УДК 519.2

СТАТИСТИЧЕСКОЕ ПРОГНОЗИРОВАНИЕ ДИНАМИКИ ЭПИДЕМИОЛОГИЧЕСКИХ ПОКАЗАТЕЛЕЙ ЗАБОЛЕВАЕМОСТИ COVID-19 В РЕСПУБЛИКЕ БЕЛАРУСЬ

Ю. С. ХАРИН [1), 2)], ***В. А. ВОЛОШКО*** [1), 2)],
О. В. ДЕРНАКОВА [1)], ***В. И. МАЛЮГИН*** [2)], ***А. Ю. ХАРИН*** [1), 2)]

[1)] *Научно-исследовательский институт прикладных проблем математики и информатики БГУ, пр. Независимости, 4, 220030, г. Минск, Беларусь*
[2)] *Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*

Рассматривается актуальная задача статистического прогнозирования динамики основных эпидемиологических показателей пандемии COVID-19 в Республике Беларусь на базе наблюдаемых временных рядов. Для решения этой задачи предлагаются пять методов: метод прогнозирования на основе «скользящих трендов»; локально-медианный метод прогнозирования; метод прогнозирования на основе дискретных временных рядов; метод прогнозирования на основе векторной эконометрической модели коррекции ошибок; метод последовательного статистического анализа. Разработаны алгоритмы вычисления точечных и интервальных прогнозов для ключевых эпидемиологических показателей. Представлены численные результаты компьютерного прогнозирования на примере Республики Беларусь.

Ключевые слова: прогнозирование; вероятностная модель; временной ряд; точечный прогноз; интервальный прогноз; COVID-19.

Благодарность. Исследование выполнено при финансовой поддержке Министерства образования Республики Беларусь. Авторы выражают благодарность кандидату физико-математических наук С. Н. Сталевской за разработку программы к разделу «Метод прогнозирования на основе “скользящих трендов”» данной статьи.

Образец цитирования:
Харин ЮС, Волошко ВА, Дернакова ОВ, Малюгин ВИ, Харин АЮ. Статистическое прогнозирование динамики эпидемиологических показателей заболеваемости COVID-19 в Республике Беларусь. *Журнал Белорусского государственного университета. Математика. Информатика.* 2020; 3:36–50.
https://doi.org/10.33581/2520-6508-2020-3-36-50

For citation:
Kharin YuS, Valoshka VA, Dernakova OV, Malugin VI, Kharin AYu. Statistical forecasting of the dynamics of epidemiological indicators for COVID-19 incidence in the Republic of Belarus. *Journal of the Belarusian State University. Mathematics and Informatics.* 2020;3:36–50. Russian.
https://doi.org/10.33581/2520-6508-2020-3-36-50

Авторы:
Юрий Семенович Харин – член-корреспондент НАН Беларуси, доктор физико-математических наук, профессор; директор[1)], профессор кафедры математического моделирования и анализа данных факультета прикладной математики и информатики[2)].
Валерий Анатольевич Волошко – кандидат физико-математических наук; старший научный сотрудник лаборатории математических методов защиты информации[1)], доцент кафедры математического моделирования и анализа данных факультета прикладной математики и информатики[2)].
Оксана Владимировна Дернакова – младший научный сотрудник сектора компьютерного анализа данных лаборатории прикладной информатики.
Владимир Ильич Малюгин – кандидат физико-математических наук; доцент кафедры математического моделирования и анализа данных факультета прикладной математики и информатики.
Алексей Юрьевич Харин – доктор физико-математических наук, доцент; заведующий кафедрой теории вероятностей и математической статистики факультета прикладной математики и информатики[2)], главный научный сотрудник лаборатории статистического анализа и моделирования[1)].

Authors:
Yuriy S. Kharin, corresponding member of the National Academy of Sciences of Belarus, doctor of science (physics and mathematics), full professor; director[a] and professor at the department of mathematical modeling and data analysis, faculty of applied mathematics and computer science[b].
kharin@bsu.by
Valery A. Valoshka, PhD (physics and mathematics); senior researcher at the laboratory of mathematical methods of information security[a] and associate professor at the department of mathematical modeling and data analysis, faculty of applied mathematics and computer science[b].
valoshka@bsu.by
Oksana V. Dernakova, junior researcher at the sector of computer data analysis, laboratory of applied informatics.
dernakova@bsu.by
Vladimir I. Malugin, PhD (physics and mathematics); associate professor at the department of mathematical modeling and data analysis, faculty of applied mathematics and computer science.
malugin@bsu.by
Alexey Yu. Kharin, doctor of science (physics and mathematics), docent; head of the department of probability theory and mathematical statistics, faculty of applied mathematics and computer science[b], and chief researcher at the laboratory of statistical analysis and modeling[a].
kharinay@bsu.by

STATISTICAL FORECASTING OF THE DYNAMICS OF EPIDEMIOLOGICAL INDICATORS FOR COVID-19 INCIDENCE IN THE REPUBLIC OF BELARUS

Yu. S. KHARIN [a, b], ***V. A. VALOSHKA*** [a, b],
O. V. DERNAKOVA [a], ***V. I. MALUGIN*** [b], ***A. Yu. KHARIN*** [a, b]

[a] *Research Institute for Applied Problems of Mathematics and Informatics, Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*
[b] *Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*
Corresponding author: Yu. S. Kharin (kharin@bsu.by)

The paper is devoted to the urgent problem of statistical forecasting for the dynamics of the main epidemiological indicators for the COVID-19 pandemic in the Republic of Belarus based on the observed time series. To solve this problem, five methods are proposed: forecasting method based on «moving trends»; local-median forecasting method; forecasting method based on discrete time series; forecasting method based on the vector econometric error correction model; method of sequential statistical analysis. Algorithms for computation of point and interval forecasts for the main epidemiological indicators have been developed. The numerical results of computer forecasting are presented on the example of the Republic of Belarus.

Keywords: forecasting; probability model; time series; point forecast; interval forecast; COVID-19.

Acknowledgements. This research is supported by the Ministry of Education of the Republic of Belarus. The authors are grateful to S. N. Staleuskaya, PhD (physics and mathematics), for development of the computer program for numeric results in section «Forecasting method based on "moving trends"» of this paper.

Введение

На данный момент имеется обширный класс математических моделей, разработанных для прогнозирования распространения заболеваний, смертности и выздоровления. Их можно классифицировать следующим образом: искусственные нейронные сети и машинное обучение; пространственно-временные авторегрессионные эпидемиологические модели; аналитические и смешанные пространственно-временные модели (SIR-модели и их аналоги); дискретные условно авторегрессионные временные ряды с длинной памятью; марковские процессы; прогнозирование на основе имитационного моделирования. Обзор существующих подходов представлен в работе [1]. Актуальной проблемой человечества является преодоление пандемий, включая COVID-19 в настоящее время [2; 3].

Прогнозирование эпидемиологических показателей в условиях пандемии характеризуется следующими особенностями: 1) стохастичностью процессов; 2) нестационарностью процессов; 3) существенной зависимостью от состояния системы здравоохранения и применяемых стратегий лечения; 4) оперативностью получения и представления прогнозов. С учетом этих особенностей в данной статье используются вероятностно-статистические модели динамики эпидемиологических процессов и методы робастного статистического прогнозирования [4–9].

Регистрируемые статистические данные и постановка задач прогнозирования динамики эпидемиологических показателей

Введем следующие обозначения: t – дискретное время (номер дня (точнее – суток), считая от некоторого начального дня); T – длительность наблюдения до начала прогнозирования; τ – длина интервала (горизонта) прогнозирования; $\mathfrak{T}_T = \{T+1, T+2, \ldots, T+\tau\}$ – временной промежуток (горизонт прогнозирования на основе T единиц наблюдения); ξ_t – зарегистрированное к моменту t число человек (с начала вспышки) с положительным анализом тестов на COVID-19 (количество выявленных зараженных COVID-19), $\xi_t \in N$; η_t – зарегистрированное к моменту t число пациентов (с начала вспышки), выписанных с отрицательным анализом тестов, у которых ранее был подтвержден диагноз «COVID-19», $\eta_t \in N$; ζ_t – зарегистрированное с начала вспышки количество умерших пациентов с выявленной инфекцией COVID-19, $\zeta_t \in N$; $\mu_t = \xi_t - (\eta_t + \zeta_t)$ – зарегистрированное число зараженных на момент t; $\nu_t = \xi_t - \xi_{t-1}$ – зарегистрированное число новых зараженных в день t $(\xi_0 ::= 0)$, $\nu_t \in N_0 = N \cup \{0\}$; $\alpha_t = \dfrac{\nu_t}{\eta_t - \eta_{t-1}}$ – отношение числа вновь инфицированных к числу выписанных в день t (показатель

нагрузки лечебных учреждений); $\beta_t = \frac{\mu_t}{\eta_t}$ – отношение числа зараженных на момент t к числу всех выздоровевших.

Зарегистрированные к моменту прогнозирования T статистические данные представляют собой временные ряды:

$$\Xi_1^T = \{\xi_1, \xi_2, \ldots, \xi_T\},\ \mathrm{Z}_1^T = \{\zeta_1, \zeta_2, \ldots, \zeta_T\},\ \mathrm{H}_1^T = \{\eta_1, \eta_2, \ldots, \eta_T\},\ \mathrm{N}_1^T = \{\nu_1, \nu_2, \ldots, \nu_T\}. \quad (1)$$

По построению среди всех указанных временных рядов функционально независимы только три ряда: Ξ_1^T, H_1^T, Z_1^T, а временные ряды N_1^T, $\{\mu_t\}$, $\{\alpha_t\}$, $\{\beta_t\}$ функционально выражаются через них. Отметим, что ξ_t, η_t, ζ_t представляют собой накопленные с начального момента времени величины, поэтому временные ряды Ξ_1^T, H_1^T, Z_1^T являются неубывающими относительно времени[1].

Задача заключается в построении точечных и интервальных прогнозов для временных рядов (1) и связанных с ними на 1, 2, …, τ шагов вперед следующего вида.

1. Точечные прогнозы:

$$\Xi_{T+1}^{T+\tau} = \{\hat{\xi}_{T+1}, \ldots, \hat{\xi}_{T+\tau}\},\ \mathrm{Z}_{T+1}^{T+\tau} = \{\hat{\zeta}_{T+1}, \ldots, \hat{\zeta}_{T+\tau}\},$$

$$\mathrm{H}_{T+1}^{T+\tau} = \{\hat{\eta}_{T+1}, \ldots, \hat{\eta}_{T+\tau}\},\ \mathrm{N}_{T+1}^{T+\tau} = \{\hat{\nu}_{T+1}, \ldots, \hat{\nu}_{T+\tau}\}.$$

2. Интервальные прогнозы:

$$\xi_{T+1} \in (\xi_{T+1}^-, \xi_{T+1}^+), \ldots, \xi_{T+\tau} \in (\xi_{T+\tau}^-, \xi_{T+\tau}^+),\ \eta_{T+1} \in (\eta_{T+1}^-, \eta_{T+1}^+), \ldots, \eta_{T+\tau} \in (\eta_{T+\tau}^-, \eta_{T+\tau}^+),$$

$$\zeta_{T+1} \in (\zeta_{T+1}^-, \zeta_{T+1}^+), \ldots, \zeta_{T+\tau} \in (\zeta_{T+\tau}^-, \zeta_{T+\tau}^+),\ \nu_{T+1} \in (\nu_{T+1}^-, \nu_{T+1}^+), \ldots, \nu_{T+\tau} \in (\nu_{T+\tau}^-, \nu_{T+\tau}^+),$$

где знаками –/+ в верхнем индексе помечены нижняя и верхняя границы прогнозного интервала соответственно.

Метод прогнозирования на основе «скользящих трендов»

Известно [1], что распространение инфекций описывается экспоненциальным ростом числа зараженных $O(\rho^t)$, где $\rho \geq 0$ – среднее число человек, заражаемых больным в течение суток (при $\rho > 1$ инфекция экспоненциально растет). В связи с этим процессы ξ_t, η_t, ζ_t на начальном этапе эпидемии целесообразно рассматривать в логарифмической шкале.

Опишем предлагаемый метод прогнозирования на примере задачи прогнозирования временного ряда $\{\xi_t\}$ на начальном этапе эпидемии. Введем в рассмотрение этот ряд в логарифмической шкале:

$$x_t = \ln \xi_t \in R,\ t \in N.$$

В связи с тем что инфекция происходит в управляемом обществе, где правительство предпринимает шаги по подавлению эпидемии, параметр ρ оказывается зависящим от времени: $\rho = \rho(t)$ (данная функция, к сожалению, неизвестна и, по-видимому, сложно поддается статистической оценке). По этой причине модель стохастической зависимости x_t от t меняется с течением времени. Учитывая отмеченные особенности, будем строить математическую модель x_t для «скользящего фрагмента» временного ряда $X_1^T = (x_1, \ldots, x_t) \in R^T$.

Введем обозначения: $2 < s < T$ – длина «скользящего фрагмента»; $X_{t-s+1}^t = (x_{t-s+1}, x_{t-s+2}, \ldots, x_t) \in R^s$ – «скользящий фрагмент» к моменту времени t, $t = s + 1, \ldots, T$. Будем предполагать, что длина s «скользящего фрагмента» достаточно мала и на этом фрагменте справедлива трендовая модель

$$x_i = \theta_0 + \theta_1 i + u_i,\ t - s + 1 \leq i \leq t, \quad (2)$$

являющаяся простой линейной регрессией, где θ_0, $\theta_1 \in R$ – параметры регрессии, θ_1 – угол наклона тренда к оси времени, характеризующий скорость экспоненциального (если $x_t = \ln \xi_t$) или линейного

[1] В настоящем исследовании статистические данные по указанным переменным получены с сайтов Министерства здравоохранения Республики Беларусь (*http://minzdrav.gov.by*), Университета Джонса Хопкинса (*https://www.jhu.edu*), Worldometer (*https://www.worldometers.info/coronavirus/#countries*). В виде таблиц эти данные можно скачать по ссылке: *https://github.com/CSSEGISandData/COVID-19*.

(если $x_t = \xi_t$) роста; $u_t \in R$ – случайная погрешность с нулевым математическим ожиданием $E\{u_t\} = 0$ и конечной дисперсией $\sigma^2 = D\{u_t\}$.

По выборке X_{t-s+1}^{t}, используя стандартное статистическое программное обеспечение (например, язык R и его библиотеки), вычислим следующие статистики: $\bar{x}(t) = \sum\limits_{i=s+1}^{t} \frac{x_i}{s}$ – выборочное среднее; $\hat{\theta}_0^{(t)}$, $\hat{\theta}_1^{(t)}$ – МНК-оценки параметров θ_0 и θ_1 соответственно; $r_{\min}^2(t) = \sum\limits_{i=t-s+1}^{t} \left(x_i - \hat{\theta}_0^{(t)} - \hat{\theta}_1^{(t)} i \right)^2$ – остаточная сумма квадратов;

$$\hat{\sigma}^2(t) = \frac{r_{\min}^2(t)}{s-2} \tag{3}$$

– несмещенная оценка дисперсии; $R^2(t) = \dfrac{\sum\limits_{i=s+1}^{t} \left(\hat{\theta}_0^{(t)} - \hat{\theta}_1^{(t)} i - \bar{x}(t) \right)^2}{\sum\limits_{i=s+1}^{t} \left(x_i - \bar{x}(t) \right)^2}$ – коэффициент детерминации модели. Чем ближе $R^2(t)$ к единице, тем адекватнее модель (2). Это позволяет выбрать длину фрагмента s из условия заданной адекватности модели:

$$R^2(t) \geq 0{,}9.$$

Точечные оценки для будущих значений $x_{T+1}, \ldots, x_{T+\tau}$ с учетом (2) имеют вид

$$\hat{x}_{T+j} = \hat{\theta}_0^{(T)} + \hat{\theta}_1^{(T)}(T+j),\ j = 1, 2, \ldots, \tau, \tag{4}$$

и основываются на оценках параметров (θ_0, θ_1) по «скользящему фрагменту» X_{T-s+1}^{T}.

Построим теперь интервальные прогнозы для будущих неизвестных значений $x_{T+1}, \ldots, x_{T+\tau}$. Примем следующие обозначения: $1 - \varepsilon$ – доверительная вероятность интервальных прогнозов, где $\varepsilon \in (0; 1)$ – задаваемый доверительный уровень (обычно $\varepsilon \in \{0{,}05;\ 0{,}10\}$); $t_{s-2}(1-\varepsilon) > 0$ – квантиль уровня $1 - \varepsilon$ стандартного t-распределения Стьюдента с $s - 2$ степенями свободы [10]; $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} ::=$

$::= \begin{pmatrix} \sum\limits_{i=T-s+1}^{T} 1 & \sum\limits_{i=T-s+1}^{T} i \\ \sum\limits_{i=T-s+1}^{T} i & \sum\limits_{i=T-s+1}^{T} i^2 \end{pmatrix}$ – заданная квадратная (2×2)-матрица, которая имеет следующий явный вид:

$$A = \begin{pmatrix} s & \dfrac{s(2T-s+1)}{2} \\ \dfrac{s(2T-s+1)}{2} & \dfrac{T(T+1)(2T+1) - (T-s)(T-s+1)(2T-2s+1)}{6} \end{pmatrix}. \tag{5}$$

С помощью (5) можно получить явный вид обратной матрицы $A^{-1} = (\bar{a}_{ij})$, который здесь не приводится из-за громоздкости.

Теорема 1. *Если имеет место локальная трендовая модель (2) для фрагмента $X_{T-s+1}^{T} \in R$ с независимыми случайными погрешностями $u_{T-s+1}, \ldots, u_{T+\tau}$, одинаково распределенными по нормальному закону*

$$\mathcal{L}\{u_t\} = \mathcal{N}_1(0, \sigma^2) \tag{6}$$

с неизвестной дисперсией σ^2, то $(1-\varepsilon) \cdot 100$ %-интервальные прогнозы для $x_{T+1}, \ldots, x_{T+\tau}$ имеют следующий вид: с вероятностью $1 - \varepsilon$

$$x_{T+j} \in \left(x_{T+j}^{-},\ x_{T+j}^{+} \right),\ j = 1, \ldots, \tau, \tag{7}$$

где доверительные границы вычисляются по формулам

$$x_{T+j}^{\pm} = \hat{x}_{T+j} \pm t_{s-2}\left(1-\frac{\varepsilon}{2}\right)\hat{\sigma}(T)\sqrt{1+\bar{a}_{11}+2\bar{a}_{12}(T+j)+\bar{a}_{22}(T+j)^2}, \tag{8}$$

$\hat{x}_{T+j}$ определяется выражением (4), а $\hat{\sigma}(T)$ – формулой (3).

Доказательство. Для построения $(1-\varepsilon)\cdot 100$ %-доверительного интервала для x_{T+j} при неизвестной дисперсии σ^2 применим метод стьюдентизации [11]. Для этого введем вспомогательные случайные величины

$$\lambda_1(j) = \frac{\hat{x}_{T+j} - x_{T+j}}{\sqrt{1+\bar{a}_{11}+2\bar{a}_{12}(T+j)+\bar{a}_{22}(T+j)^2}},$$

$$\lambda_2(j) = \frac{r_{\min}^2(T)}{\sigma^2} = \frac{(s-2)\hat{\sigma}^2(T)}{\sigma^2}, \tag{9}$$

$$\lambda(j) = \frac{\lambda_1(j)}{\sqrt{\lambda_2(j)/(s-2)}}.$$

Согласно [11] $\lambda_1(j)$, $\lambda_2(j)$ независимы, причем $\lambda_2(j)$ имеет стандартное χ^2-распределение с $s-2$ степенями свободы:

$$\mathcal{L}\{\lambda_2(j)\} = \chi_{s-2}^2. \tag{10}$$

В силу (2), (4), (6) и теоремы о линейности МНК-оценок $\hat{\theta}_0^{(T)}$, $\hat{\theta}_1^{(T)}$ [11] получаем

$$\mathcal{L}\{\hat{x}_{T+j} - x_{T+j}\} = \mathcal{N}(0,\, b^2(j)),$$

$$b^2(j) = D\{\hat{x}_{T+j} - x_{T+j}\} = D\{\hat{\theta}_0^{(T)} + \hat{\theta}_1^{(T)}(T+j)\} + D\{x_{T+j}\} =$$

$$= D\{\hat{\theta}_0^{(T)}\} + 2(T+j)Cov\{\hat{\theta}_0^{(T)},\, \hat{\theta}_1^{(T)}\} + (T+j)^2 D\{\hat{\theta}_1^{(T)}\} + \sigma^2. \tag{11}$$

Согласно работе [11] и матрице (5) ковариационная матрица оценок $\hat{\theta}_0^{(T)}$, $\hat{\theta}_1^{(T)}$ равна

$$Cov\left\{\begin{pmatrix}\hat{\theta}_0^{(T)}\\ \hat{\theta}_1^{(T)}\end{pmatrix}, \begin{pmatrix}\hat{\theta}_0^{(T)}\\ \hat{\theta}_1^{(T)}\end{pmatrix}\right\} = \sigma^2 \cdot A^{-1}.$$

Подставляя это выражение в (11), получаем

$$b^2(j) = \sigma^2\left(1+\bar{a}_{11}+2\bar{a}_{12}(T+j)+\bar{a}_{22}(T+j)^2\right). \tag{12}$$

Из (11), (12) заключаем, что

$$\mathcal{L}\{\lambda_1(j)\} = \mathcal{N}_1(0,\, 1).$$

Тогда из (9), (10) и (12) следует, что случайная величина $\lambda(j)$ имеет стандартное t-распределение Стьюдента с $s-2$ степенями свободы. Поэтому

$$P\left\{-t_{s-2}\left(1-\frac{\varepsilon}{2}\right) < \lambda(j) < +t_{s-2}\left(1-\frac{\varepsilon}{2}\right)\right\} = 1-\varepsilon. \tag{13}$$

Подставляя в (13) вместо члена $\lambda(j)$ его выражение из (9) и разрешая двустороннее неравенство относительно прогнозируемой случайной величины x_{T+j}, приходим к (7), (8).

Для иллюстрации на рис. 1 представлены результаты прогнозирования показателя ν_t для $s = 12$.

Локально-медианный метод прогнозирования

Локально-медианный метод прогнозирования разработан в монографии [12] как робастный метод прогнозирования временных рядов [13–15].

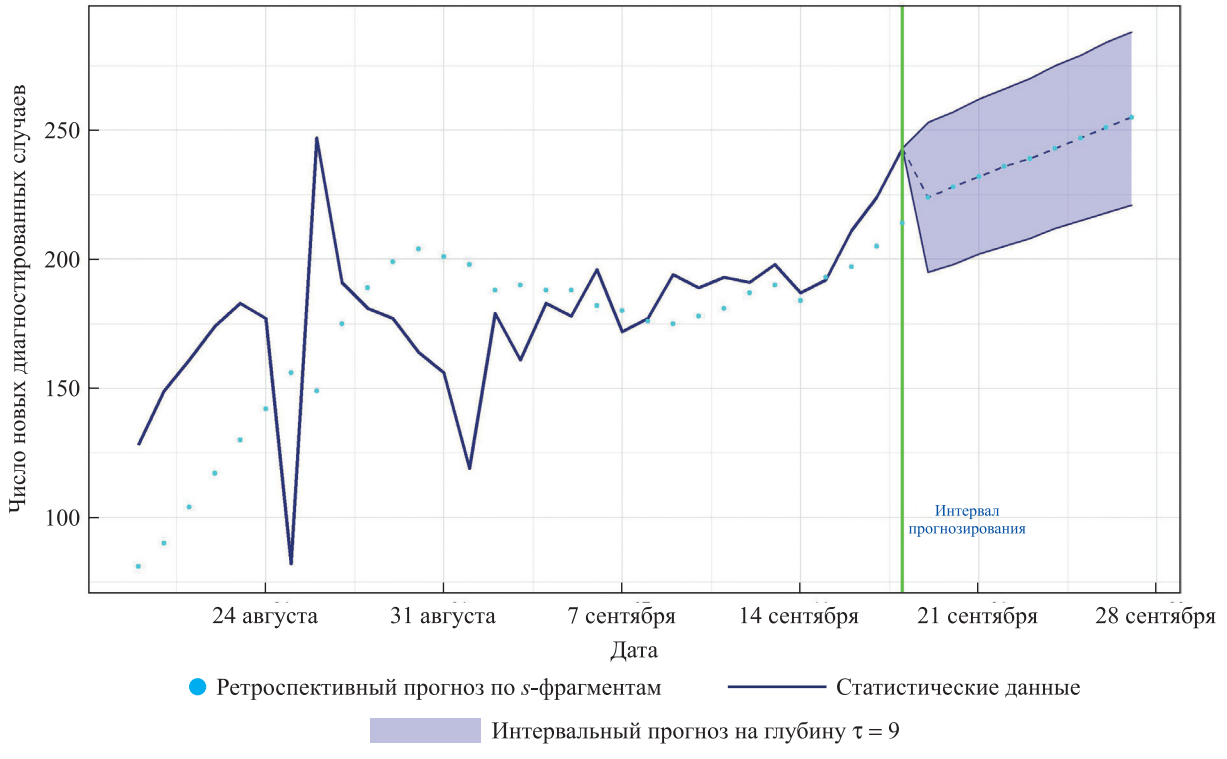

Рис. 1. Результаты прогнозирования для зарегистрированного числа ν_t новых зараженных в день t с помощью метода «скользящих трендов»

Fig. 1. Forecasting results for the registered number ν_t of new infections on day t using the «moving trends» method

Выберем ближайший к моменту прогнозирования T фрагмент $X_{T-s+1}^{T} = \left(x_{T-s+1}, x_{T-s+2}, \ldots, x_T\right) \in R^s$ длины s, где $2 < s \le T$ (если $s = T$, то рассматривается весь наблюдаемый временной ряд X_1^T), и будем считать, что для него адекватна линейная трендовая модель (2). Предположение (6) о гауссовском распределении случайных погрешностей $\{u_t\}$ здесь использовать не будем. Примем следующие обозначения: $N_s = \{T-s+1, \ldots, T\}$ – множество s используемых моментов времени; $\Gamma^{(l)} = \left\{t_1^{(l)}, t_2^{(l)}, \ldots, t_m^{(l)}\right\} \subset N_s$ – подмножество m ($2 \le m < s$) упорядоченных по возрастанию моментов времени $T-s+1 \le t_1^{(l)} < t_2^{(l)} < \ldots < t_m^{(l)} \le T$, $l = 1, 2, \ldots, L$; $L = C_s^m$ – число всех различных подмножеств $\Gamma^{(l)}$; $X^{(l)} = \left\{x_{t_1^{(l)}}, x_{t_2^{(l)}}, \ldots, x_{t_m^{(l)}}\right\}'$ – l-й m-вектор-столбец наблюдений для моментов времени $\Gamma^{(l)}$; $\Psi^{(l)} = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ t_1^{(l)} & t_2^{(l)} & \ldots & t_m^{(l)} \end{pmatrix}'$ – транспонированная $(m \times 2)$-матрица.

Для построения точечных и интервальных прогнозов для $x_{T+1}, \ldots, x_{T+\tau}$ применим локально-медианный метод [12]. Используя принятые выше обозначения, введем вначале семейство L локальных оценок параметров $\theta = \left(\theta_0, \theta_1\right)'$ модели (2):

$$\hat{\theta}^{(l)} = \begin{pmatrix} \hat{\theta}_0^{(l)} \\ \hat{\theta}_1^{(l)} \end{pmatrix} = \left(\Psi^{(l)\prime}\Psi^{(l)}\right)^{-1} \Psi^{(l)\prime} X^{(l)}, \; l = 1, \ldots, L. \tag{14}$$

По локальным оценкам (14) с учетом модели (2) построим далее семейство L локальных прогнозов будущего состояния:

$$\hat{x}_{T+j}^{(l)} = \hat{\theta}_0^{(l)} + \hat{\theta}_1^{(l)}\left(T + j\right), \; l = 1, \ldots, L. \tag{15}$$

Локально-медианный прогноз определяется как выборочная медиана локальных прогнозов (15):

$$\hat{x}_{T+j} = S\left(X_{T-s+1}^{T}\right) ::= med\left\{\hat{x}_{T+j}^{(1)}, \ldots, \hat{x}_{T+j}^{(L)}\right\}. \tag{16}$$

Заметим, что обращение матриц в (14) корректно, так как по построению матрица $\Psi^{(l)}$ имеет ранг 2. Параметр алгоритма m определяет мощность подмножеств $\Gamma^{(l)}$. От него зависят точность и вычислительная сложность алгоритма (14)–(16).

Для построения $(1-\varepsilon)\cdot 100$ %-интервальных прогнозов для $x_{T+1}, \ldots, x_{T+\tau}$ упорядочим полученную согласно (16) выборку L локальных прогнозов в порядке возрастания их величин и сформируем вариационный ряд

$$\hat{x}_{T+j(1)} \le \hat{x}_{T+j(2)} \le \ldots \le \hat{x}_{T+j(L)}.$$

Медиана (16) является «средним» членом в этом вариационном ряду:

$$\hat{x}_{T+j} = \begin{cases} \hat{x}_{T+j(k+1)}, & L = 2k+1 \text{ (нечетное)}, \\ \dfrac{\hat{x}_{T+j(k)} + \hat{x}_{T+j(k+1)}}{2}, & L = 2k \text{ (четное)}. \end{cases}$$

Для нахождения границ $(1-\varepsilon)\cdot 100$ % -доверительного интервала для x_{T+j} отбросим

$$K = \left[\frac{\varepsilon L}{2} + 1\right]$$

наименьших и K наибольших членов вариационного ряда. Оставшиеся в вариационном ряду крайние члены и определяют границы $(1-\varepsilon)\cdot 100$ %-доверительного интервала для x_{T+j} ($j = 1, \ldots, \tau$): с вероятностью $1-\varepsilon$

$$x_{T+j} \in \left(x^{-}_{T+j}, x^{+}_{T+j}\right), \; x^{-}_{T+j} = \hat{x}_{T+j(K+1)}, \; x^{+}_{T+j} = \hat{x}_{T+j(L-K)}.$$

Отметим, что по выборке локальных прогнозов $\hat{x}^{(1)}_{T+j}, \ldots, \hat{x}^{(L)}_{T+j}$ можно построить гистограмму распределения прогнозов и с ее помощью определять «шансы» каждого прогноза x, т. е. построить вероятностный прогноз.

Для иллюстрации локально-медианного метода на рис. 2 представлены результаты прогнозирования показателя ν_t для $s = 7$, $m = 5$.

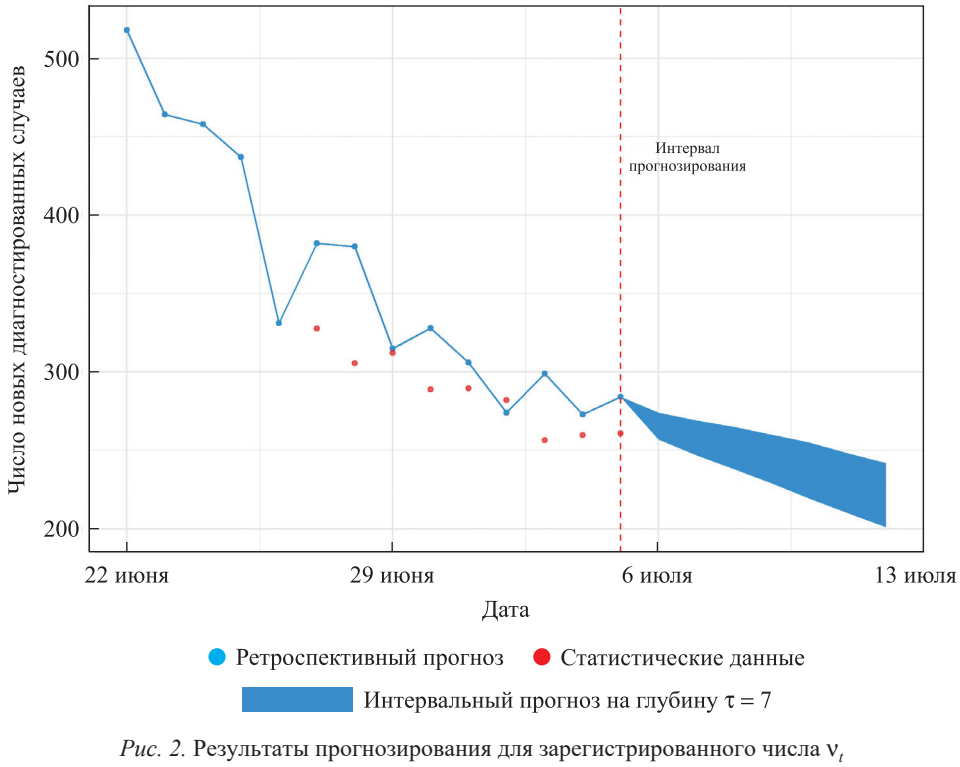

Рис. 2. Результаты прогнозирования для зарегистрированного числа ν_t новых зараженных в день t с помощью локально-медианного метода

Fig. 2. Forecasting results for the registered number ν_t of new infections on day t using the local median method

Прогнозирование на основе моделей дискретных временных рядов

Введем в рассмотрение вероятностную версию известной SIR-модели [16]. Будем использовать обозначения: $t \in \mathrm{Z}$ – дискретное время (номер суток); $N \in \mathrm{N}$ – численность населения страны (считаем ее постоянной, витальная динамика (смертность, рождаемость) не учитывается, что согласуется с низкой смертностью от COVID-19 в Республике Беларусь); $x_t \in \mathrm{N}$ – общее число инфицированных к моменту t ($x_t \le N$); y_t – общее число выздоровевших к моменту t ($y_t \le x_t$). Соответственно, разность $x_t - y_t$ равна числу активных инфицированных в момент t. В обозначениях SIR-модели $S_t = N - x_t$, $I_t = x_t - y_t$, $R_t = y_t$. Будем также использовать оператор левого дискретного дифференцирования (соседних разностей) $\Delta x_t = x_t - x_{t-1}$. Тогда Δx_t – число инфицированных за сутки, Δy_t – число выздоровевших за сутки. Определим следующую вероятностную версию SIR-модели:

$$\Delta x_{t+1} \,\big|\, F_t \sim Bi\left(N - x_t,\, \beta \frac{x_t - y_t}{N}\right) \approx \Pi\left(\beta \frac{(N - x_t)(x_t - y_t)}{N}\right), \tag{17}$$

$$\Delta y_{t+1} \,\big|\, F_t \sim Bi\big(x_t - y_t,\, \gamma\big) \approx \Pi\big(\gamma\big(x_t - y_t\big)\big), \tag{18}$$

$$\Delta x_{t+1} \perp \Delta y_{t+1} \,\big|\, F_t, \tag{19}$$

где $A \sim B$ означает «случайная величина A распределена по закону B»; F_t – сигма-алгебра, порожденная случайными величинами $\left(x_\tau,\, y_\tau\right)_{\tau \le t}$ (предыстория в момент t); $A \,\big|\, F_t$ – условное распределение вероятностей случайной величины A при фиксированной предыстории в момент t; $A \perp B \,\big|\, F_t$ означает «случайные величины A и B условно независимы при фиксированной предыстории в момент t»; $Bi\big(n,\, p\big)$ – биномиальное распределение с параметрами $p \in [0, 1]$ и $n \in \mathrm{N}$; $\Pi\big(\lambda\big)$ – пуассоновское распределение с параметром $\lambda > 0$. Приближенные равенства вида $Bi\big(n,\, p\big) \approx \Pi\big(np\big)$ между биномиальным и пуассоновским распределениями в уравнениях (17), (18) применимы в случае выполнения соответствующей асимптотики: n достаточно велико, p достаточно мало. Модель (17)–(19), в которой используются только биномиальные распределения, будем для краткости называть моделью BBSIR, где первая буква B относится к уравнению (17), вторая – к уравнению (18). Если в одном из уравнений вместо биномиального распределения используется его пуассоновская аппроксимация, соответствующую букву в названии модели будем заменять на P. Если в обоих уравнениях (17), (18) применяется пуассоновское приближение, получаем, соответственно, модель PPSIR. Параметр β характеризует интенсивность заражений, параметр γ – интенсивность выздоровлений. Величину $\frac{1}{\gamma}$ принято интерпретировать как среднюю продолжительность заболевания человека – от заражения до выздоровления.

Пусть теперь наблюдается эпидемиологический процесс $\left(x_t,\, y_t\right)_{t=t_1}^{t_2}$ длительностью $T = t_2 - t_1 + 1$ дней. Построим логарифмическую функцию правдоподобия модели BBSIR на основе марковских свойств (17), (18):

$$L_{\mathrm{BBSIR}}\big(\beta,\, \gamma\big) = \sum_{t=t_1}^{t_2-1}\left(\ln\binom{N - x_t}{\Delta x_{t+1}} + \ln\binom{x_t - y_t}{\Delta y_{t+1}} + \Delta y_{t+1} \ln\gamma + \big(x_t - y_{t+1}\big)\ln\big(1 - \gamma\big) + \right.$$

$$\left. + \Delta x_{t+1} \ln\left(\beta \frac{x_t - y_t}{N}\right) + \big(N - x_{t+1}\big)\ln\left(1 - \beta \frac{x_t - y_t}{N}\right)\right). \tag{20}$$

Из (20) следует, что функция двух переменных $L_{\mathrm{BBSIR}}\big(\beta,\, \gamma\big)$ распадается на сумму функций, зависящих от каждой переменной в отдельности, и задача максимизации $L_{\mathrm{BBSIR}}\big(\beta,\, \gamma\big)$ разбивается на две подзадачи максимизации однопараметрических функций. Максимизация по γ дает оценку максимального правдоподобия (ОМП):

$$\hat{\gamma} = y_{t_2} - \frac{y_{t_1}}{\sum\limits_{t=t_1}^{t_2-1} x_t - \sum\limits_{t=t_1+1}^{t_2} y_t}. \tag{21}$$

Зависящее от β слагаемое в (20) может быть приближенно максимизировано, например, полным перебором по дискретной сетке на отрезке допустимых значений:

$$\beta \in \left[0, \frac{N}{\max\{x_t - y_t : t_1 \le t < t_2\}}\right].$$

В случае модели PBSIR логарифмическая функция правдоподобия примет вид

$$L_{\text{PBSIR}}(\beta, \gamma) = \sum_{t=t_1}^{t_2-1}\left(\ln\binom{x_t - y_t}{\Delta y_{t+1}} - \ln(\Delta x_{t+1}!) + \Delta y_{t+1}\ln\gamma + (x_t - y_{t+1})\ln(1-\gamma) + \right.$$
$$\left. + \Delta x_{t+1}\ln\left(\beta\frac{(N-x_t)(x_t-y_t)}{N}\right) - \beta\frac{(N-x_t)(x_t-y_t)}{N}\right). \tag{22}$$

ОМП параметра γ для модели PBSIR вычисляется так же, как и для модели BBSIR, по формуле (21). ОМП параметра β для модели PBSIR получается приравниванием к нулю производной по β функции (22):

$$\hat{\beta}_{\text{PBSIR}} = N\frac{x_{t_1} - x_{t_2}}{\sum_{t=t_1}^{t_2-1}(N-x_t)(x_t-y_t)}.$$

По аналогии с рассмотренными моделями BBSIR и PBSIR могут быть построены ОМП параметров моделей BPSIR и PPSIR.

Рассмотрим теперь модификацию модели PBSIR, в которой параметры β и γ (интенсивности заражений и выздоровлений соответственно) зависят от времени детерминированным образом:

$$\Delta x_{t+1}\big|F_t \sim \Pi\left(\beta_{t+1}\frac{(N-x_t)(x_t-y_t)}{N}\right),$$
$$\Delta y_{t+1}\big|F_t \sim Bi(x_t - y_t, \gamma_{t+1}),$$
$$\beta_t = \exp(b_t),\ b_t = \sum_{i=1}^{m_\beta}\psi_i^\beta(t)a_i^\beta = \langle\psi^\beta(t), a^\beta\rangle, \tag{23}$$
$$\gamma_t = \Lambda(c_t),\ c_t = \sum_{i=1}^{m_\gamma}\psi_i^\gamma(t)a_i^\gamma = \langle\psi^\gamma(t), a^\gamma\rangle,\ \Lambda(z) = \frac{1}{1+e^{-z}},$$

где b_t и c_t – канонические параметры условных пуассоновского и биномиального распределений (17), (18), двойственные параметрам β_t и γ_t соответственно; $\Lambda(\cdot)$ – логистическая функция распределения; $\langle u, \nu\rangle = \sum_i u_i\nu_i$ – скалярное произведение действительнозначных векторов $u = (u_i)$ и $\nu = (\nu_i)$; $\{\psi_i^\beta\}_{i=1}^{m_\beta}$ – базис из $m_\beta \in \mathbb{N}$ линейно независимых на исследуемом отрезке $t_1 \le t \le t_2$ функций $\psi_i^\beta : \mathbb{Z} \to R$ для задания интенсивности заражений β_t (интенсивность заражений γ_t задается аналогично базисом $\{\psi_i^\gamma\}_{i=1}^{m_\gamma}$); $\psi^\beta(t) = (\psi_i^\beta(t))_{i=1}^{m_\beta}$ – m_β-вектор одновременных значений базисных функций $\{\psi_i^\beta\}_{i=1}^{m_\beta}$ в момент t (аналогично $\psi^\gamma(t) = (\psi_i^\gamma(t))_{i=1}^{m_\gamma}$); векторы коэффициентов $a^\beta = (a_i^\beta)_{i=1}^{m_\beta}$, $a^\gamma = (a_i^\gamma)_{i=1}^{m_\gamma}$ – параметры модифицированной модели PBSIR (19), (23), которую далее будем называть моделью TPBSIR (приставка T означает зависимость параметров β и γ от времени).

Логарифмическая функция правдоподобия для модели TPBSIR аналогично (22) распадается на сумму трех слагаемых:

$$L_{\text{TPBSIR}}(a^\beta, a^\gamma) = L^0_{\text{TPBSIR}} + L^\beta_{\text{TPBSIR}}(a^\beta) + L^\gamma_{\text{TPBSIR}}(a^\gamma),$$
$$L^0_{\text{TPBSIR}} = \sum_{t=t_1}^{t_2-1}\left(\ln\binom{x_t - y_t}{\Delta y_{t+1}} - \ln(\Delta x_{t+1}!) + \Delta x_{t+1}\ln\left(\frac{(N-x_t)(x_t-y_t)}{N}\right)\right), \tag{24}$$

$$L^{\beta}_{\text{TPBSIR}}\left(a^{\beta}\right)=\sum_{t=t_1}^{t_2-1}\left(\Delta x_{t+1}b_{t+1}-\frac{\left(N-x_t\right)\left(x_t-y_t\right)}{N}\exp\left(b_{t+1}\right)\right),$$

$$L^{\gamma}_{\text{TPBSIR}}\left(a^{\gamma}\right)=\sum_{t=t_1}^{t_2-1}\left(\Delta y_{t+1}c_{t+1}-\left(x_t-y_t\right)\ln\left(1+\exp\left(c_{t+1}\right)\right)\right).$$

Согласно свойствам канонических параметров экспоненциальных распределений каждое слагаемое в (24) есть выпуклая функция от соответствующего канонического параметра b_{t+1}, а следовательно, и от a^{β}, поскольку b_{t+1} линейно зависит от a^{β}. Поэтому функция $L^{\beta}_{\text{TPBSIR}}\left(a^{\beta}\right)$ выпукла, более того, она строго выпукла в силу линейной независимости базисных функций $\left\{\psi^{\beta}_i\right\}_{i=1}^{m_{\beta}}$, и, значит, $L^{\beta}_{\text{TPBSIR}}\left(a^{\beta}\right)$ имеет единственный локальный и глобальный максимум, который может быть найден методом градиентного подъема. Аналогично выводится строгая выпуклость и единственность локального и глобального максимума функции $L^{\gamma}_{\text{TPBSIR}}\left(a^{\gamma}\right)$.

Модель TPBSIR применима для разбиения наблюдаемого эпидемиологического процесса на фазы, такие как рост, плато, спад. Для этого могут использоваться кусочно-заданные базисные функции $\left\{\psi^{\beta}_i\right\}$, $\left\{\psi^{\gamma}_i\right\}$, равные нулю за пределами своей фазы. Границы фаз при этом становятся дискретными параметрами модели, для которых методом перебора строятся оценки максимального правдоподобия. Для сокращения перебора налагаются дополнительные ограничения, например запрет слишком коротких фаз.

Прогнозирование на основе описанной модели TPBSIR производится с использованием имитационного моделирования. Для построения прогноза на τ дней вперед необходимо, чтобы базисные функции $\left\{\psi^{\beta}_i\right\}$, $\left\{\psi^{\gamma}_i\right\}$ были определены при $t = t_2 + 1, \ldots, t_2 + \tau$. Тогда согласно модели (23) строятся K траекторий $\left\{x^i_t, y^i_t\right\}_{t=t_2+1}^{t_2+\tau}$, $i = 1, \ldots, K$. Прогноз значения $x_{t_2+\tau'}$, $1 \le \tau' \le \tau$, может быть построен как потраекторное среднее или медиана (для $y_{t_2+\tau'}$ аналогично):

$$\hat{x}_{t_2+\tau'}=K^{-1}\sum_{i=1}^{K}x^i_{t_2+\tau'},\ \tilde{x}_{t_2+\tau'}=x^{\left([K/2]\right)}_{t_2+\tau'},$$

где $[z]$ означает целую часть z; $x^{(1)}_{t_2+\tau'} \le x^{(2)}_{t_2+\tau'} \le \ldots \le x^{(K)}_{t_2+\tau'}$ – вариационный ряд, составленный из значений $\left\{x^i_{t_2+\tau'}\right\}_{i=1}^{K}$. Доверительный интервал $\left[x^-_{t_2+\tau'}, x^+_{t_2+\tau'}\right]$ для $x_{t_2+\tau'}$ с уровнем значимости $0 < \alpha < 1$ имеет границы $x^{\pm}_{t_2+\tau'}=x^{\left([K(1\pm\alpha)/2]\right)}_{t_2+\tau'}$.

Теорема 2. *При фиксированной предыстории условный среднеквадратический риск прогнозирования на 1 день вперед на основе модели* TPBSIR *имеет вид*

$$E\left\{\left(\hat{x}_{t_2+1}-x_{t_2+1}\right)^2\middle|F_{t_2}\right\}=\beta_{t_2+1}\frac{\left(N-x_{t_2}\right)\left(x_{t_2}-y_{t_2}\right)}{N},$$

$$E\left\{\left(\hat{y}_{t_2+1}-y_{t_2+1}\right)^2\middle|F_{t_2}\right\}=\left(x_{t_2}-y_{t_2}\right)\gamma_{t_2+1}\left(1-\gamma_{t_2+1}\right).$$

Доказательство. Достаточно воспользоваться (17)–(19) и свойствами биномиального и пуассоновского распределений.

Опыт применения модели TPBSIR для прогнозирования параметров распространения COVID-19 в Республике Беларусь на разных этапах эпидемии, начиная с апреля 2020 г., показал, что при прогнозировании на неделю ($\tau = 7$ дней) среднеквадратическая ошибка, как правило, не превышала 50 человек. На рис. 3 представлены прогнозы параметров распространения COVID-19 в Республике Беларусь на основе модели TPBSIR с базисом аффинных функций $\left\{\psi^{\beta}_i\right\}=\{1, t\}$, $m_{\beta} = 2$, для интенсивности заражений и базисом аффинных функций с недельной периодической компонентой $\left\{\psi^{\gamma}_i\right\}=\left\{1, t, \sin\left(\frac{2\pi t}{7}\right), \cos\left(\frac{2\pi t}{7}\right)\right\}$, $m_{\gamma} = 4$, для интенсивности выздоровлений. Добавление периодической компоненты обусловлено обнаруженной недельной периодичностью в наблюдаемой интенсивности выздоровлений от COVID-19 в Республике Беларусь. Долгосрочный прогноз, изображенный на рис. 3, был построен 29 августа 2020 г. на даты до конца сентября 2020 г.

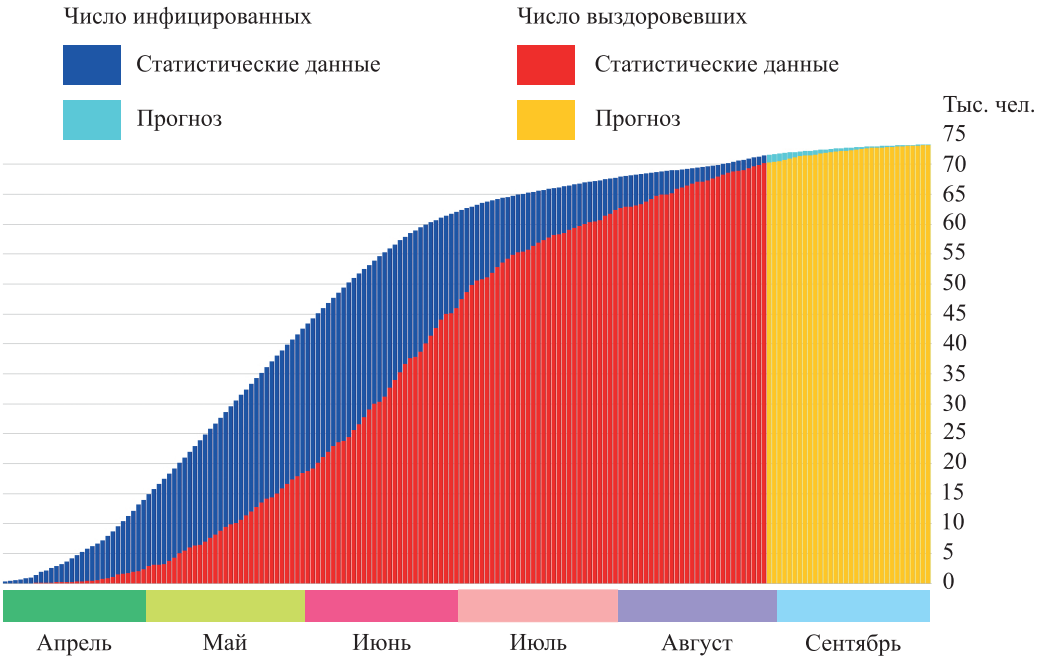

Рис. 3. Долгосрочный прогноз параметров распространения COVID-19 в Республике Беларусь на основе модели TPBSIR до 30 сентября 2020 г. включительно (дата прогнозирования – 29 августа 2020 г.)

Fig. 3. Long-term forecast of the dynamics for COVID-19 dissemination in Republic of Belarus based on the TPBSIR model for the dates up to 30 September 2020 (forecast is made on 29 August 2020)

Прогнозирование на основе эконометрической модели коррекции ошибок

Для моделирования эпидемиологического процесса COVID-19 в Беларуси разработана эконометрическая векторная модель коррекции ошибок (*vector error correction model*) VECM COVID-19 RB. Она базируется на теории коинтеграции нестационарных интегрированных временных рядов, предполагающей существование между ними при определенных условиях долгосрочной равновесной зависимости, которая учитывается при построении краткосрочных прогнозов [17].

Модель VECM COVID-19 RB основана на близких с известной моделью эпидемиологического процесса SIR (*susceptible – infectious – recovered*) [18] предположениях, главным из которых является предположение о существовании долгосрочной равновесной зависимости для устойчивого состояния эпидемиологического процесса вида

$$I(t)+R(t)+S(t)=N, \tag{25}$$

где (для момента времени t) $I(t)$ – численность инфицированных индивидов; $R(t)$ – численность переболевших индивидов; $S(t)$ – численность восприимчивых к инфекции индивидов; N – численность всей популяции.

Модель VECM COVID-19 RB отличается от SIR-модели следующими основными особенностями:

1) она является не детерминированной, а стохастической и не предполагает наличие управляемых параметров;

2) в силу незначительной доли умерших (менее 1 % от общего числа заражений за весь период наблюдения) переменная $R(t)$ соответствует числу всех закрытых случаев заражения, т. е. включает выздоровевших и умерших. С учетом этого тождество (25) допускает интерпретацию

$$I(t)+R(t)=N-S(t)=T(t),$$

где $T(t)$ – общее число случаев заражения (*total infected*) в момент t;

3) условие долгосрочной коинтеграционной зависимости, связывающее переменные $I(t)$, $R(t)$, оценивается и тестируется в процессе построения модели;

4) моделированию и прогнозированию подлежат ежедневные изменения переменных $I(t)$, $R(t)$, т. е. их первые разности $\Delta I(t)$, $\Delta R(t)$, а соответствующие им уравнения используются для построения краткосрочных прогнозов;

5) все параметры модели неизвестны и оцениваются в процессе построения модели по ежедневным значениям переменных $I(t)$, $R(t)$.

Помимо краткосрочных прогнозов, с помощью модели VECM COVID-19 RB проводился долгосрочный анализ динамики эпидемиологического процесса в целях оценивания «поворотной точки» (когда текущее число зараженных равно числу закрытых случаев заражения), а также возможных сроков завершения острой стадии процесса (когда число закрытых случаев близко к общему числу заражений). Оба эти события были предсказаны на основе разработанной модели примерно за месяц до их наступления. В данной статье для иллюстрации представлена модель для первой волны COVID-19 в Беларуси.

Модель коррекции ошибок при сделанных предположениях включает три уравнения: коинтеграционное уравнение (*cointegration equation*), описывающее долгосрочную зависимость (*long run relation*) между временными рядами $I(t)$, $R(t)$, и два уравнения краткосрочных зависимостей (*short run relations*) для их первых разностей, которые соответствуют ежедневным изменениям и используются для построения краткосрочных прогнозов временных рядов $I(t)$, $R(t)$. Прогнозы для общего числа случаев заражения T вычисляются суммированием прогнозных значений для $I(t)$, $R(t)$.

Для построения векторной модели коррекции ошибок применяется подход Йохансена [19]. Для временных рядов $x_{1,t} \equiv R(t)$, $x_{2,t} \equiv I(t)$ с помощью расширенного теста Дики – Фуллера (*augmented Dickey – Fuller (ADF) unit root test*) установлена интегрированность второго порядка, т. е. наличие как стохастических трендов, так и детерминированного квадратичного тренда. По этой причине для тестирования коинтегрированности $R(t)$ и $I(t)$ используется спецификация модели VECM, предполагающая наличие квадратичных трендов во временных рядах $R(t)$, $I(t)$ [20]. При тестировании коинтеграции временных рядов и оценивании модели это учитывается добавлением линейного тренда t и константы c в уравнения для долгосрочной и краткосрочных зависимостей. Поскольку временные ряды ежедневных изменений (первые разности) $\Delta x_{1,t} \equiv \Delta R(t)$, $\Delta x_{2,t} \equiv \Delta I(t)$ имеют семидневные циклические колебания, то в уравнения для $\Delta x_{1,t}$, $\Delta x_{2,t}$ включены лаговые переменные с лагом 7. При построении и применении модели используются следующие временные интервалы: максимальный период оценивания – с 18 мая по 20 августа; прогнозный период (краткосрочные прогнозы) – с 21 по 31 августа; анализ долгосрочной динамики – с 21 августа по 21 сентября. Оцененные уравнения модели для моментов времени t ($t = 1, 2, \ldots$) имеют следующий вид.

1. Отклонения от долгосрочной зависимости (disequilibrium errors) для лага $t - 1$

$$\widehat{\xi}_{t-1} = x_{1,t-1} - 0{,}311\,3x_{2,t-1} - 679{,}10t + 21\,857{,}73.$$

2. Краткосрочные зависимости

$$\Delta x_{1,t} = -0{,}094\,2\widehat{\xi}_{t-1} - 1{,}627\,0\Delta x_{1,t-1} - 0{,}790\,5\Delta x_{1,t-2} -$$
$$- 1{,}918\,7\Delta x_{2,t-1} - 0{,}585\,5\Delta x_{2,t-2} + 0{,}360\,7\Delta x_{1,t-7} - \mathbf{0{,}152\,8}\Delta x_{2,t-7} - 30{,}93t + 4244{,}80,$$

$$\Delta x_{2,t} = 0{,}073\,4\widehat{\xi}_{t-1} + 1{,}830\,4\Delta x_{1,t-1} + 0{,}851\,0\Delta x_{1,t-2} -$$
$$- 2{,}112\,1\Delta x_{2,t-1} + 0{,}637\,4\Delta x_{2,t-2} - 0{,}451\,3\Delta x_{1,t-7} + \mathbf{0{,}068\,1}\Delta x_{2,t-7} + \mathbf{23{,}96}t + 3284{,}57.$$

Значения скорректированного коэффициента детерминации R^2_{adj} для уравнений $\Delta x_{1,t}$, $\Delta x_{2,t}$ равны 0,701 471 и 0,695 933 соответственно. На основании модифицированного теста Льюнга – Бокса (*residual portmanteau test for autocorrelations*) остатки не коррелированы вплоть до лага 6 (для обоих уравнений) и имеют нормальный закон (для первого уравнения). Жирным шрифтом в уравнениях выделены оценки параметров, статистически не значимые на уровне 0,05.

Долгосрочная динамика эпидемиологического процесса COVID-19 исследуется с помощью модели, оцененной на расширенном временном интервале с 18 мая по 20 августа, учитывающей взаимосвязь переменных за более долгий период, чем в случае краткосрочного прогнозирования. Представленная на рис. 4 прогнозная динамика моделируемого процесса с 21 августа по 21 сентября свидетельствует о том, что в момент построения прогноза (20 августа) ожидалось затухание эпидемии с относительно невысоким уровнем новых заражений до начала сентября. Со второй декады сентября прогнозировался незначительный рост числа новых заражений. При этом предполагалось, что в рассматриваемый период времени будут отсутствовать новые факторы роста эпидемиологического процесса.

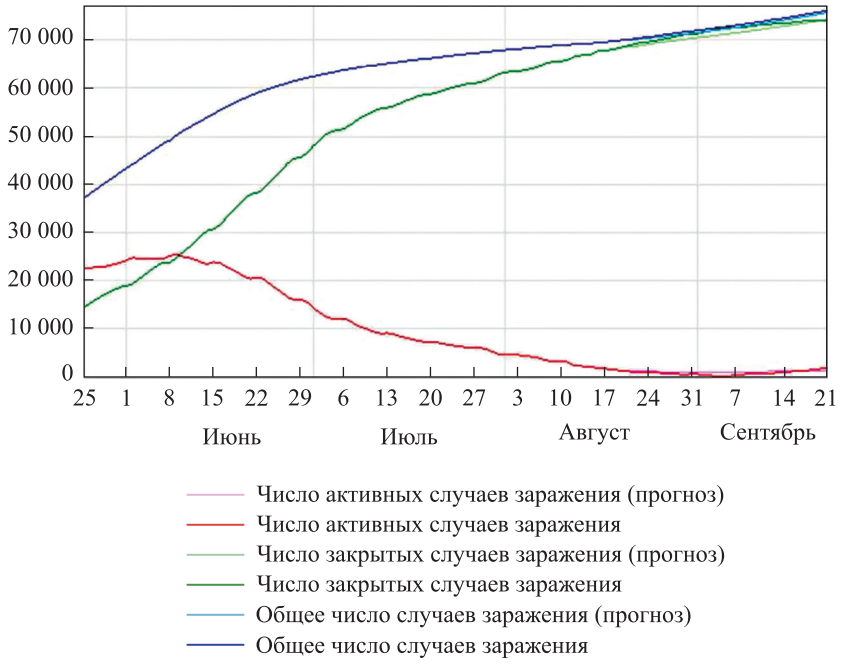

Рис. 4. Долгосрочная динамика эпидемиологического процесса COVID-19 с 21 августа по 21 сентября 2020 г.

Fig. 4. Long-term dynamics of the epidemiologic process for COVID-19 from 21 August to 21 September 2020

Применение последовательного статистического анализа для исследования текущих тенденций заболеваемости

Последовательный статистический анализ, предложенный американским математиком А. Вальдом, позволяет строить статистические выводы, основываясь лишь на минимально необходимом количестве наблюдений, которое не фиксируется заранее, а определяется в зависимости от поступающих случайных наблюдений так, чтобы обеспечивать требуемую точность метода (например, малые значения вероятностей ошибочных решений), и, как следствие, само является случайной величиной [21]. Это обстоятельство затрудняет теоретический анализ эффективности последовательных статистических процедур, однако позволяет «экономно» использовать наблюдения и останавливать процесс принятия решений сразу, как только обеспечивается заданная точность по тем данным, которые наблюдаются в конкретной исследуемой ситуации.

Рассмотрим применение последовательного статистического анализа для решения задач исследования динамики заболеваемости COVID-19.

Для мониторинга текущих тенденций заболеваемости будем полагать, что случайные наблюдения $\nu_1, \nu_2, \ldots$ зарегистрированных чисел новых зараженных в дни 1, 2, … описываются следующей простейшей вероятностной моделью:

$$\nu_t = \nu + \theta t + \lambda_t, \; t = 1, 2, \ldots, \tag{26}$$

где ν – заданный уровень для «стационарной» ситуации; $\theta \geq 0$ – параметр (его значение неизвестно), отвечающий за тренд – тенденцию развития эпидемиологической ситуации на рассматриваемом коротком промежутке времени; $t = 1$ – момент начала мониторинга; λ_t, $t \in N$, представляют собой независимые одинаково распределенные случайные величины, описывающие случайные колебания заболеваемости с нулевым математическим ожиданием $E\{\lambda_t\} = 0$. Относительно значения параметра θ имеются две гипотезы: $H_0 : \theta = 0$ (эпидемиологическая ситуация находится в «стационарной» стадии плато), $H_1 : \theta = \theta_1 > 0$ (уровень заболеваемости начал расти, где θ_1 определяется, например, из условия достижения трендом ко дню τ некоторого критического уровня заболеваемости). Пусть заданы максимально допустимые значения вероятностей ошибок 1-го рода (принята H_1 при справедливой гипотезе H_0) и 2-го рода (принимается H_0, когда верна H_1). Соответствующий последовательный статистический тест проверки гипотез H_0, H_1 в этом случае построен и исследован в работе [22]. Кроме того, теория, представленная в статье [22], позволяет рассмотреть в (26) зависимости, отличные от линейных. Анализ эффективности указанного теста дает возможность использовать описанный подход для отслеживания наметившихся отклонений от «стационарной» ситуации и быстрого реагирования на них.

В общем случае, когда необходим более детальный анализ, может потребоваться формулировка гипотез в следующем виде:

$$H_0 : \theta \in [0, \theta_0], \; H_1 : \theta \geq \theta_1 \; (\theta_0 < \theta_1).$$

В такой постановке методология построения соответствующего последовательного теста и анализа его эффективности представлена в работе [23]. Вместо трендовой модели (26), предполагающей лишь абстрагированный от причин анализ числа случаев заболеваемости, можно применить построенные в монографии [21] последовательные тесты для модели марковской зависимости $\{\nu_t\}$ порядка p.

Библиографические ссылки

1. Кондратьев МА. Методы прогнозирования и модели распространения заболеваний. *Компьютерные исследования и моделирование.* 2013;5(5):863–882. DOI: 10.20537/2076-7633-2013-5-5-863-882.

2. Hirk R, Kastner G, Vana L. Investigating the dark figure of COVID-19 cases in Austria: borrowing from the decode genetics study in Iceland. *Austrian Journal of Statistics*. 2020;49(5):1–17. DOI: 10.17713/ajs.v49i4.1142.

3. Fanelli D, Piazza F. Analysis and forecast of COVID-19 spreading in China, Italy and France. *Chaos, Solitons and Fractals*. 2020;134:109841. DOI: 10.1016/j.chaos.2020.109761.

4. Kharin Yu. *Robustness in statistical forecasting*. New York: Springer; 2013. 356 p.

5. Kharin YuS, Voloshko VA, Medved EA. Statistical estimation of parameters for binary conditionally nonlinear autoregressive time series. *Mathematical Methods of Statistics*. 2018;27:103–118. DOI: 10.3103/S1066530718020023.

6. Волошко ВА, Харин ЮС. Семибиномиальные условно нелинейные авторегрессионные модели дискретных случайных последовательностей: вероятностные свойства и статистическое оценивание параметров. *Дискретная математика.* 2019; 31(1):72–98. DOI: 10.4213/dm1561.

7. Kharin Yu, Zhurak M. Analysis of spatio-temporal data based on Poisson conditional autoregressive model. *Informatica*. 2015; 26(1):67–87. DOI: 10.15388/Informatica.2015.39.

8. Maevskii VV, Kharin YuS. Robust regressive forecasting under functional distortions in a model. *Automation and Remote Control*. 2002;63(11):1803–1820. DOI: 10.1023/A:1020959432568.

9. Pashkevich MA, Kharin YuS. Robust estimation and forecasting for beta-mixed hierarchical models of grouped binary data. *Statistics and Operations Research Transactions*. 2004;28(2):125–160.

10. Большев ЛН, Смирнов НВ. *Таблицы математической статистики.* Москва: Наука; 1983. 512 с.

11. Харин ЮС, Зуев НМ, Жук ЕЕ. *Теория вероятностей, математическая и прикладная статистика.* Минск: БГУ; 2011. 465 с.

12. Харин ЮС. *Оптимальность и робастность в статистическом прогнозировании.* Минск: БГУ; 2008. 265 с.

13. Kharin Yu. Statistical analysis of discrete-valued time series by parsimonious high-order Markov chains. *Austrian Journal of Statistics*. 2020;49(4):76–88. DOI: 10.17713/ajs.v49i4.1132.

14. Kedem B, Fokianos K. *Regression models for time series analysis.* Wiley: Hoboken; 2002. 326 p.

15. Малюгин ВИ. *Методы анализа многомерных эконометрических моделей с неоднородной структурой.* Минск: БГУ; 2014. 351 с.

16. Colizza V, Barrat A, Barthelemy M, Vespignani A. Predictability and epidemic pathways in global outbreaks of infectious diseases: the SARS case study. *BMC Medicine*. 2007;5:34. DOI: 10.1186/1741-7015-5-34.

17. Engle RF, Granger CWJ. Co-integration and error correction: representation, estimation and testing. *Econometrica*. 1987;55(2): 251–276. DOI: 10.2307/1913236. JSTOR 1913236.

18. Kermack WO, McKendrick AG. Contributions to the mathematical theory of epidemics – I. *Bulletin of Mathematical Biology*. 1991;53(1–2):33–55. DOI: 10.1007/BF02464423.

19. Харин ЮС, Малюгин ВИ, Харин АЮ. *Эконометрическое моделирование.* Минск: БГУ; 2003. 313 с.

20. Johansen S. *Likelihood-based inference in cointegrated vector autoregressive models.* 2nd edition. Oxford: Oxford University Press; 1995. 267 p.

21. Харин АЮ. *Робастность байесовских и последовательных статистических решающих правил.* Минск: БГУ; 2013. 207 с.

22. Kharin A, Tu TT. Performance and robustness analysis of sequential hypotheses testing for time series with trend. *Austrian Journal of Statistics*. 2017;46(3–4):23–36. DOI: 10.17713/ajs.v46i3-4.668.

23. Kharin AYu. An approach to asymptotic robustness analysis of sequential tests for composite parametric hypotheses. *Journal of Mathematical Sciences*. 2017;227(2):196–203. DOI: 10.1007/s10958-017-3585-z.

References

1. Kondratyev MA. [Forecasting methods and models of disease spread]. *Komp'yuternye issledovaniya i modelirovanie*. 2013; 5(5):863–882. Russian. DOI: 10.20537/2076-7633-2013-5-5-863-882.

2. Hirk R, Kastner G, Vana L. Investigating the dark figure of COVID-19 cases in Austria: borrowing from the decode genetics study in Iceland. *Austrian Journal of Statistics*. 2020;49(5):1–17. DOI: 10.17713/ajs.v49i4.1142.

3. Fanelli D, Piazza F. Analysis and forecast of COVID-19 spreading in China, Italy and France. *Chaos, Solitons and Fractals*. 2020;134:109841. DOI: 10.1016/j.chaos.2020.109761.

4. Kharin Yu. *Robustness in statistical forecasting*. New York: Springer; 2013. 356 p.

5. Kharin YuS, Voloshko VA, Medved EA. Statistical estimation of parameters for binary conditionally nonlinear autoregressive time series. *Mathematical Methods of Statistics*. 2018;27:103–118. DOI: 10.3103/S1066530718020023.

6. Valoshka VA, Kharin YuS. [Semibinomial conditionally nonlinear autoregression models of discrete random sequences: probabilistic properties and statistical estimation of parameters]. *Diskretnaya matematika*. 2019;31(1):72–98. Russian. DOI: 10.4213/dm1561.

7. Kharin Yu, Zhurak M. Analysis of spatio-temporal data based on Poisson conditional autoregressive model. *Informatica*. 2015; 26(1):67–87. DOI: 10.15388/Informatica.2015.39.

8. Maevskii VV, Kharin YuS. Robust regressive forecasting under functional distortions in a model. *Automation and Remote Control*. 2002;63(11):1803–1820. DOI: 10.1023/A:1020959432568.

9. Pashkevich MA, Kharin YuS. Robust estimation and forecasting for beta-mixed hierarchical models of grouped binary data. *Statistics and Operations Research Transactions*. 2004;28(2):125–160.

10. Bol'shev LN, Smirnov NV. *Tablitsy matematicheskoi statistiki* [Mathematical statistics tables]. Moscow: Nauka; 1983. 512 p. Russian.

11. Kharin YuS, Zuev NM, Zhuk EE. *Teoriya veroyatnostei, matematicheskaya i prikladnaya statistika* [Probability theory, mathematical and applied statistics]. Minsk: Belarusian State University; 2011. 465 p. Russian.

12. Kharin YuS. *Optimal'nost' i robastnost' v statisticheskom prognozirovanii* [Optimality and robustness in statistical forecasting]. Minsk: Belarusian State University; 2008. 265 p. Russian.

13. Kharin Yu. Statistical analysis of discrete-valued time series by parsimonious high-order Markov chains. *Austrian Journal of Statistics*. 2020;49(4):76–88. DOI: 10.17713/ajs.v49i4.1132.

14. Kedem B, Fokianos K. *Regression models for time series analysis.* Wiley: Hoboken; 2002. 326 p.

15. Malugin VI. *Metody analiza mnogomernykh ekonometricheskikh modelei s neodnorodnoi strukturoi* [Methods for analyzing multivariate econometric models with a heterogeneous structure]. Minsk: Belarusian State University; 2014. 351 p. Russian.

16. Colizza V, Barrat A, Barthelemy M, Vespignani A. Predictability and epidemic pathways in global outbreaks of infectious diseases: the SARS case study. *BMC Medicine*. 2007;5:34. DOI: 10.1186/1741-7015-5-34.

17. Engle RF, Granger CWJ. Co-integration and error correction: representation, estimation and testing. *Econometrica*. 1987;55(2): 251–276. DOI: 10.2307/1913236. JSTOR 1913236.

18. Kermack WO, McKendrick AG. Contributions to the mathematical theory of epidemics – I. *Bulletin of Mathematical Biology*. 1991;53(1–2):33–55. DOI: 10.1007/BF02464423.

19. Kharin YuS, Malugin VI, Kharin AYu. *Ekonometricheskoe modelirovanie* [Econometric modeling]. Minsk: Belarusian State University; 2003. 313 p. Russian.

20. Johansen S. *Likelihood-based inference in cointegrated vector autoregressive models.* 2nd edition. Oxford: Oxford University Press; 1995. 267 p.

21. Kharin AYu. *Robastnost' baiesovskikh i posledovatel'nykh statisticheskikh reshayushchikh pravil* [Robustness of Bayesian and sequential statistical decision rules]. Minsk: Belarusian State University; 2013. 207 p. Russian.

22. Kharin A, Tu TT. Performance and robustness analysis of sequential hypotheses testing for time series with trend. *Austrian Journal of Statistics*. 2017;46(3–4):23–36. DOI: 10.17713/ajs.v46i3-4.668.

23. Kharin AYu. An approach to asymptotic robustness analysis of sequential tests for composite parametric hypotheses. *Journal of Mathematical Sciences*. 2017;227(2):196–203. DOI: 10.1007/s10958-017-3585-z.

Статья поступила в редколлегию 08.10.2020.
Received by editorial board 08.10.2020.