, the estimation problem (10), (11) with the added constraints (14) and $\alpha_{12} = 0$ was solved. As a result, a spline that describes smooth passage from NCC growth to a plateau, a horizontal section (see Fig. 3) with highly significant nonzero estimates of its parameters, was obtained: $\hat{\alpha}_{01} = 340.69$; $\hat{\alpha}_{11} = 11.56$; and $\hat{\alpha}_{02} = 467.81$. The estimate of the covariance matrix $\mathbf{V}$ was calculated according to the initial data (16), where matrix $\mathbf{X}_2$ is a column of ones.

3. CONSTRUCTING THE SWITCHING REGRESSION ON THE INTERVAL I_2

The beginning of the interval $I_2 = [04.23.20, 06.21.20]$ coincides with the beginning of the plateau, the switching point defined on I_1, plus 1. Then leaving the plateau towards increase or decrease of the NCC is possible. A repeated change in the direction of the NCC dynamics is possible. Therefore (as well as in Sec. 2), we will suppose in the estimation problem that $k = 2$. Let us establish $\tau_{02} = 1$, $\tau_{12} = 60$. Thus, time reference in I_2 begins with one. Solving the estimation problem for I_2, we obtain estimates of two switching points: $\hat{t}_1 = 37$ (05.29.20) and $\hat{t}_2 = 48$ (06.09.20), as well as estimates of parameters of three regressions that form a switching regression:

$$\begin{array}{cccc} \hat{\alpha}_{01} = 493.65; & \hat{\alpha}_{11} = -3.335; & \hat{\alpha}_{02} = 121.71; & \hat{\alpha}_{12} = 8.498; \\ (0) & (0) & (0.471) & (0.033) \end{array} \qquad (18)$$

$$\begin{array}{cc} \hat{\alpha}_{03} = -102.3; & \hat{\alpha}_{13} = 15.275 \\ (0.585) & (0) \end{array}$$

(digits in brackets mean the significance of the parameter, which was determined in the same way as in Sec. 2).

According to (18), at point $t = 37$ the plateau has ended. It began at $t = 1$ and has a small declination with a significant angular coefficient whose estimate −3.335. According to Sec. 2, its value is insignificant and the estimate is 1.374; it is possible to explain such a discrepancy by a great amount of data on the plateau on the interval I_2.

Straight lines of the second and third regressions, according to (18), have angular coefficients of identical signs, and the angular coefficient of the second straight line is insignificant at the 1%-level. Therefore, the null hypothesis about equality of parameters of the specified regressions was advanced and was tested with the help of statistics (13), where $T = 23$ and $n = 2$. It was obtained that $F^* = 4.97$. Since $F_{0.05}(2, 19) = 3.52$, the null hypothesis was rejected.

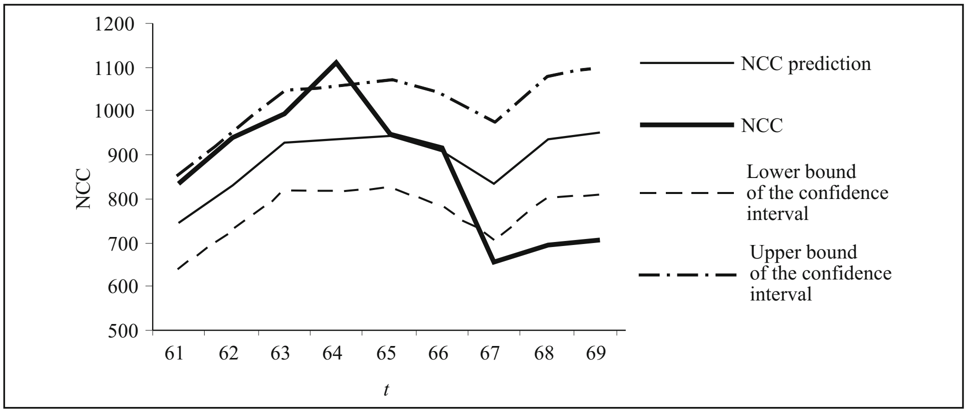

Fig. 4. NCC prediction based on the results of estimation on I_2.

Let us now predict the NCC based on the obtained switching regression on the considered interval, which means extrapolation of the straight line of the third regression. From here, we get the pointwise prediction of the NCC

$$\hat{y}_t = \hat{\alpha}_{03} + \hat{\alpha}_{13} t + \hat{K}_{i(t)} = \mathbf{w}'_t \hat{\boldsymbol{\alpha}}_3 + \hat{K}_{i(t)}, \;\; \hat{\boldsymbol{\alpha}}_3 = [\hat{\alpha}_{03} \; \hat{\alpha}_{13}]', \;\; \mathbf{w}_t = [1 \;\; t]', \;\; t > T, \tag{19}$$

where the prime denotes transposition.

Due to Assumptions 1 and 2, we get the interval prediction

$$\hat{y}_t - u_p(q)\hat{\sigma}_f(t) \le y_t \le \hat{y}_t + u_p(q)\hat{\sigma}_f(t), \;\; t > T, \tag{20}$$

where $u_p(q)$ is the $100p$%-point of the Student distribution with the number of degrees of freedom $q = T - (k+1)n$; $\hat{\sigma}_f(t)$ is the estimate of the mean square deviation of the prediction. We will find it based on (1); as a result, we get $y_t = \mathbf{w}'_t \boldsymbol{\alpha}_3^0 + K_{i(t)}^0 + E_t$, $\boldsymbol{\alpha}_3^0 = [\alpha_{03}^0 \;\; \alpha_{13}^0]'$, $t > T$. This equality and (19) yield the prediction error: $\hat{y}_t - y_t = \mathbf{w}'_t(\hat{\boldsymbol{\alpha}}_3 - \boldsymbol{\alpha}_3^0) + (\hat{K}_{i(t)} - K_{i(t)}^0) - E_t$. From here, since $\hat{\boldsymbol{\alpha}}_3$ and $\hat{K}_{i(t)}$ do not depend on E_t, neglecting the dependence of $\hat{K}_{i(t)}$ on $\hat{\boldsymbol{\alpha}}_3$, we obtain the variance of the NCC prediction $\sigma_f^2(t) = \mathbf{w}'_t \mathbf{K}_3 \mathbf{w}_t + \sigma^2(\hat{K}_{i(t)}) + \sigma^2$, where $\mathbf{K}_3$ is a covariance matrix $\hat{\boldsymbol{\alpha}}_3$; $\sigma^2(\hat{K}_{i(t)})$ is the variance $\hat{K}_{i(t)}$. Replacing in it all the quantities with their estimates, we obtain

$$\hat{\sigma}_f^2(t) = \mathbf{w}'_t \hat{\mathbf{K}}_3 \mathbf{w}_t + \hat{\sigma}^2(\hat{K}_{i(t)}) + s^2. \tag{21}$$

Figure 4 shows the results of the prediction for $t = 61, \ldots, 69$ (06.22.20–06.30.20) calculated based on (19). The boundaries of the confidence interval for the prediction are defined by means of (20) and (21) for $p = 0.05$, $q = 60 - 6 = 54$. According to the figure, the true NCC value does not get into the confidence interval at points t equal to 64, 67, 68, and 69, which can testify about the ocurrence of one or two new switching points in the neighborhood of these time intervals and hence, necessity to pass to a new estimation interval for I_3.

Average prediction error on the time interval [61, 66], where there was only one NCC value that did not get into the confidence interval, was determined as the ratio of the estimate of the root mean square deviation of the prediction to the average true value of the NCC on the interval $[61, 66]$. It was 9.9%. As calculations have shown, the prediction error can be reduced to 7.4% if we make the estimate $\hat{\alpha}_{03}$ significant, i.e., increase its accuracy. This can be attained by jointing the second and third regressions at the point $t = 48$.

For comparison, average relative error of the prediction for eight days for the model obtained on the interval I_1 was 19%. Such a large error can be explained by a low accuracy of estimation of regular oscillations of NCC on the small number of data equal to 31.

4. CONSTRUCTING THE SWITCHING REGRESSION ON THE INTERVAL I_3

On the interval $I_3 = [06.10.20, 07.07.20]$, days are indexed as a continuation of indexing on the previous interval. Therefore, $\tau_{03} = 49$: the beginning of I_3 is at the last switching point on the interval I_2 plus 1. The end of the interval $\tau_{13} = 76$. For the case of possible decrease or increase in the NCC revealed in Sec. 3, we suppose $k = 2$.

The solution of the estimation problem (10), (11) determines the estimates of switching points $\hat{t}_1 = 63$ (06.24.20) and $\hat{t}_2 = 69$ (06.30.20) and estimates of regression parameters:

$$\underset{(0.034)}{\hat{\alpha}_{01} = -411.77}; \; \underset{(0)}{\hat{\alpha}_{11} = 21.11}; \; \underset{(0)}{\hat{\alpha}_{02} = 4810.11}; \; \underset{(0)}{\hat{\alpha}_{12} = -59.55}; \qquad (22)$$

$$\underset{(0)}{\hat{\alpha}_{03} = 2277.1}; \; \underset{(0.059)}{\hat{\alpha}_{13} = -20.55},$$

where the notation in brackets have the same sense as in (18).

According to (22), all the angular coefficients, except for $\hat{\alpha}_{13}$, are highly significant, at the point $\hat{t}_1$ sharp decrease has begun, followed by slow decrease at $\hat{t}_2$. In view of the proximity of these points, the hypothesis about equality of parameters of the second and third regressions was tested, assuming that $\hat{t}_1 = 63$ is approximately known since sharp decrease of the NCC has begun at its neighborhood. According to the criterion (13), this assumption was rejected at the 5%-level: $F^* = 5.36$, $F_{0.05}(2,9) = 4.26$. Sharp decrease in the NCC can be due to liquidation of one or several centers of infection. Insignificance of $\hat{\alpha}_{13}$ can be explained by the fact that infection dynamics probably has gone out to the next plateau for a short while.

CONCLUSIONS

In the paper, we have considered a model of the dynamics of coronavirus cases in the form of a switching regression whose switching points are unknown.

We have proposed stepwise solution of the problem of its creation. First, at each step, we solved the estimation problem for two switching points based on observations on some small time interval I_j, $j = 1, 2, 3$. Then we carried out statistical analysis of the constructed part of regression, which considerably simplified the estimation problem. As a result, the required regression was determined on three sequential time intervals: $I_1 \setminus I_1 \cap I_2$, $I_2 \setminus I_2 \cap I_3$, and I_3. Completely, the line of this regression was obtained in the form of a piecewise linear function of time since the ends of these intervals coincide with the last switching points of the previous interval plus one or with the ends of the observation interval (see Fig. 1).

The use of prediction allows obtaining a satisfactory approximation to NCC if the process is between two switching points, and determining if it hits the domain where a new switching point can be. The latter is important in making a decision to step up (ease) the quarantine and related measures.

As is shown in the paper, without a priori information about the number of switching points, five such points are found and the pattern of NCC variation (growth, decay, stabilization) to the right from these points based on a small number of observations available in this domain is determined. Thus, switching regression allows not only obtaining a short-term prediction of the dynamics of the epidemy, but also promptly determining the trend of its development.

Note that even in case of 87 observations at the very beginning of the research, finding all the switching points simultaneously when their number is not known would be a challenge. The idea of stepwise estimation has simplified the solution.

REFERENCES

1. J. Bai and P. Perron, “Estimating and testing linear models with multiple structural changes,” Econometrica, Vol. 66, No. 1, 47–78 (1998).
2. J. Bai and P. Perron, “Computation and analysis of multiple structural change models,” J. of Applied Econometrics, Vol. 18, Iss. 1, 1–22 (2003).

3. A. Casini and P. Perron, Structural Breaks in Time Series, Preprint, Boston University (2018). URL: https://arxiv.org/abs/1805.03807.
4. R. Bellman and R. Roth, “Curve fitting by segmented straight lines,” J. of the American Statistical Association, Vol. 64, 1079–1084 (1969).
5. A. S. Korkhin, “Constructing a switching regression with unknown switching points,” Cybern. Syst. Analysis, Vol. 54, No. 3, 443–455 (2018). https://doi.org/10.1007/s10559-018-0045-9.
6. P. S. Knopov and A. S. Korkhin, “Continuous-time switching regression method with unknown switching points,” Cybern. Syst. Analysis, Vol. 56, No. 1, 68–80 (2020). https://doi.org/10.1007/s10559-020-00222-z.
7. Coronavirus Infection (COVID-19), Statistics in Ukraine. © Google LLC (2020).
8. A. S. Korkhin and E. P. Minakova, Computer Statistics, Pt. 2 [in Russian], Nats. Girnychyi Universytet, Dnipro (2009).
9. P. S. Knopov and A. S. Korkhin, “Estimation of regression model parameters with specific constraints,” in: Regression Analysis under a Priori Parameter Restrictions, Ch. 1, Springer, New York (2012), pp. 1–28.
10. A. N. Golodnikov, P. S. Knopov, and V. A. Pepelyaev, “Estimation of reliability parameters under incomplete primary information,” Theory and Decision, Vol. 57, No. 4, 331–344 (2004).
11. V. S. Mikhalevich, P. S. Knopov, and A. N. Golodnikov, “Mathematical models and methods of riks assessment in ecologically hazardous industries,” Cybern. Syst. Analysis, Vol. 30, No. 2, 259–273 (1994).
12. Yu. M. Ermoliev and P. S. Knopov, “Method of empirical means in stochastic programming problems,” Cybern. Syst. Analysis, Vol. 42, No. 6, 773–785 (2006).
13. P. S. Knopov and V. A. Pepelyaev, “Nonparametric estimate of almost periodic signals,” Cybern. Syst. Analysis, Vol. 43, No. 3, 362–367 (2007).
14. I. V. Sergienko and V. P. Shylo, “Problems of discrete optimization: Challenges and main approaches to solve them,” Cybern. Syst. Analysis, Vol. 42, No. 4, 465–482 (2006).
15. S. Butenko, P. Pardalos, I. Sergienko, V. Shylo, and P. Stetsyuk, “Estimating the size of correcting codes using extremal graph problems,” in: C. Pearce and E. Hunt (eds.), Optimization, Springer Optimization and Its Applications, Vol. 32, Springer, New York (2009), pp. 227–243.
16. G. C. Chow, “Tests of equality between sets of coefficients in two linear regressions,” Econometrica, Vol. 28, No. 3, 591–605 (1960).
17. D. R. Cox and D. V. Hinkley, Theoretical Statistics, Chapman and Hall/CRC (1978).
18. A. S. Korkhin and E. P. Minakova, Computer Statistics, Pt. 1 [in Russian], Nats. Gornyi Universitet, Dnepropetrovsk (2008).