
ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ИНФОРМАТИКИ

THEORETICAL FOUNDATIONS OF COMPUTER SCIENCE

УДК 004.4

МЕТОДЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ В ИССЛЕДОВАНИЯХ ЭПИДЕМИИ COVID-19

**О. В. СЕНЬКО^{1), 2)}, А. В. КУЗНЕЦОВА^{1), 3)}, Е. М. ВОРОНИН¹⁾,
О. А. КРАВЦОВА^{1), 4)}, Л. Р. БОРИСОВА⁵⁾, И. Л. КИРИЛЮК⁶⁾, В. Г. АКИМКИН¹⁾**

¹⁾Центральный научно-исследовательский институт эпидемиологии Роспотребнадзора,
ул. Новогиреевская, 3а, 111123, г. Москва, Россия

²⁾Федеральный исследовательский центр «Информатика и управление»
Российской академии наук, ул. Вавилова, 44, корп. 2, 119333, г. Москва, Россия

³⁾Институт биохимической физики им. Н. М. Эмануэля
Российской академии наук, ул. Косыгина, 4, 119334, г. Москва, Россия

⁴⁾Московский государственный университет им. М. В. Ломоносова,
Ленинские горы, 1, 119991, г. Москва, Россия

⁵⁾Финансовый университет при Правительстве Российской Федерации,
пр. Ленинградский, 49/2, 125167, г. Москва, Россия

⁶⁾Институт экономики Российской академии наук, пр. Нахимовский, 32, 117218, г. Москва, Россия

Представлен оригинальный метод поиска связи хода эпидемии с социально-экономическими, демографическими и климатическими факторами. В рамках предложенного метода проведена иерархическая агломеративная кластеризация 110 стран мира по кривым темпа роста COVID-19 за период с января 2020 по август 2021 г. Выделены четыре крупных кластера с единообразными кривыми, включающих 11, 39, 17 и 13 стран соответственно. Еще 30 стран не вошли ни в один из кластеров. Методами машинного обучения в выделенных кластерах выявлены различия социально-экономических, демографических и географо-климатических показателей. Наиболее важными показателями, по которым кластеры отличаются друг от друга, стали амплитуда температур в течение года,

Образец цитирования:

Сенько ОВ, Кузнецова АВ, Воронин ЕМ, Кравцова ОА, Борисова ЛР, Кирилук ИЛ, Акимкин ВГ. Методы интеллектуального анализа данных в исследованиях эпидемии COVID-19. *Журнал Белорусского государственного университета. Математика. Информатика*. 2022;1:83–96.
<https://doi.org/10.33581/2520-6508-2022-1-83-96>

For citation:

Senko OV, Kuznetsova AV, Voronin EM, Kravtsova OA, Borisova LR, Kirilyuk IL, Akimkin VG. Methods of intellectual data analysis in COVID-19 research. *Journal of the Belarusian State University. Mathematics and Informatics*. 2022;1:83–96. Russian.
<https://doi.org/10.33581/2520-6508-2022-1-83-96>

Сведения об авторах см. на с. 95–96.

Information about the authors see p. 95–96.



высокотехнологичный экспорт, коэффициент Джини, численность городского населения и населения в целом, индекс чистых бартерных условий торговли, рост населения, средняя температура января, территория (площадь суши), количество погибших в результате стихийных бедствий, коэффициент рождаемости, длина береговой линии, запасы нефти, доля населения в городских агломерациях с численностью населения более 1 млн человек и др. Данный подход (применение кластеризации в сочетании с классификацией методами логико-статистического анализа) ранее никем не использовался. Найденные закономерности позволят более точно проводить прогнозирование эпидемиологического процесса в странах, принадлежащих к разным кластерам. Дополнение представленного подхода авторегрессионными моделями позволит автоматизировать прогноз и повысить его точность.

Ключевые слова: кластерный анализ; методы машинного обучения; статистика; эпидемиологический процесс; COVID-19.

METHODS OF INTELLECTUAL DATA ANALYSIS IN COVID-19 RESEARCH

O. V. SENKO^{a, b}, A. V. KUZNETSOVA^{a, c}, E. M. VORONIN^a,
O. A. KRAVTSOVA^{a, d}, L. R. BORISOVA^c, I. L. KIRILYUK^f, V. G. AKIMKIN^a

^aCentral Research Institute of Epidemiology, Federal Service for Surveillance on Consumer Rights
Protection and Human Wellbeing, 3a Novogireevskaya Street, Moscow 111123, Russia

^bFederal Research Center «Computer Science and Control», Russian Academy of Sciences,
44 Vavilova Street, 2 building, Moscow 119333, Russia

^cInstitute of Biochemical Physics named after N. M. Emanuel, Russian Academy of Sciences,
4 Kosygina Street, Moscow 119334, Russia

^dLomonosov Moscow State University, 1 Leninskie Gory, Moscow 119991, Russia

^eFinancial University under the Government of the Russian Federation,
49/2 Leningradskii Avenue, Moscow 125167, Russia

^fInstitute of Economics, Russian Academy of Sciences,
32 Nakhimovskii Avenue, Moscow 117218, Russia

Corresponding author: A. V. Kuznetsova (azforus@yandex.ru)

The paper presents an original method for solving the problem of finding a connection between the course of the epidemic and socio-economic, demographic and climatic factors. The method was applied to solve this problem for 110 countries of the world using a set of corresponding curves of the COVID-19 growth rate for the period from January 2020 to August 2021. Hierarchical agglomerative clustering was applied. Four large clusters with uniform curves were identified – 11, 39, 17 and 13 countries, respectively. Another 30 countries were not included in any cluster. Using machine learning methods, we identified the differences in socio-economic, demographic and geographical and climatic indicators in the selected clusters of countries of the world. The most important indicators by which the clusters differ from each other are amplitude of temperatures throughout the year, high-tech exports, Gini coefficient, size of the urban population and the general population, index of net barter terms of trade, population growth, average January temperature, territory (land area), number of deaths due to natural disasters, birth rate, coastline length, oil reserves, population in urban agglomerations with a population of more than 1 million etc. This approach (the use of clustering in combination with classification by methods of logical-statistical analysis) has not been used by anyone before. The found patterns will make it possible to more accurately predict the epidemiological process in countries belonging to different clusters. Supplementing this approach with autoregressive models will automate the forecast and improve its accuracy.

Keywords: cluster analysis; machine learning methods; statistics; epidemiological process; COVID-19.

Введение

Хорошо известно, что масштабные эпидемии могут оказывать значительное влияние на социально-экономические и демографические показатели [1]. Менее изучено обратное влияние данных показателей на развитие эпидемий. Целью настоящей работы является исследование влияния на эпидемиологический процесс социально-экономических, демографических и климатических факторов, которые далее мы будем называть внешними факторами. Необходимость учета климатических факторов связана с описанным в литературе влиянием сезонности и непосредственно температуры воздуха на заболеваемость острыми респираторными заболеваниями [2]. Знание объективных закономерностей связи хода эпидемиологического процесса с перечисленными внешними факторами, несомненно, должно способствовать пониманию механизмов развития эпидемий и выработке эффективных профилактических мер.

Задачей исследования было создание эффективной информационной технологии для поиска статистически значимых закономерностей, связывающих ход эпидемиологического процесса с внешними факторами. Одним из путей поиска подобных закономерностей является изучение связи внешних факторов с такими глобальными показателями, как общее число случаев заболеваний и (или) общая смертность по совокупности стран мира.

Однако такой анализ затрудняется различными по странам стандартами сбора эпидемиологических данных и динамическим характером процесса. Наиболее полно эпидемиологический процесс описывается кривыми, показывающими зависимость ежесуточного прироста числа случаев заболевания от времени.

Альтернативный способ выявления связи между внешними факторами и динамикой эпидемиологического процесса основан на выделении групп стран со сходной формой эпидемических кривых при значительных различиях между группами. Далее выделенные группы сравниваются между собой по перечисленным внешним факторам с использованием статистических тестов и других средств сравнения групп по наборам показателей в целях выявления статистически значимых различий. Для выделения в совокупности данных групп объектов, имеющих похожие описания, обычно применяются методы кластерного анализа. Поскольку оптимальное число выделяемых групп заранее неизвестно, а общее число объектов невелико, то наиболее подходящей технологией кластерного анализа является иерархическая агломеративная кластеризация. Важная характеристика алгоритма кластеризации – используемая мера сходства между объектами.

Сходство двух эпидемических кривых в смысле синхронности подъема или затухания эпидемии хорошо оценивается с помощью стандартного коэффициента корреляции Пирсона между одновременными суточными приростами заболеваемости. Сохранение коэффициента корреляции Пирсона между двумя переменными при произвольных линейных преобразованиях этих переменных позволяет предположить относительную устойчивость данной меры к методам определения суточного прироста заболеваемости, применяемым в разных странах. В настоящей работе для установления связи между внешними факторами и динамикой эпидемиологического процесса используется подход, основанный на выделении групп стран с единообразной формой эпидемических кривых в смысле меры сходства, задаваемой коэффициентами корреляции Пирсона. Затем выделенные группы сравниваются между собой по перечисленным внешним факторам в целях выявления статистически значимых различий. Дополнительным преимуществом предлагаемого подхода является возможность применения его непосредственно для краткосрочного прогнозирования.

Методы кластерного анализа широко используются для анализа данных, связанных с развитием эпидемии COVID-19, различными исследователями. Так, в работе [3] выделялись группы районов Индии, однородные по показателям плотности населения, количеству госпиталей для пациентов с COVID-19, числу подтвержденных случаев COVID-19. В исследовании [4] кластеризация проводилась по эпидемическим кривым с использованием евклидова расстояния для оценки различия между кривыми. В статье [5] территориальные кластеры внутри материкового Китая выделялись с применением коэффициента пространственной корреляции Морана, связь с метеорологическими, экологическими и социально-экономическими факторами устанавливалась с помощью линейной регрессии с географической привязкой. В работе [6] использовалась иерархическая кластеризация (а именно *average-link clustering*) на данных о заболеваемости и смертности и оценивался индекс перехода для прогнозирования тенденции новой волны заболеваемости как расстояние между ближайшими кластерами. В публикации [7] применялся метод *k*-средних для разбиения стран на группы, где набор признаков включает социальные, экономические и связанные со здоровьем и окружающей средой показатели, а также измерялся коэффициент корреляции Пирсона между эпидемическими кривыми заболеваемости или смертности и выбранными характеристиками. В исследовании [8] проводилась кластеризация стран ЕС с использованием метода Уорда и метода *k*-средних.

Таким образом, рассматриваемый в настоящей работе подход отличается от способов использования кластерного анализа для исследования развития эпидемии COVID-19 в работах [3–8].

Материалы и методы исследования

Исходными данными для кластеризации стали кривые темпа роста COVID-19 за период с января 2020 по август 2021 г.¹

В базу данных для классификации по кластерам вошли показатели 110 стран. В базе социально-экономического статуса стран и погодно-климатических условий было 138 показателей.

¹Данные о заболеваемости COVID-19 по странам мира получены из источника: https://datalens.yandex/7o7is1q6ikh23?tab=0Ze&utm_source=cbmain&state=6057afc92382.

Анализ данных включал следующие этапы.

Этап 1. Вычислялась мера сходства между всевозможными парами эпидемических кривых. Более точный учет сходства или различия между эпидемическими кривыми S_i и S_j достигался через подбор лага l из отрезка $[0, 20]$, при котором коэффициент корреляции Пирсона оказывался максимальным. С этой целью для каждого l вычислялись коэффициент корреляции ρ_l^+ между рядами $S_i(0), \dots, S_i(n-l)$ и $S_j(l), \dots, S_j(n)$ и коэффициент корреляции ρ_l^- между рядами $S_i(l), \dots, S_i(n)$ и $S_j(0), \dots, S_j(n-l)$. Максимальный коэффициент корреляции из набора $\rho_0^+, \rho_0^-, \dots, \rho_{20}^+, \rho_{20}^-$ использовался в качестве меры близости $\rho(S_i, S_j)$ между эпидемическими кривыми S_i и S_j .

Этап 2. Проводилась иерархическая агломеративная кластеризация. В качестве меры сходства двух групп эпидемических кривых G' и G'' использовалось среднее значение меры сходства между эпидемическими кривыми из разных групп:

$$P(G', G'') = \frac{1}{m' m''} \sum_{i=1}^{m'} \sum_{j=1}^{m''} \rho(C_i, C_j).$$
 Процесс слияния кластеров прекращался, если мера сходства P между любыми двумя кластерами в текущей кластеризации не оказывалась ниже 0,5.

Этап 3. Производилось сравнение полученных кластеров по перечисленным выше внешним показателям. С этой целью применялся стандартный тест Манна – Уитни, а также метод оптимальных достоверных разбиений (ОДР).

Метод ОДР, описанный ранее в работе [9], основан на поиске разбиений интервалов значений отдельных признаков или двумерных областей допустимых значений пар признаков, позволяющих достичь максимального разделения сравниваемых групп. Разбиения строятся по выборке $\tilde{S} = \{(y_1, x_1), \dots, (y_m, x_m)\}$, где $y_1, \dots, y_m$ – значения бинарной переменной Y , являющейся индикатором сравниваемых групп G_1 и G_2 ; $x_1, \dots, x_m$ – векторы значений потенциальных объясняющих переменных $X_1, \dots, X_N$. Предположим, что имеется разбиение R , состоящее из k элементов $q_1, \dots, q_k$. Качество разделения групп G_1 и G_2 разбиением R оценивается с помощью функционала

$$F(\tilde{S}, R) = \sum_{i=1}^k (v_i - v_0)^2 m_i,$$
 где m_i – число объектов из выборки $\tilde{S}$; v_0 – доля объектов из группы G_1 во всей выборке $\tilde{S}$; v_i – доля объектов из группы G_1 внутри элемента q_i . Разбиение R_0 считается оптимальным, если значение функционала $F(\tilde{S}, R)$ для него оказывается не ниже значения $F(\tilde{S}, R)$ для произвольного разбиения R внутри семейства $\tilde{R}$, по которому производится поиск.

Поиск осуществляется внутри трех семейств разбиений, имеющих разный уровень сложности. Простейшее семейство I состоит из разбиений интервалов значений отдельных признаков с одной граничной точкой, семейство II – из разбиений интервалов значений отдельных признаков с двумя граничными точками, семейство III – из разбиений двумерной области допустимых значений пары признаков на четыре подобласти с границами, параллельными координатным осям.

Процедура статистической верификации, основанная на перестановочном тесте, позволяет оценить значимость различий между группами по отдельным признакам, а также значимость различий по каждому из признаков для двумерных разбиений.

Перестановочный тест базируется на проверке нулевой гипотезы о независимом от объясняющих переменных процессе генерации значений целевой переменной Y . Будем предполагать также, что значения переменной Y порождаются независимо из распределения, тождественного маргинальному эмпирическому распределению Y на выборке $\tilde{S}$. Сгенерируем значения переменной Y для каждого вектора объясняющих переменных согласно рассматриваемому процессу. В результате получим новую случайную выборку $\tilde{S}^r = \{(y_1^r, x_1), \dots, (y_m^r, x_m)\}$. Введем дополнительные ограничения на описанный процесс генерации, используя дополнительную селекцию случайных выборок с требованием равенства маргинальных распределений $\{y_1^r, \dots, y_m^r\}$ исходному маргинальному распределению $\{y_1, \dots, y_m\}$. Предположение о генерации данных последним процессом составляет нулевую гипотезу перестановочного теста, которая также эквивалентна гипотезе о генерации случайных выборок с помощью случайных перестановок значений Y относительно фиксированных позиций векторов $\bar{x}_1, \dots, \bar{x}_m$. То есть случайная выборка будет представлена в виде $\{(y_{f(1)}, \bar{x}_1), \dots, (y_{f(m)}, \bar{x}_m)\}$, где $\{f(1), \dots, f(m)\}$ – случайная перестановка натуральных чисел из множества $\{1, \dots, m\}$.

В качестве статистики критерия в перестановочном тесте используется функция от данных, которая количественно описывает выраженность связи между целевой переменной Y и объясняющими переменными.

В методе ОДР статистикой критерия выступает величина $F(\tilde{S}, R_0)$, т. е. значение функционала $F(\tilde{S}, R)$ на оптимальном разбиении R_0 .

Подсчет p -значений при использовании простейшей одномерной модели производится исходя из проверки описанной выше нулевой гипотезы о полной независимости Y от объясняющих переменных. Предположим, что сгенерировано множество $\tilde{f}_N$ из N случайных перестановок натуральных чисел из набора $\{1, \dots, m\}$. Тогда оценка p -значения может быть вычислена по формуле

$$p = \frac{\sum_{f \in \tilde{f}_N} I\left[F(\tilde{S}_f^r, R_0^f) \geq F(\tilde{S}, R_0) \mid \tilde{S}_f^r = \left\{\left(y_{f(1)}, \bar{x}_1\right), \dots, \left(y_{f(m)}, \bar{x}_m\right)\right\}\right]}{N},$$

где R_0^f – оптимальное разбиение для выборки $\tilde{S}_f^r$.

Для более сложных семейств II и III вычисление p -значений осложняется требованием полной верификации моделей. Невыполнение данного требования обычно приводит к появлению на выходе метода ОДР чрезмерно большого числа избыточных разбиений [9].

Модель считается полностью верифицированной, если статистическую верификацию на достаточном уровне значимости прошли все ее элементы. Для верификации элементов разбиений в методе ОДР используется принцип бритвы Оккама, трактуемый как возможность принятия более сложной модели, если она позволяет опровергнуть нулевую гипотезу об исчерпывающем описании существующих в анализируемых данных закономерностей с помощью более простой модели [10].

Этап 4. Для интегральной оценки различий между сравниваемыми кластерами применялись методы распознавания. Нами использовался набор из девяти методов распознавания, включая такие известные статистические методы, как метод k -ближайших соседей, линейный дискриминант Фишера, логистическая регрессия; метод опорных векторов; ансамбли решающих деревьев, генерируемые с помощью процедур адаптивного и градиентного бустинга; нейронные сети; оригинальный авторский метод – статистически взвешенные синдромы (СВС) [11].

Метод СВС основан на процедуре взвешенного голосования по системам так называемых синдромов – областей признакового пространства, содержащих преимущественно объекты одного из классов.

В качестве интегральной меры различий между кластерами по наборам внешних факторов применялась величина ROC AUC для коллективного решения (ансамбля). Использование интегральной меры позволяет ранжировать различия между кластерами по уровню их выраженности. Причиной использования коллективных решений является их более высокая обобщающая способность, что связано со взаимной компенсацией ошибок различных алгоритмов.

Результаты и их обсуждение

Результаты кластеризации. С использованием описанного выше метода проведена кластеризация 110 стран мира по кривым суточного прироста случаев заболеваний. В итоге были выявлены четыре крупных кластера, каждый из которых содержал не менее 11 стран. Всего в крупные кластеры вошло 80 стран. Еще 30 стран сформировали мелкие кластеры, включающие не более 4 стран, или не образовали кластеров вообще.

Представление общих кривых заболеваемости по кластерам. В 1-й кластер вошли 11 стран – Великобритания, Россия, США, Португалия, Панама, Нигерия, Мексика, Израиль, Ирландия, ОАЭ, ЮАР. Как видно, данный кластер не имеет явной географической локализации.

Сглаженная усредненная эпидемическая кривая для 1-го кластера приведена на рис. 1. Как следует из рис. 1, высокая интенсивность эпидемии отмечалась с октября 2020 по март 2021 г.

Во 2-й кластер вошли 39 стран – Австрия, Германия, Италия, Сербия, Болгария, Румыния, Северная Македония, Венгрия, Иордания, Польша, Украина, Канада, Молдова, Сирия, Босния и Герцеговина, Норвегия, Финляндия, Албания, Эстония, Черногория, Ливан, Словакия, Чехия, Азербайджан, Литва, Хорватия, Грузия, Дания, Беларусь, Латвия, Словения, Нидерланды, Армения, Люксембург, Бельгия, Кения, Швейцария, Швеция, Палестина. Географически они локализованы главным образом в Европе. Из рис. 2 видно существование двух выраженных пиков на интервале с октября 2020 по май 2021 г.

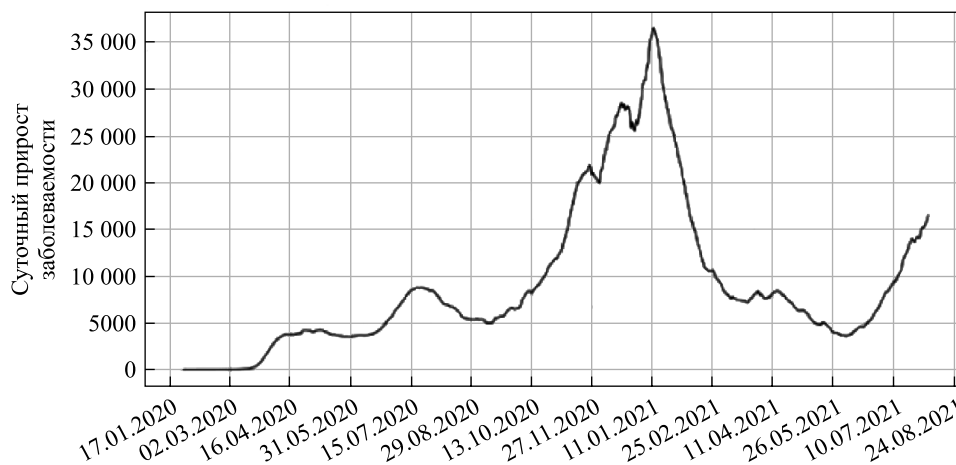


Рис. 1. Сглаженная с использованием метода скользящего среднего усредненная эпидемическая кривая для 1-го кластера на интервале от 01.02.2020 до 04.08.2021

Fig. 1. Smoothed using the moving average method, the averaged epidemic curve for the 1st cluster in the interval from 01.02.2020 to 04.08.2021

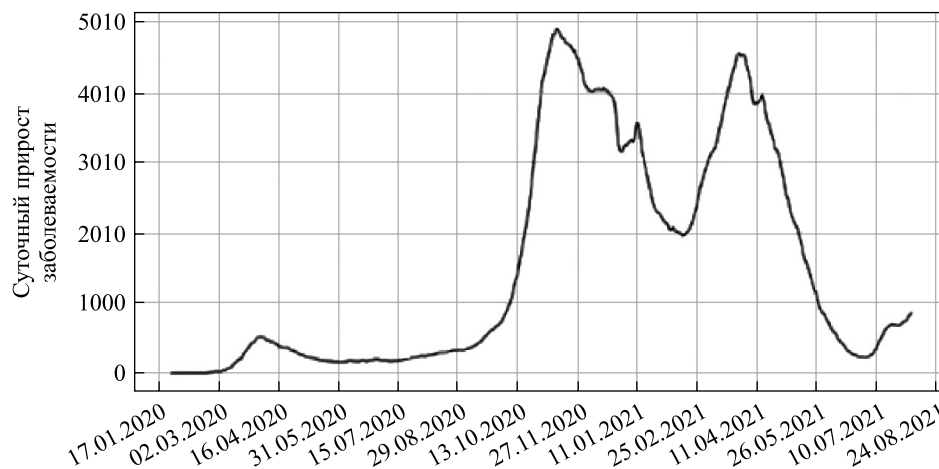


Рис. 2. Сглаженная с использованием метода скользящего среднего усредненная эпидемическая кривая для 2-го кластера на интервале от 01.02.2020 до 04.08.2021

Fig. 2. Smoothed using the moving average method, the averaged epidemic curve for the 2nd cluster in the interval from 01.02.2020 to 04.08.2021

В 3-й кластер (рис. 3) вошли 17 стран – Алжир, Марокко, Бангладеш, Вьетнам, Индонезия, Куба, Малайзия, Таиланд, Лаос, Мьянма, Тунис, Ливия, Греция, Иран, Кипр, Япония, Республика Корея. Преимущественно это страны с тропическим климатом.

В 4-й кластер вошли 13 стран – Аргентина, Колумбия, Парагвай, Уругвай, Камбоджа, Шри-Ланка, Бразилия, Ирак, Кувейт, Филиппины, Индия, Непал, Монголия. Значительная часть стран из 4-го кластера географически локализована в Южной Америке. Из рис. 4 можно отметить высокую интенсивность эпидемии с апреля 2020 по август 2021 г.

В малые кластеры (отделены точкой с запятой) вошли следующие группы стран: Катар, Саудовская Аравия, Сингапур, Таджикистан; Боливия, Египет, Чили; Гватемала, Коста-Рика, Оман; Мадагаскар, Перу, Эфиопия; Казахстан, Турция. Не образовали кластеров вообще 15 стран – Австралия, Гондурас, Исландия, Испания, Йемен, Кыргызстан, Китай, Никарагуа, Новая Зеландия, Пакистан, Судан, Танзания, Узбекистан, Франция, Эквадор.

Метод ОДР и критерий Манна – Уитни были использованы для сравнения кластеров между собой по перечисленным внешним факторам. На рис. 5 приведен пример одномерного разбиения из семейства I, разделяющего страны из 1-го и 2-го кластеров по индексу Джини, являющемуся показателем степени расслоения общества для некоторой страны.

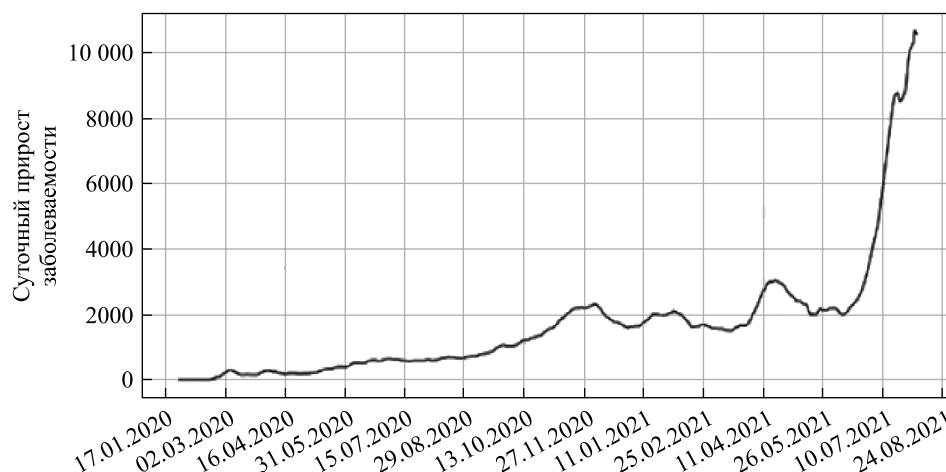


Рис. 3. Сглаженная с использованием метода скользящего среднего усредненная эпидемическая кривая для 3-го кластера на интервале от 01.02.2020 до 04.08.2021

Fig. 3. Smoothed using the moving average method, the averaged epidemic curve for the 3rd cluster in the interval from 01.02.2020 to 04.08.2021

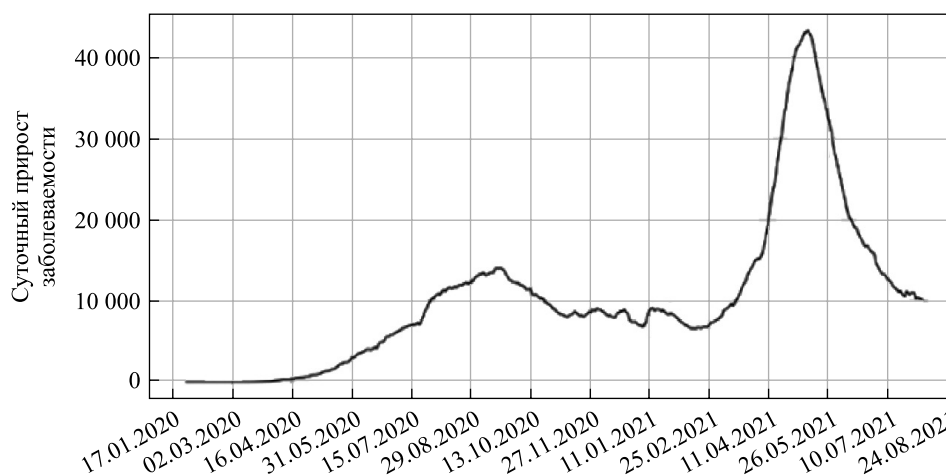


Рис. 4. Сглаженная с использованием метода скользящего среднего усредненная эпидемическая кривая для 4-го кластера на интервале от 01.02.2020 до 04.08.2021

Fig. 4. Smoothed using the moving average method, the averaged epidemic curve for the 4th cluster in the interval from 01.02.2020 to 04.08.2021

Значимость различий между кластерами (классами распознавания) оценивается с помощью перестановочного теста на уровне $p = 0,0005$.

Левее границы 36,95 находятся 29 стран из 2-го кластера (90,6 %) и только 3 страны из 1-го кластера. Справа от этой границы находятся 7 стран из 1-го кластера и лишь 3 страны из 2-го кластера (9,4 %). Результаты сравнения 1-го и 2-го кластеров с использованием однофакторной модели из семейства I приведены в табл. 1.

В табл. 1 представлены оптимальные границы, рассчитанные с помощью метода ОДР. В столбце «квадрант I» даны количества объектов 1-го и 2-го кластеров со значением соответствующего показателя слева от границы, в столбце «квадрант II» – количества объектов 1-го и 2-го кластеров со значением соответствующего показателя справа от границы. В скобках указана их доля в процентах.

В столбце «ПТ» представлены p -значения, рассчитанные с помощью перестановочного теста. Все показатели ранжированы по p -значению перестановочного теста. На первом месте, как видно из табл. 1, находятся коэффициент Джини и доля населения в городских агломерациях с численностью населения более 1 млн человек.

В столбце «U-тест» представлены p -значения, полученные с помощью стандартного теста Манна – Уитни.



Рис. 5. Оптимальное разделение по коэффициенту Джини.
Крестиками обозначены страны 1-го кластера, кружочками – страны 2-го кластера
Fig. 5. Optimal partition of Gini index ranges.
The crosses indicate the countries of the 1st cluster, circles – the countries of the 2nd cluster

Таблица 1

Одномерные разбиения для распознавания 1-го и 2-го кластеров стран мира

Table 1

One-dimensional partitions for recognising the 1st and 2nd clusters of countries in the world

№ п/п	Показатели	Граница	Квадрант I		Квадрант II		F	ПТ	U-тест
			1-й кластер	2-й кластер	1-й кластер	2-й кластер			
1	Коэффициент Джини	36,95	3 (30 %)	29 (90,6 %)	7 (70 %)	3 (9,4 %)	15,069	0,0005	0,1073
2	Доля населения в городских агломерациях с численностью населения более 1 млн чел. в 2020 г., %	35,695	4 (36,4 %)	24 (92,3 %)	7 (63,6 %)	2 (7,7 %)	12,787	0,002	0,5258
3	Последствия стихийных бедствий (землетрясения, засухи, наводнения, штормы), всего погибших в 1993–2002 гг.	369	5 (50 %)	31 (100 %)	5 (50 %)	0 (0 %)	17,222	0,003	0,6425
4	Береговая линия, км	694,5	1 (9,1 %)	25 (71,4 %)	10 (90,9 %)	10 (28,6 %)	12,947	0,003	0,1804
5	Амплитуда температур в течение года (max – min)	13,735	4 (66,7 %)	0 (0 %)	2 (33,3 %)	19 (100 %)	14,476	0,004	0,06171
6	Население в 2020 г., млн чел.	106,0699	7 (63,6 %)	35 (100 %)	4 (36,4 %)	0 (0 %)	13,636	0,005	0,03717
7	Индекс чистых бартерных условий торговли (2000 = 100)	154,0499	7 (63,6 %)	35 (100 %)	4 (36,4 %)	0 (0 %)	13,636	0,005	0,002466

Продолжение табл. 1
Continuation of the table 1

№ п/п	Показатели	Граница	Квадрант I		Квадрант II		F	ПТ	U-тест
			1-й кластер	2-й кластер	1-й кластер	2-й кластер			
8	Население в 2014 г., млн чел.	46,3349	1 (16,7 %)	17 (89,5 %)	5 (83,3 %)	2 (10,5 %)	11,51	0,005	0,0172
9	Городское население, млн чел.	32	1 (16,7 %)	17 (89,5 %)	5 (83,3 %)	2 (10,5 %)	11,51	0,006	0,04187
10	Высокотехнологичный экспорт, %	22,35	7 (63,6 %)	33 (100 %)	4 (36,4 %)	0 (0 %)	12,9	0,007	0,303
11	Территория (площадь суши) в 2020 г., млн км ²	0,7451	6 (54,5 %)	34 (97,1 %)	5 (45,5 %)	1 (2,9 %)	13,099	0,007	0,000281
12	Население столицы в 2014 г., млн чел.	5,847	3 (50 %)	19 (100 %)	3 (50 %)	0 (0 %)	10,364	0,008	0,001142
13	Экспорт товаров и услуг, % ВВП	30,7999	6 (54,5 %)	3 (8,6 %)	5 (45,5 %)	32 (91,4 %)	10,996	0,011	0,001954
14	Рост населения, % в год	1,0	4 (36,4 %)	31 (88,6 %)	7 (63,6 %)	4 (11,4 %)	12,265	0,012	0,4694
15	Последствия стихийных бедствий (землетрясения, засухи, наводнения, штормы), всего погибших в 2003–2012 гг.	81	4 (40 %)	28 (90,3 %)	6 (60 %)	3 (9,7 %)	10,903	0,015	0,000947
16	Импорт товаров и услуг, % ВВП	32,2999	6 (54,5 %)	3 (8,6 %)	5 (45,5 %)	32 (91,4 %)	10,996	0,016	0,5196
17	Орошаемые земли, тыс. га	4125	3 (50 %)	19 (100 %)	3 (50 %)	0 (0 %)	10,364	0,016	1,4E–05
18	Средняя температура января в 1961–1999 гг., °C	8,6915	3 (50 %)	19 (100 %)	3 (50 %)	0 (0 %)	10,364	0,021	0,03669
19	Пахотные земли в 2018 г., млн га	11,865	6 (54,5 %)	33 (94,3 %)	5 (45,5 %)	2 (5,7 %)	10,022	0,023	0,06803
20	Территория (площадь суши) в 2014 г., млн км ²	0,8969	2 (33,3 %)	18 (94,7 %)	4 (66,7 %)	1 (5,3 %)	10,316	0,025	0,001142
21	Количество погибших в результате стихийных бедствий в 2003–2012 гг., чел.	47	0 (0 %)	13 (72,2 %)	6 (100 %)	5 (27,8 %)	9,061	0,026	0,000463
22	Доля территории, засеянной зерновыми культурами, от пахотных земель	0,27275	5 (45,5 %)	2 (5,7 %)	6 (54,5 %)	33 (94,3 %)	10,022	0,027	0,08072
23	Ожидаемая продолжительность жизни, лет	63,6999	3 (27,3 %)	0 (0 %)	8 (72,7 %)	35 (100 %)	9,989	0,027	0,004673
24	Ирригация, тыс. га	4212	7 (70 %)	35 (100 %)	3 (30 %)	0 (0 %)	11	0,028	0,3214
25	Газ, млрд м ³	3765,5	8 (72,7 %)	35 (100 %)	3 (27,3 %)	0 (0 %)	9,989	0,03	0,001255
26	Население столицы в 2020 г., млн чел.	6,328	8 (72,7 %)	35 (100 %)	3 (27,3 %)	0 (0 %)	9,989	0,036	0,06327
27	Количество погибших в результате стихийных бедствий в среднем за 2003–2012 гг., чел. на 100 тыс. чел. населения	18	1 (16,7 %)	15 (83,3 %)	5 (83,3 %)	3 (16,7 %)	8,625	0,038	0,03952

Окончание табл. 1
Ending table 1

№ п/п	Показатели	Граница	Квадрант I		Квадрант II		F	ПТ	U-тест
			1-й кластер	2-й кластер	1-й кластер	2-й кластер			
28	Среднегодовая смертность в результате стихийных бедствий в 2003–2012 гг., чел. на 1 млн чел. населения	0,1798	1 (16,7 %)	15 (83,3 %)	5 (83,3 %)	3 (16,7 %)	8,625	0,041	0,09846
29	Земли под зерновыми культурами в 2018 г., млн га	16,1749	8 (72,7 %)	35 (100 %)	3 (27,3 %)	0 (0 %)	9,989	0,044	0,1296

Приведем и другие значимые показатели по мере убывания информативности:

- на уровне от $p < 0,005$ до $p < 0,010$ – средняя температура января в 1961–1999 гг., °С; средняя температура февраля в 1961–1999 гг., °С; средняя максимальная температура в 1961–1999 гг., °С; средняя температура декабря в 1961–1999 гг., °С; запасы нефти, млн баррелей; средняя температура марта в 1961–1999 гг., °С; средняя температура ноября в 1961–1999 гг., °С; автомобильные дороги, включая грунтовые и покрытые, км на 1000 км² территории;

- на уровне от $p < 0,01$ до $p < 0,05$ – автомобильные дороги, включая грунтовые и покрытые, км на 1000 км² территории; импорт товаров и услуг, % ВВП; год обретения независимости; среднегодовое количество пострадавших от природных катаклизмов за период 1970–2011 гг., чел. на 100 тыс. чел. населения; амплитуда температур в течение года (max – min); среднемесячный объем осадков за март, мм; средняя минимальная температура в 1961–1999 гг., °С; сельскохозяйственные угодья, % от всей территории; средняя температура сентября в 1961–1999 гг., °С; экспорт товаров и услуг, % ВВП; коэффициент рождаемости; ущерб от природных катаклизмов в среднем за 2003–2012 гг., млн долл. на 1000 км² территории; население в крупнейшем городе, % от всего городского населения; воздействие экстремальных температур в 1971–1973, 1975, 1977–2008 гг., среднее значение; среднегодовая смертность в результате стихийных бедствий в 2003–2012 гг., чел. на 100 тыс. чел. населения; среднемесячный объем осадков за сентябрь, мм; индекс чистых бартерных условий торговли (2000 = 100); средняя температура апреля в 1961–1999 гг., °С.

Из табл. 1 видна некоторая корреляция между p -значениями, рассчитанными с помощью перестановочного теста в рамках метода ОДР. Однако для многих показателей имеются существенные разнонаправленные различия.

Применение метода ОДР также позволило выявить различия между кластерами и наиболее информативные показатели по двумерным разбиениям. Значимость аналогично считали с помощью перестановочных тестов.

На рис. 6 приведено двумерное разбиение, разделяющее 1-й и 2-й кластеры по показателям «безработица» и «доля населения в городских агломерациях с численностью населения более 1 млн человек». Значимость p_X составила 0,005, значимость p_Y была на уровне 0,017.

Из рис. 6 видно, что страны 1-го кластера преобладают в квадранте I, а странам из 2-го кластера соответствуют квадранты II, III и IV. Следовательно, для стран 1-го кластера характерны невысокая безработица и значительно больший процент населения в городах-миллионниках.

Все кластеры попарно сравнивали методами машинного обучения (МО). Методы ранжировали по эффективности и формировали из них ансамбли по 3, 5 или 7 методов. После этого для каждой сравниваемой пары кластеров выбирали тот ансамбль, который давал более высокие значения ROC AUC. Результаты наиболее успешных ансамблей методов МО представлены в табл. 2.

Таблица 2

Результаты распознавания в парах кластеров (ROC AUC) по ансамблю методов машинного обучения в комплексе *Data Master Azforus*

Table 2

Recognition results in pairs of clusters (ROC AUC) for an ensemble of machine learning methods in the *Data Master Azforus* complex

Сравниваемые кластеры (число стран)	Количество показателей	Ансамбль методов	ROC AUC
1-й кластер (11) – 2-й кластер (35)	130	AdaBoost, DT, GBM	0,9506
1-й кластер (11) – 3-й кластер (16)	130	AdaBoost, DT, KNN, LDA, SVM, SWS, GBM	0,8977

Окончание табл. 2
Ending table 2

Сравниваемые кластеры (число стран)	Количество показателей	Ансамбль методов	ROC AUC
1-й кластер (11) – 4-й кластер (11)	136	AdaBoost, KNN, SWS	0,9132
2-й кластер (35) – 3-й кластер (16)	130	AdaBoost, DT, LDA, SWS, GBM	0,9625
2-й кластер (33) – 4-й кластер (12)	128	AdaBoost, DT, SVM, SWS, GBM	0,9545
3-й кластер (16) – 4-й кластер (11)	132	AdaBoost, DT, SVM	0,7159

Примечание. AdaBoost – адаптивный бустинг; LDA – линейный дискриминант Фишера; KNN – метод k -ближайших соседей; SVM – метод опорных векторов; DT – решающие деревья; SWS – статистически взвешенные синдромы; GBM – градиентный бустинг.

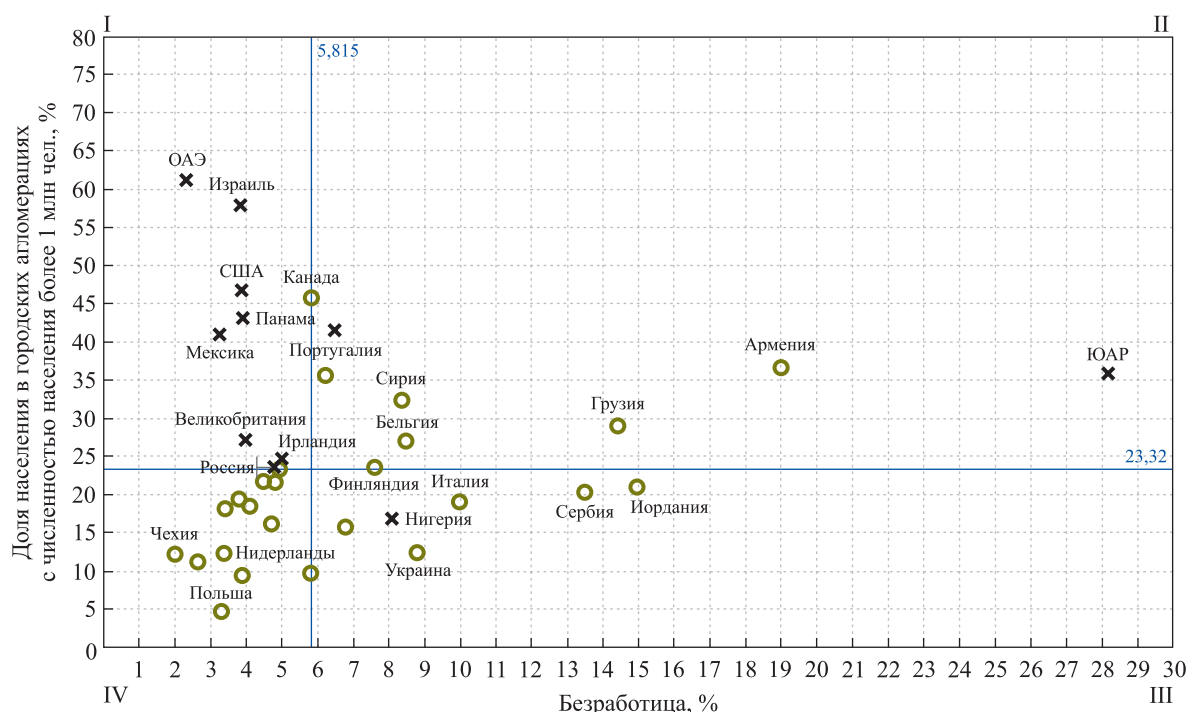


Рис. 6. Двумерное оптимальное разделение: ось x – безработица, ось y – доля населения в городских агломерациях страны с численностью населения более 1 млн чел.
Крестиками обозначены страны 1-го кластера, кружочками – страны 2-го кластера

Fig. 6. Two-dimensional optimal partitions: x axis – unemployment, y axis – population in urban agglomerations with a population of more than 1 million.
The crosses indicate the countries of the 1st cluster, circles – the countries of the 2nd cluster

Распознавание 1-го и 2-го кластеров методами МО (*Data Master Azforus*) показало следующий результат: ROC AUC = 0,95. Ансамбль методов включал адаптивный бустинг (ROC AUC = 0,895), деревья решений (ROC AUC = 0,775) и градиентный бустинг (ROC AUC = 0,732) и имел невысокую чувствительность (0,545) вследствие малого числа объектов в 1-м кластере, но при этом обладал высокой специфичностью (0,921).

Наилучший результат распознавания получили при сравнении 2-го и 3-го кластеров: ROC AUC = 0,9625. Это объясняется большим числом объектов и территориальным расположением вошедших в кластеры стран (европейские страны во 2-м кластере и страны с тропическим климатом в 3-м кластере). Самый слабый результат распознавания получили при сравнении 3-го и 4-го кластеров: ROC AUC = 0,7159. Это обусловлено малым числом объектов в 4-м кластере и тем, что сравниваемые кластеры включают страны со сходным климатом и экономикой.

Несмотря на то что метод CBC не вошел в некоторые ансамбли, он имеет преимущество перед остальными методами МО: это единственный метод, который выявляет наиболее значимые показатели с точки зрения разделения сравниваемых классов, ранжирует их, наглядно показывает на диаграммах рассеяния расположение объектов относительно выставленных границ разбиения. Все это позволяет опреде-

лить набор наиболее информативных (ключевых) показателей и оценить расположение объекта в кругу его соседей. Другие методы не обладают таким свойством (прозрачностью), что приводит к проблеме «черного ящика». Использование методов ОДР и СВС выводит анализ данных на новый уровень.

Заключение

Проведенные исследования подтвердили эффективность предложенной технологии изучения связи хода эпидемии COVID-19 с социально-экономическими, демографическими и климатическими факторами. На это указывает частичная связь найденных кластеров с географической локализацией. Вторым доказательством объективности кластеризации является высокая статистическая значимость различий между кластерами по важнейшим показателям, включая коэффициент Джини (расслоение по доходам), долю населения в городских агломерациях с численностью населения более 1 млн человек, амплитуду температур в течение года, ВВП на душу населения, импорт товаров и услуг, индекс чистых бартерных условий торговли, ожидаемую продолжительность жизни, рост населения, детскую смертность и др. Ранее показано [12], что ряд перечисленных показателей имеют статистически значимую связь с особенностями структуры базовых институтов стран мира. Поэтому в дальнейших исследованиях предполагается более детальное выявление зависимости особенностей протекания пандемии от особенностей социально-экономических институтов разных стран.

Следует отметить важность разрабатываемой технологии для решения задачи прогнозирования хода эпидемии. Такой прогноз может производиться по группе стран, в которую входит анализируемая страна.

Для успешного прогнозирования могут быть использованы модель ARIMA и различные модели авторегрессии. Авторегрессионные модели дают разное предсказание и разную точность предсказания в зависимости от того, проводилась ли массовая вакцинация, которая кардинально влияет на прогноз динамики заболеваемости COVID-19 [13]. Недостатком авторегрессионных моделей является неопределенность прогноза для моментов времени, удаленных от временного интервала, по которому была построена модель. В связи с этим в дальнейших исследованиях предполагается изучить прогностические возможности предложенной в работе [14] технологии, позволяющей получить прогнозирующую функцию в виде комбинации зависящих от времени комплексных экспоненциальных функций. Иными словами, прогноз будет экстраполирован на произвольный промежуток времени с помощью нелинейной функции, вычисляемой по интервалу наблюдений. Дальнейшим направлением исследований предполагается также включение в разрабатываемую технологию методов оценивания статистической значимости кластеризации.

Библиографические ссылки

1. Романюха АА, Санникова ТЕ, Дрынов ИД. Возникновение эпидемий острых респираторных заболеваний. *Вестник Российской академии наук*. 2011;81(2):122–126.
2. Борисова ЛР, Фридман МН. Некоторые аспекты влияния пандемии коронавируса на экономику. *Самоуправление*. 2021; 5:147–152.
3. Sengupta P, Ganguli B, SenRoy S, Chatterjee A. An analysis of COVID-19 clusters in India. *BMC Public Health*. 2021;21:631. DOI: 10.1186/s12889-021-10491-8.
4. Zarikas V, Pouloupoulos SG, Gareiou Z, Zervas E. Clustering analysis of countries using the COVID-19 cases dataset. *Data in Brief*. 2020;31:105787. DOI: 10.1016/j.dib.2020.105787.
5. Mengyang Liu, Mengmeng Liu, Zhiwei Li, Yingxuan Zhu, Yue Liu, Xiaonan Wang, et al. The spatial clustering analysis of COVID-19 and its associated factors in mainland China at the prefecture level. *Science of the Total Environment*. 2021;777:145992. DOI: 10.1016/j.scitotenv.2021.145992.
6. Rios RA, Nogueira T, Coimbra DB, Lopes TJS, Abraham A, de Mello RF. Country transition index based on hierarchical clustering to predict next COVID-19 waves. *Scientific Reports*. 2021;11(1):15271. DOI: 10.1038/s41598-021-94661-z.
7. Rizvi SA, Umair M, Cheema MA. Clustering of countries for COVID-19 cases based on disease prevalence, health systems and environmental indicators. *Chaos, Solitons & Fractals*. 2021;151:111240. DOI: 10.1016/j.chaos.2021.111240.
8. Brzyska J, Szamrej-Baran I. Classification of the EU countries according to the vulnerability of their economies to the impact of COVID-19 pandemic. *European Research Studies Journal*. 2021;XXIV(2B):967–978. DOI: 10.35808/ersj/2318.
9. Kuznetsova AV, Kostomarova IV, Senko OV. Modification of the method of optimal valid partitioning for comparison of patterns related to the occurrence of ischemic stroke in two groups of patients. *Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications*. 2014;24(1):114–123. DOI: 10.1134/S105466181401009X.
10. Senko OV, Dzyba DS, Pigarova EA, Rozhinskaya LYa, Kuznetsova AV. A method for evaluating validity of piecewise-linear models. In: Fred ALN, Filipe J, editors. *Proceedings of the 6th International Conference on Knowledge Discovery and Information Retrieval; 2014 October 21–24; Rome, Italy*. [S. l.]: Science and Technology Publications; 2014. p. 437–442. DOI: 10.5220/0005156904370443.
11. Senko OV, Kuznetsova AV. A recognition method based on collective decision making using systems of regularities of various types. *Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications*. 2010;20(2):152–162. DOI: 10.1134/S1054661810020069.
12. Кирилук ИЛ, Волинский АИ, Круглова МС, Кузнецова АВ, Рубинштейн АА, Сенько ОВ. Эмпирическая проверка теории институциональных матриц методами интеллектуального анализа данных. *Компьютерные исследования и моделирование*. 2015;7(4):923–939. DOI: 10.20537/2076-7633-2015-7-4-923-939.

13. Борисова ЛР. Исследование динамики заболеваемости коронавирусной инфекцией в Москве. В: *Современные проблемы физико-математических наук. Материалы VII Всероссийской научно-практической конференции с международным участием; 18–21 ноября 2021 г.; Орел, Россия*. Орел: Орловский государственный университет имени И. С. Тургенева; 2021. с. 217–220.
14. Smirnov VYu, Kuznetsova AV. Approximation of experimental data by solving linear difference equations with constant coefficients (in particular, by exponentials and exponential cosines). *Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications*. 2017;27(2):175–183. DOI: 10.1134/S1054661817020109.

References

1. Romanyukha AA, Sannikova TE, Drynov ID. [Emergence of epidemics of acute respiratory diseases]. *Vestnik Rossiiskoi akademii nauk*. 2011;81(2):122–126. Russian.
2. Borisova LR, Fridman MN. Some aspects of the impact of the coronavirus pandemic on the economy. *Samoupravlenie*. 2021; 5:147–152. Russian.
3. Sengupta P, Ganguli B, SenRoy S, Chatterjee A. An analysis of COVID-19 clusters in India. *BMC Public Health*. 2021;21:631. DOI: 10.1186/s12889-021-10491-8.
4. Zarikas V, Pouloupoulos SG, Gareiou Z, Zervas E. Clustering analysis of countries using the COVID-19 cases dataset. *Data in Brief*. 2020;31:105787. DOI: 10.1016/j.dib.2020.105787.
5. Mengyang Liu, Mengmeng Liu, Zhiwei Li, Yingxuan Zhu, Yue Liu, Xiaonan Wang, et al. The spatial clustering analysis of COVID-19 and its associated factors in mainland China at the prefecture level. *Science of the Total Environment*. 2021;777:145992. DOI: 10.1016/j.scitotenv.2021.145992.
6. Rios RA, Nogueira T, Coimbra DB, Lopes TJS, Abraham A, de Mello RF. Country transition index based on hierarchical clustering to predict next COVID-19 waves. *Scientific Reports*. 2021;11(1):15271. DOI: 10.1038/s41598-021-94661-z.
7. Rizvi SA, Umair M, Cheema MA. Clustering of countries for COVID-19 cases based on disease prevalence, health systems and environmental indicators. *Chaos, Solitons & Fractals*. 2021;151:111240. DOI: 10.1016/j.chaos.2021.111240.
8. Brzyska J, Szamrej-Baran I. Classification of the EU countries according to the vulnerability of their economies to the impact of COVID-19 pandemic. *European Research Studies Journal*. 2021;XXIV(2B):967–978. DOI: 10.35808/ersj/2318.
9. Kuznetsova AV, Kostomarov IV, Senko OV. Modification of the method of optimal valid partitioning for comparison of patterns related to the occurrence of ischemic stroke in two groups of patients. *Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications*. 2014;24(1):114–123. DOI: 10.1134/S105466181401009X.
10. Senko OV, Dzyba DS, Pigarova EA, Rozhinskaya LYa, Kuznetsova AV. A method for evaluating validity of piecewise-linear models. In: Fred ALN, Filipe J, editors. *Proceedings of the 6th International Conference on Knowledge Discovery and Information Retrieval; 2014 October 21–24; Rome, Italy*. [S. l.]: Science and Technology Publications; 2014. p. 437–442. DOI: 10.5220/0005156904370443.
11. Senko OV, Kuznetsova AV. A recognition method based on collective decision making using systems of regularities of various types. *Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications*. 2010;20(2):152–162. DOI: 10.1134/S1054661810020069.
12. Kirilyuk IL, Volynsky AI, Kruglova MS, Kuznetsova AV, Rubinstein AA, Senko OV. Empirical testing of institutional matrices theory by data mining. *Computer Research and Modeling*. 2015;7(4):923–939. Russian. DOI: 10.20537/2076-7633-2015-7-4-923-939.
13. Borisova LR. [Study of the dynamics of the incidence of coronavirus infection in Moscow]. In: *Sovremennye problemy fiziko-matematicheskikh nauk. Materialy VII Vserossiiskoi nauchno-prakticheskoi konferentsii s mezhdunarodnym uchastiem; 18–21 noyabrya 2021 g.; Orel, Rossiya* [Modern problems of physical and mathematical sciences. Materials of the 7th All-Russian scientific and practical conference with international participation; 2021 November 18–21; Orel, Russia]. Orel: Orel State University named after I. S. Turgenev; 2021. p. 217–220. Russian.
14. Smirnov VYu, Kuznetsova AV. Approximation of experimental data by solving linear difference equations with constant coefficients (in particular, by exponentials and exponential cosines). *Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications*. 2017;27(2):175–183. DOI: 10.1134/S1054661817020109.

Получена 31.12.2021 / исправлена 04.03.2022 / принята 04.03.2022.
Received 31.12.2021 / revised 04.03.2022 / accepted 04.03.2022.

Авторы:

Олег Валентинович Сенько – доктор физико-математических наук, профессор; старший научный сотрудник научной группы математических методов и эпидемиологического прогнозирования¹⁾, ведущий научный сотрудник Вычислительного центра имени А. А. Дородницына Российской академии наук²⁾.

Анна Викторовна Кузнецова – кандидат биологических наук; старший научный сотрудник научной группы математических методов и эпидемиологического прогнозирования¹⁾, старший научный сотрудник лаборатории математической биофизики³⁾.

Authors:

Oleg V. Senko, doctor of science (physics and mathematics), full professor; senior researcher at the scientific group of mathematical methods and epidemiological forecasting^a and leading researcher at the Dorodnicyn Computing Centre, Russian Academy of Sciences^b.

senkoov@mail.ru

<https://orcid.org/0000-0002-5586-3503>

Anna V. Kuznetsova, PhD (biology); senior researcher at the scientific group of mathematical methods and epidemiological forecasting^a and senior researcher at the laboratory of mathematical biophysics^c.

azforus@yandex.ru

<https://orcid.org/0000-0002-0297-7013>

Евгений Михайлович Воронин – кандидат медицинских наук; руководитель научной группы математических методов и эпидемиологического прогнозирования.

Ольга Анатольевна Кравцова – статистик научной группы математических методов и эпидемиологического прогнозирования¹⁾, аспирантка кафедры математических методов прогнозирования факультета вычислительной математики и кибернетики⁴⁾. Научный руководитель – О. В. Сенько.

Людмила Робертовна Борисова – кандидат физико-математических наук; доцент Департамента математики.

Игорь Леонидович Кирилук – научный сотрудник сектора институционально-эволюционной экономики Центра институционально-эволюционной экономики и прикладных проблем воспроизводства.

Василий Геннадьевич Акимкин – академик Российской академии наук, доктор математических наук, профессор; директор.

Evgeny M. Voronin, PhD (medicine); head of the scientific group of mathematical methods and epidemiological forecasting.

emvoronin@yandex.ru

<https://orcid.org/0000-0001-5925-7757>

Olga A. Kravtsova, methodological expert at the scientific group of mathematical methods and epidemiological forecasting^a and postgraduate student at the department of mathematical methods of forecasting, faculty of computational mathematics and cybernetics^d.

dbri.olga@gmail.com

<https://orcid.org/0000-0002-7757-5334>

Ludmila R. Borisova, PhD (physics and mathematics); associate professor at the Department of Mathematics.

borisovalr@mail.ru

<https://orcid.org/0000-0002-5757-0341>

Igor L. Kirilyuk, researcher at the subcenter «Institutional and evolutionary economics», Center for Institutional and Evolutionary Economics and Applied Problems of Reproduction.

igokir@rambler.ru

<https://orcid.org/0000-0002-8935-9241>

Vasiliy G. Akimkin, academician of the Russian Academy of Sciences, doctor of science (medicine), full professor; director.

crie@pcr.ru

<https://orcid.org/0000-0003-4228-9044>
