

ПОСТРОЕНИЕ И ВЕРИФИКАЦИЯ СТАТИСТИЧЕСКИХ РЕЙТИНГОВ, ХАРАКТЕРИЗУЮЩИХ ТИПОЛОГИЮ ПАНДЕМИИ COVID-19, С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ

В. И. Малюгин, А. К. Корниевич

Белорусский государственный университет, Минск, Беларусь
E-mail: Malugin@bsu.by, kosasha998@gmail.com

В статье представляются результаты решения следующих задач: статистическая классификация стран по интенсивности пандемии COVID-19; построение страновых и интегральных рейтингов, характеризующих интенсивность пандемии на страновом и много-страновом уровнях; оценка связи страновых рейтингов с темпами экономического роста. Предлагаемые статистические показатели пандемии COVID-19 могут быть полезны для оценки степени интенсивности эпидемического процесса, а также его влияния на экономические показатели стран.

CONSTRUCTION AND VERIFICATION OF STATISTICAL RATINGS CHARACTERIZING THE TYPOLOGY OF THE COVID-19 PANDEMIC USING MACHINE LEARNING ALGORITHMS

V. I. Malugin, A. K. Kornievich

The paper presents the results of solving the following problems: statistical classification of countries according to the intensity of the COVID-19 pandemic; construction country and integral ratings, that characterize the intensity of the pandemic at the country and multi-country levels and their statistical verification; assessment of the relationship between country ratings and countries economic growth rates. The proposed statistical indicators of the COVID-19 pandemic may be useful for assessing the intensity of the epidemic process, as well as its impact on the economic performance of countries.

1. Цели исследования и используемые данные

Проблеме анализа пандемии COVID-19 в различных аспектах отводится большое место в мировой научной литературе [1]. Одним из активно развивающихся направлений является разработка методов статистического моделирования и краткосрочного прогнозирования пандемии COVID-19 на уровне отдельных стран [2, 3]. Ряд исследований посвящены задачам анализа различных аспектов пандемии COVID-19 на много-страновом уровне [4]. Исследования, связанные со статистическим анализом, моделированием и прогнозированием эпидемического процесса в Республике Беларусь проводились в НИИ Прикладных проблем математики и информатики (НИИ ППМИ) Белорусского государственного университета на регулярной основе в течение первых двух волн эпидемии с апреля 2020 г. по декабрь 2021 г. [5].

В данной статье представляются новые результаты анализа пандемии COVID-19 в много-страновом аспекте. В работе используются ежедневные зна-

чения основных показателей пандемии с апреля 2020 г. по апрель 2022 г. для 30 стран евро-азиатского региона, включая: 21 страну Европейского Союза (Австрия, Болгария, Хорватия, Чехия, Дания, Эстония, Финляндия, Франция, Германия, Греция, Венгрия, Ирландия, Италия, Латвия, Литва, Нидерланды, Польша, Румыния, Словения, Испания, Швейцария), а также Азербайджан, Армению, Беларусь, Великобританию, Грузию, Казахстан, Молдову, Россию и Турцию.

В список доступных на регулярной основе показателей пандемии входят [5]:

- общее число подтвержденных случаев заражения (*Total Infections – $T(t)$*),
- число новых активных случаев заражения (*active Infections – $I(t)$*),
- число выздоровевших (*Recovered cases – $R(t)$*),
- число случаев смерти (*Deaths cases – $D(t)$*),
- число закрытых случаев заражения (*Closed cases – $C(t)=R(t) + D(t)$*).

Проводимые исследования, преследуют следующие основные цели:

- статистический много-страновой анализ с использованием алгоритмов машинного обучения интенсивности протекания пандемии COVID-19;
- оценка влияния типологии протекания эпидемического процесса на темпы экономического роста стран в период первой и второй волн пандемии.

2. Постановка задачи и используемый подход

Постановка задачи. Имеются ежедневные панельные данные, включающие значения N показателей эпидемического процесса для выборки n стран в моменты времени t :

$$x_{i,t} = (x_{i1,t}, \dots, x_{iN,t})' \in \mathfrak{R}^N \quad (i = 1, \dots, n, t = 1, \dots, T).$$

Предполагается, что панельные данные имеют неоднородную кластерную структуру, обусловленную скрытыми факторами, влияющими на интенсивность и характер протекания эпидемического процесса. По этому свойству страны можно отнести к одному из L классов. Номер класса описывается дискретной случайной величиной $d_{it} \in \{1, \dots, L\}$ для страны i в момент времени t . Номера классов $\{d_{it}\}$ интерпретируются как *страновые рейтинги эпидемического процесса*.

На основе исходных показателей пандемии COVID-19 в рамках проведенного исследования сформированы следующие классификационные признаки:

- 1) отношение количества закрытых случаев заражения к общему количеству зараженных (*Close to Total*);
- 2) отношение количества закрытых случаев к числу активных случаев заражения (отношение *Closed to Active*);
- 3) отношение текущего числа заражений к предыдущему (*Daily Rate*);
- 4) уровень смертности – доля умерших от общего числа заболевших (*Death Rate*).

Задача анализа типологии эпидемического процесса состоит в разбиении рассматриваемой выборки стран на $L \geq 2$ однородных подвыборок (*классов*) стран, различающихся в пространстве классификационных признаков по характеру и степени интенсивности эпидемиологического процесса.

Используемый подход. Поскольку обучающая классифицированная выборка наблюдений отсутствует и количество классов L неизвестно, для решения сформулированной выше задачи используются *алгоритмы машинного обучения*, предназначенные для статистической классификации многомерных данных в режиме самообучения. В [6, 7] для решения подобных задач был предложен подход, основанный на использовании алгоритмов кластерного анализа панельных данных в *пространственном (cross-sectional)* представлении.

Этот подход включает следующие шаги:

1) предварительный статистический анализ данных;
2) цензурирование данных и приведение значений классификационных признаков к интервалу $(0,1)$ таким образом, чтобы значения признаков, близкие к нулю, соответствовали более благоприятному течению эпидемического процесса и наоборот;

3) кластерный анализ панельных данных в пространственном представлении с помощью двух алгоритмов [8]:

- алгоритма иерархического кластерного анализа для оценивания числа классов L ;

- алгоритма L -средних для классификации наблюдений;

4) расчет и анализ статистических показателей пандемии и страновых рейтингов.

Решением сформулированной выше задачи анализа типологии пандемии с помощью алгоритмов кластерного анализа является оцененная классификационная матрица $D = \{d_{i,t}\} (i = 1, \dots, n, t = 1, \dots, T)$.

3. Результаты много-странового анализа типологии пандемии

Оценка качества классификационных признаков. На основе рис. 1 и табл. 1 можно сделать вывод о хорошей различительной (дискриминационной) способности используемых классификационных признаков на протяжении первой и второй волны эпидемии. Значения классификационных признаков (вертикальная ось), соответствующие центрам кластеров 1, 2 и 3, могут интерпретироваться как показатели среднестатистической страны из классов 1, 2 и 3 соответственно. При этом значения, близкие к нулю соответствуют классу 1 с наименьшей интенсивности эпидемического процесса.

Значения, близкие к единице, соответствуют классу 3 с наибольшей интенсивности эпидемического процесса. Класс 2 занимает промежуточное положение. Данные результаты согласуются с табл. 1: меньшие средние значения классификационных признаков соответствуют более благоприятному течению эпидемического процесса и наоборот.

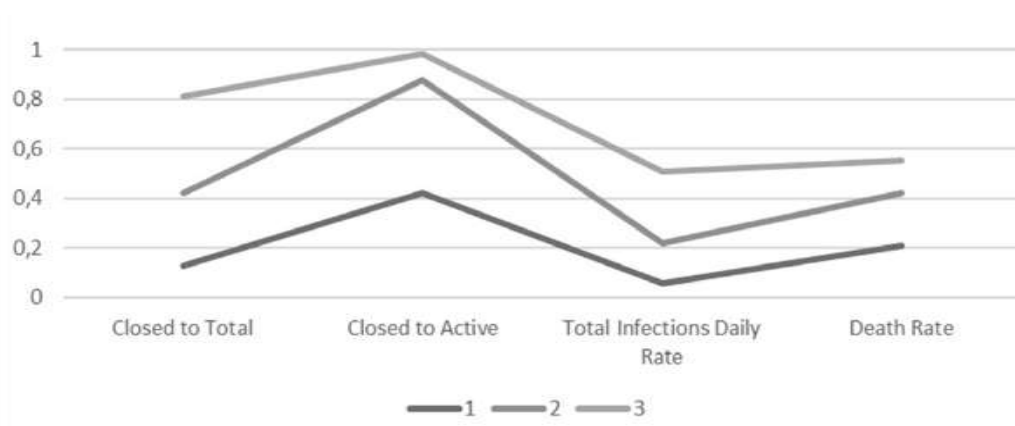


Рис. 1. Значения классификационных признаков для центров кластеров 1, 2, 3

Таблица 1

Средние значения классификационных признаков предлагаемые статистические показатели пандемии COVID-19

Кластер	Классификационные признаки			
	Closed to Total	Closed to Active	Total Infections Daily Rate	Death Rate
1	0.126	0.422	0.058	0.21
2	0.422	0.88	0.22	0.42
3	0.814	0.986	0.51	0.552

На основании оцененной классификационной матрицы $D = \{d_{i,t}\} (i = 1, \dots, n, t = 1, \dots, T)$ построены следующие статистические показатели COVID-19:

– ежедневный рейтинг страны i ($i = 1, \dots, n$) в момент времени t (*Daily Country Rating*):

$$DCR_{it} = d_{it} \in \{1, 2, 3\};$$

– средний рейтинг страны i за рассматриваемый временной интервал (*Average Country Rating*):

$$ACR_i = \frac{1}{T} \sum_{t=1}^T d_{it} \in [1, 3], i = 1, \dots, n,$$

– интегральный много-страновой индикатор пандемии COVID-19 (*Integral Multi-county Indicator*) в момент времени t :

$$IMI_t = \frac{1}{n} \sum_{i=1}^n d_{it} \in [1, 3], t = 1, \dots, T.$$

Верификация классов и страновых рейтингов. Для верификации связи страновых рейтингов с экономическими показателями на временном интервале с 5 марта 2020 г. до 18 апреля 2021 г. по рейтингу ACR сформированы три группы стран. В группы 1, 2 и 3 включаются страны с низкими (от 1 до 1.5), средними (начиная с 1.5 до 2.5) и высокими значениями ACR (начиная с 2.5 до 3) соответственно. Анализ рейтингов стран, отнесенных к различным группам, позволяет сделать следующие выводы: в первой группе (9 стран, включая Беларусь, Грузию, Турцию и 6 стран ЕС) с более низкой интенсивностью протека-

ния эпидемического процесса в рассматриваемый период доминирующими являются страновые рейтинги 1 и 2. Большинство стран отнесены ко второй группе (17 стран, включая Россию и 13 стран ЕС). Для этой группы стран имеет место различный характер эпидемического процесса в разные периоды времени с преобладанием средней степени интенсивности (рейтинг 2). Преобладающим для стран из третьей группы (4 страны – Великобритания, Венгрия, Франция, Италия) является рейтинг 3.

Оценка устойчивости типологии эпидемии в различных странах. Для оценивания устойчивости типологии эпидемии в различных странах проведено сравнение результатов классификации стран для двух временных интервалов: *a)* исходный интервал: апрель 2020 г.–апрель 2021 г.; *b)* пролонгированный интервал до апреля 2022 г. Согласно рис. 2, во второй год интенсивность эпидемии (начиная с апреля 2021 г.) в большинстве стран сохранилась на прежнем уровне либо снизилась, что обусловило снижение для них рейтинга АСР в целом. Для отдельных стран в этот период наблюдался рост интенсивности эпидемии.

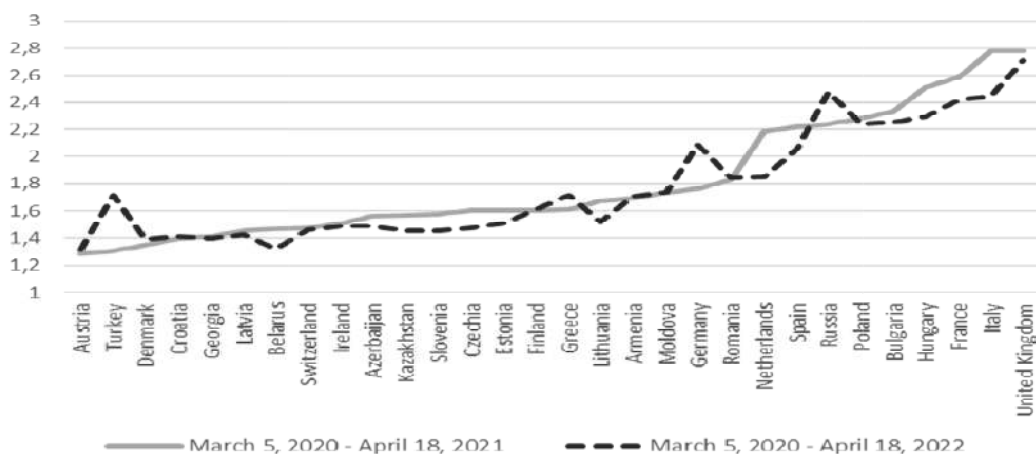


Рис. 2. Значения рейтинга АСР: интервал *a* (непрерывная линия), интервал *b* (пунктирная линия)

Верификация интегрального индикатора. Интегральный индикатор пандемии ІМІ принимает значения на интервале от 1 до 3 и описывает общую динамику пандемии COVID-19 в рассматриваемом регионе. В силу существенных различий в характере протекания эпидемии в разных странах, он имеет иррегулярную случайную периодичность. Согласно рис. 3, индикатор ІМІ находится в интервале от 1,5 до 2,1 пунктов и имеют тенденцию к снижению, начиная с апреля 2021 г.

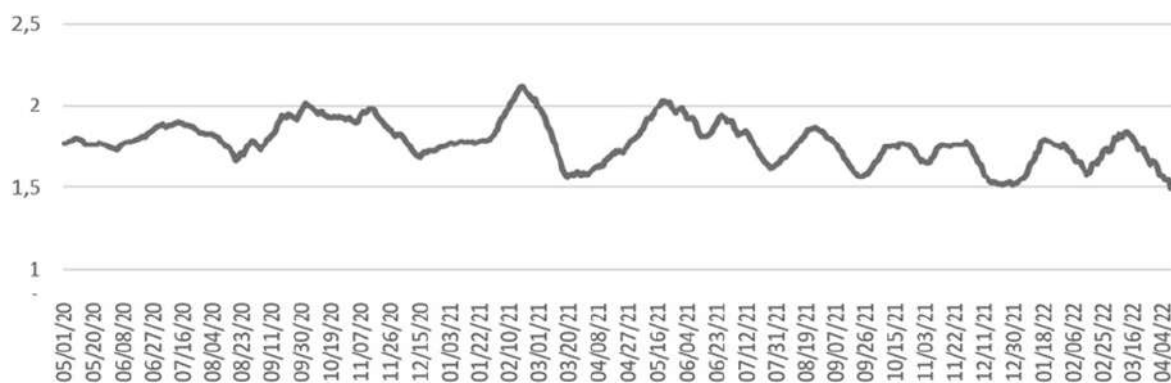


Рис. 3. Динамика индикатора ІМІ:
март 2020 – апрель 2022 г., 30-дневное сглаживание

4. Оценка влияния COVID-19 на экономический рост

В табл. 2 представлены средние значения годовых темпов роста ВВП для ранее полученных групп стран 1, 2 и 3 с относительно низкой, средней и высокой интенсивностью эпидемического процесса соответственно [9], а на рис. 4 – динамика темпов роста реального ВВП Беларуси, России и ЕС на конец 2020 г.

Таблица 2

Годовые темпы снижения роста ВВП
для трех групп стран (на конец 2020 г.)

Группа	Годовые темпы роста ВВП (переменная GGDP)
1	-3,610
2	-4,063
3	-7,716

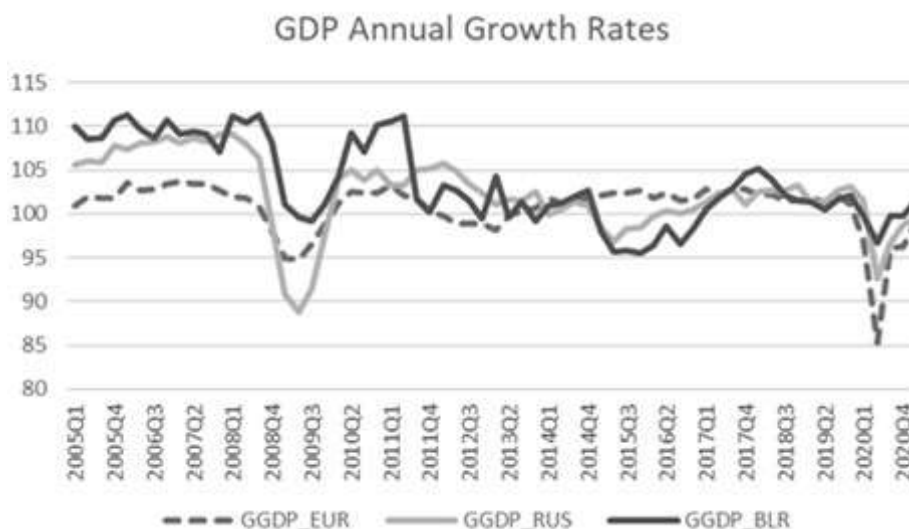


Рис. 4. Влияние COVID-19 на темпы роста реального ВВП в Беларуси, России и ЕС

Анализ годовых темпов роста ВВП на конец 2020 г. для Беларуси (первая группа), России (вторая группа) и ЕС свидетельствует о том, что более значительный спад экономической активности наблюдался в странах Европейского

союза, большинство из которых относятся к группам 2 и 3.

Заключение

Проведенный анализ типологии COVID-19 на основе официально зарегистрированных случаев заражения можно сделать следующие выводы:

- в странах с более высокой интенсивности эпидемического процесса в целом наблюдается больший спад экономической активности;
- поскольку интенсивность эпидемического процесса в многом обусловлена проводимой в стране антиковидной политикой, то анализ страновых рейтингов может в определенной степени характеризовать эффективность принимаемых антиковидных мер в рассматриваемый период времени;
- предложенные статистические показатели пандемии COVID-19 могут быть полезны для оценки степени интенсивности эпидемического процесса, а также его влияния на экономические показатели стран.

СПИСОК ЛИТЕРАТУРЫ

1. *Cao Longbing, Liu Qing*. COVID-19 Modeling: A Review. *SSRN papers*, 2021. [Electronic resource]. URL: <https://ssrn.com/abstract=3899127> (date of access: 10.10.2022).
2. *Doornik J. A., Castle J. L. Hendry D. F.* Short-term forecasting of the coronavirus pandemic // *International Journal of Forecasting*. 2022. Vol. 38/2. P. 453-466.
3. *Fantazzini Dean*. Short-term forecasting of the COVID-19 pandemic using Google Trends data: Evidence from 158 countries // *Applied Econometrics, Russian Presidential Academy of National Economy and Public Administration (RANEPA)*. Vol. 59. P. 33-54.
4. *Jang S. Y., Hussain-Alkhateeb L., Rivera Ramirez, T. et al.* Factors shaping the COVID-19 epidemic curve: a multi-country analysis. *BMC Infect Dis* 21, 2020. [Electronic resource]. URL: <https://doi.org/10.1186/s12879-021-06714-3>. (date of access: 11.09.2022).
5. *Харин Ю. С., Волошко В. А., Дернакова О. В., Малюгин В. И., Харин А. Ю.* Вероятностно-статистическое моделирование и прогнозирование эпидемиологических процессов COVID-19 в Республике Беларусь // *Журнал Белорусского государственного университета. Математика, информатика*. 2020. № 3. С. 36-50.
6. *Worldometers.info* Mode of access. [Electronic resource]. URL: <https://www.worldometers.info/coronavirus>. (date of access: 27.02.2022).
7. *Malugin V. I., Hryn N. V., Novopoltsev A. Yu.* Statistical analysis and econometric modeling of the creditworthiness of non-financial companies // *Int. J. Computational Economics and Econometrics*. 2014. Vol. 4 (1/2). P. 130-147.
8. *Малюгин В. И., Гринь Н. В., Милевский П. С.* Система статистических кредитных рейтингов предприятий: методика построения, верификации и применения // *Банковский Вестник. Исследования банка*. 2013. № 5. 73 с.
9. *Харин Ю. С., Малюгин В. И., Абрамович М. С.* Математические и компьютерные основы статистического моделирования и анализа данных. Минск: БГУ, 2008. 455 с.
10. *The World Bank Group* Mode of access: [Electronic resource]. URL: <https://data.worldbank.org/> (date of access: 02.03.2022).