

**ПОСТАНАЛИЗ ДАННЫХ ПО ЗАБОЛЕВАЕМОСТИ COVID-19
МЕТОДОМ SSA НА ПРИМЕРЕ НОВОСИБИРСКОЙ ОБЛАСТИ**

Петракова В. С. ^{1, 2}

¹ Институт вычислительного моделирования СО РАН.

² Институт математики им. Соболева СО РАН.

**SSA POST-ANALYSIS OF COVID-19 NOVOSIBIRSK REGION
INCIDENCE MAGNITUDE**

Petrakova V. S. ^{a, b}

^a Institute of Computational Modeling SB RAS.

^b Sobolev Institute of Mathematics SB RAS.

Резюме

Статья посвящена преимущественно пост-анализу эпидемиологических данных о Covid-19 в Новосибирске за период с 2020 по 2023 годы с использованием метода сингулярного спектрального анализа (Singular spectrum analysis, SSA). Предложен литературный обзор работ, посвященный анализу данных различного типа, описывающих эпидемиологическую ситуацию во время пандемии в различных регионах мира. Показано, что преобладающая часть работ, написана и посвящена описанию первого начального этапа развития эпидемии (2020 г). Работ по пост-анализу данных о многолетней динамике вируса Sars-Cov-2 фактически не наблюдается. Исследование подчеркивает важность пост-анализа данных для понимания динамики распространения вируса и особенностей его воздействия на здоровье населения. Результаты проведенного исследования позволяют оценить, как менялась восприимчивость населения к разным штаммам вируса, в чем отличие вспышки заболеваемости, связанной с появлением нового штамма вируса, от сезонного распространения инфекции.

Для анализа в статье используется метод SSA, применяемый к анализу временных рядов для разделения их на составляющие. Метод применялся для изучения ключевых показателей, таких как количество новых заражений, смертей, числа критических случаев, числа госпитализированных и числа пациентов, находящихся на ИВЛ в Новосибирской области. Для описанных наборов данных выделяются три основные компоненты: общий тренд, который отражает изменение скорости распространения вируса с распространением новых штаммов, а также периодические явления, связанные с штаммами и сезонностью.

Результаты показывают, что значительная часть изменений в динамике заболеваний обусловлена появлением новых штаммов, но также проявляется и "фоновая" эпидемия с сезонными колебаниями. Это подчеркивает

необходимость учитывать множество факторов, влияющих на распространение вируса, включая иммунитет, методы лечения и качество медицинской помощи. Показаны наблюдаемые взаимосвязи и временные лаги между количеством критических больных и количеством зафиксированных смертей от вируса, а также между количеством госпитализированных больных и пациентов, находящихся на ИВЛ.

Статья завершается выводами о том, что выделенный тренд показывающий, как менялось число инфицированных с развития штаммов, может быть полезен для уточнения параметров математических моделей распространения COVID-19. В качестве иллюстративного примера выбрана дифференциальная модель SEIR-HCD, которая ранее использовалась для моделирования распространения заболеваемости в Новосибирской области. Показано, что параметр скорости распространения вируса, восстановленный через по выделенному тренду, при подстановке его в модель даёт меньшую ошибку моделирования, чем восстановленный с помощью решения обратной задачи.

Ключевые слова: данные по заболеваемости Covid-19, пост-анализ данных, SSA метод, SEIR-HCD модель, эпидемиологические модели, анализ временных рядов.

Abstract

The article is aimed at conducting post-analysis of Covid-19 epidemiological data in Novosibirsk for the period from 2020 to 2023. The study emphasizes the importance of data post-analysis for understanding the dynamics of SARS-CoV-2 spread and the characteristics of its impact on public health. The study results make possible to assess of how population susceptibility to different virus strains has changed, what is the difference between an outbreak of diseases associated with the emergence of a new virus strain and its seasonal infection spread.

There was applied the SSA method for analyzing time series to separate them into components as well as studying key indicators such as the number of new infections, deaths, critical cases, hospitalizations, and ventilator-dependent patients in the Novosibirsk region. Three main components have been identified for the described data sets: a general trend that reflects changes in the rate of virus spread related to spread of new strains, as well as periodic phenomena associated with virus strains and seasonality.

The results show that a significant part of the changes in disease dynamics is accounted for by the emergence of new strains, but also due to "chronicity" epidemic with seasonal fluctuations. The observed relationships and time lags between the number of critically-ill patients and the number of recorded deaths due to COVID-19, as well as between the number of hospitalized patients and ventilator-dependent patients are shown.

Thus, it is concluded that the identified trend depicting a change between the number of infected people and development of virus strains can be useful for refining the parameters of mathematical models for COVID-19 spread. The SEIR-HCD differential model, which was previously used to simulate the disease spread in the Novosibirsk region, was chosen as an illustrative example. It is shown that the parameter of the virus spread rate, restored through the selected trend, when

introduced into the model, provides a smaller modeling error than the if it was generated using the solution of the inverse issue.

Keywords: Covid-19 incidence data, post-analysis of data, SSA method, SEIR-HCD model, epidemiological models, time series analysis.

1. Введение

Пандемия Covid-19 стала одной из самых значительных глобальных угроз здравоохранению в XXI веке, оказавшее глубокое влияние на все аспекты жизни общества, экономики и системы здравоохранения. Поиск эффективных мер по контролю и предотвращению распространения новых вирусов требует адекватной оценки прошедшей пандемии и постанализа эпидемиологических данных по распространению вируса, его клинических характеристик и последствий для здоровья населения.

Вместе с тем, работ по анализу реальных данных по пандемии Covid-19 значительно меньше, чем по разработке моделей, прогнозирующих развитие эпидемии. В статье [28] авторы провели всесторонний обзор открытых наборов данных по Covid-19. Обзор обобщал типы данных и область их применения. Авторы показали, что основным видом открытых данных были медицинские изображения, текстовые и речевые данные (записи кашля). В работе [24] проведен анализ наборов данных по распространению Covid-19 в Индии с помощью методов машинного обучения для понимания того, какая возрастная группа больше всего пострадала из-за вируса. В работе [18] авторы исследовали клинические данные и продемонстрировали использование моделей машинного обучения для прогнозирования наличия Covid-19 у пациента. В частности, через анализ клинических данных, авторы успешно дифференцировали пациентов с Covid-19 от пациентов с гриппом типа А. Авторы статьи [9] на основе официальных данных по заболеваемости делают обзор и сравнительный анализ развития эпидемиологической ситуации и политики ответных мер во Франции, Бельгии и Канаде на ранних стадиях пандемии. В статье [12] авторами приводится анализ временной динамики вспышки коронавирусной болезни за первые 3 месяца 2020 г в Китае, Италии и Франции. Анализ показывает, что кинетический параметр, описывающий скорость выздоровления, по-

видимому, одинаков независимо от страны, в то время как показатели инфицирования и смертности более зависимы от региона. Отметим без детального описания еще несколько работ [7,8, 31], в которых также ключевая роль отведена анализу эпидемиологических данных. Как показывает литературный обзор, преобладающая часть работ, посвященная анализу данных, описывающих эпидемиологическую ситуацию, написана и посвящена описанию первого начального этапа развития эпидемии (2020 г). Работ по пост-анализу данных фактически не наблюдается. В большинстве это работы, связанные с анализом текстовых сообщений в социальной сети Twitter для оценки социальных настроений общества во время процесса вакцинации [25]; экономических последствий пандемии [30]; наличия коллективного иммунитета в Австрии [27]. Пост-анализа данных о многолетней динамике вируса Sars-Cov-2 автором найдено не было. Вместе с тем результат такого анализа может дать ответы на некоторые актуальные вопросы, например,

- а) как менялась восприимчивость населения к разным штаммам вируса?
- б) как отличить новый штамм вируса от сезонного распространения инфекции?
- в) каковы ключевые тренды, паттерны и корреляции, которые могут быть использованы для прогнозирования будущих вспышек и оценки эффективности внедренных мер общественного здравоохранения?

Данная работа посвящена ответу на предложенные выше вопросы на примере SSA анализа наблюдений за распространением коронавирусной инфекции в Новосибирской области (Россия) в период с 2020 по 2022 годы. Работа организована следующим образом. В разделе 2.1 приведено описание выбранного для анализа датасета. В разделе 2.2 приведено описание метода и обзор литературы по применению метода к данным,

связанным с пандемией Covid-19. В разделе 3 приведены результаты исследования. В разделе 4 приведены полученные выводы.

2. Материалы и методы

2.1. Описание используемого для анализа набора данных

Для анализа были выбраны наблюдения за динамикой Covid-19 в Новосибирской области (Россия). Данные доступны онлайн и могут быть найдены по ссылке <https://ai-biolab.ru/data/En>. Помимо данных по Новосибирской области, по ссылке также доступны данные по заболеваемости в Москве, Санкт-Петербургу, Омске и респ. Алтай (Россия). Данные по Новосибирской области были выбраны для анализа, поскольку они являются наиболее полными. Описание датасета представлено в таблице 1.

Для анализа были выбраны следующие поля: *new_diagnoses*, *new_deaths*, *n_critical*, *hospitalised*, *ventilation*, поскольку эти поля являются ключевыми для описания динамики распространения SARS-CoV-2 в рассматриваемом регионе.

2.2. Метод анализа сингулярного спектра (Singular Spectrum Analysis)

Метод Singular Spectrum Analysis (SSA), также известный как “метод гусеницы”, представляет собой мощный инструмент для анализа, декомпозиции и прогнозирования временных рядов числовых измерений [10,13]. Этот метод был разработан в 1980-х годах и с тех пор получил широкое признание в различных областях, таких как экономика, метеорология, медицина и многие другие. Временным рядом здесь называется последовательность наблюдений за динамической системой.

Принцип работы метода SSA основывается на разложении временного ряда на набор компонент: различные виды трендов, сезонностей, циклов и случайных колебаний. При этом каждая компонента ассоциирована с вкладом сингулярного числа траекторной матрицы ряда и его векторов.

86 Так, будем рассматривать временной ряд $X_t = (x_1, \dots, x_N)$ длины N , такой
87 что $x_i \in \square, i \in \{1, \dots, N\}$. Метод сингулярного анализа заключается в
88 реализации нескольких шагов.

89 Шаг 1. Построение траекторной матрицы.

90 Пусть L ($1 < L < N$) — некоторое целое число (выбираемое
91 пользователем), называемое длиной окна, и $K = N - L + 1$. Траекторной
92 матрицей ряда X_t будем называть

$$93 \quad \mathbf{X} = \begin{bmatrix} x_1 & x_2 & x_3 & \dots & x_K \\ x_2 & x_3 & x_4 & \dots & x_{K+1} \\ x_3 & x_4 & x_5 & \dots & x_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & x_{L+2} & \dots & x_N \end{bmatrix},$$

94 где матрица $\mathbf{X}$ является ганкелевой размерностью $L \times K$.

95 Шаг 2. Сингулярное разложение траекторной матрицы.

96 Далее для траекторной матрицы получают сингулярное разложение,
97 позволяющее оценить собственные значения и собственные векторы
98 матрицы $\mathbf{X}$. Пусть $d = \text{rank}(\mathbf{X})$. Тогда $\mathbf{X}$ представимо в виде суммы
99 элементарных матриц:

$$100 \quad \mathbf{X} = \sum_{i=1}^d \sqrt{\lambda_i} U_i V_i^T = \sum_{i=1}^d \mathbf{X}_i.$$

101 Здесь $U_i, i = 0, \dots, d-1$ и $V_i, i = 0, \dots, d-1$ -- левые и правые сингулярные
102 вектора, которые по свойству SVD составляют ортонормированные
103 базисы пространства строк и пространства столбцов матрицы $\mathbf{X}$,
104 соответствующие упорядоченным по убыванию собственным числам
105 $\lambda_i, i = 0, \dots, d-1$ матрицы $\mathbf{S} = \mathbf{X}\mathbf{X}^T$. Матрицы $\mathbf{X}_i$ называют элементарными.

106 Шаг 3: Группировка элементарных матриц.

Далее элементарные матрицы с похожими свойствами могут быть сгруппированы. Множество всех индексов $\{0, \dots, d-1\}$ разбивается на m непересекающихся подмножеств $I_1, \dots, I_m$. Пусть $I = \{i_1, \dots, i_p\}$, тогда результирующая матрица X_I , соответствующая группе I , определяется как $X_I = X_{i_1} + \dots + X_{i_p}$. Тогда траекторная матрица X представляется в виде суммы сгруппированных матриц $X = X_{I_1} + \dots + X_{I_m}$.

Шаг 4: Диагональное усреднение.

Далее каждую сгруппированную матрицу X_{I_i} необходимо развернуть обратно во временной ряд. Для этого каждая матрица X_{I_j} сгруппированного разложения ганкелизуется и затем полученная ганкелева матрица трансформируется в новый временной ряд длины N . Диагональное усреднение, применённое к каждой результирующей матрице X_{I_i} , производит восстановленные ряды $\tilde{X}_t^{(k)} = (\tilde{x}_1^{(k)}, \dots, \tilde{x}_N^{(k)})$, $k = 1, \dots, m$. Таким образом, исходный ряд X_t раскладывается в сумму m восстановленных рядов:

$$X_t = \sum_{k=1}^m \tilde{X}_t^{(k)}. \quad (1)$$

Разложение (1) является главным результатом алгоритма SSA, поскольку каждая результирующая компонента $\tilde{X}_t^{(k)}$ соотносится с набором собственных значений траекторной матрицы исходного ряда и может быть интерпретируема как либо тренд, либо колебания (периодики), либо шум. Метод SSA обладает рядом преимуществ для анализа временных рядов. Во-первых, метод не требователен к структуре ряда, не требователен к предположению распределения данных, применим к нестационарным временным рядам. Во-вторых, компоненты, полученные в результате

анализа, легко интерпретируются в рамках исследуемых данных и визуализируются. В-третьих, метод достаточно устойчив к шуму в данных. К недостаткам метода относят сложность выбора гиперпараметров метода: выбор размера окна L для формирования траекторной матрицы и количества ключевых компонент ряда может быть затруднительным, зависит, в большей степени, от субъективного мнения исследователя и требует адаптации под конкретные данные. Другой сложностью является выбор набора элементарных матриц, которые могут быть сгруппированы на шаге 3. Этот выбор также определяется исследователем.

Различные модификации метода используются для заполнения пропусков в данных [32], поиска разладок [2] и прогнозирования [29]. Можно найти большое число работ по прогнозированию развития пандемии Covid-19 с использованием метода SSA в период начала развития эпидемии (см., например, работы [3, 14]). Работ по пост-анализу данных известно не много. Например, в работе [20] проведены исследования причинно-следственных связей между изменениями погоды и распространением COVID-19 через показатели мобильности населения (несоблюдение изоляции) в том числе с помощью SSA анализа. Хотелось бы отметить еще работу [11], где проведен анализ многомерных, многомасштабных клинических данных и лабораторных результатов для классификации клинических фенотипов.

3. Результаты применения метода SSA к имеющимся данным

3.1. Данные по числу новых детектированных инфицированных

Сначала представим результаты постанализа для данных по числу новых выявленных случаев заражения коронавирусом (поле *new_diagnoses* в датасете). На рисунке 1 представлен вклад i -той элементарной матрицы этого ряда, ассоциированной с i -тым сингулярным числом траекторной

матрицы, отсортированным в порядке убывания. Из рисунка видно, что максимальное собственное число вносит четверть вклада в последовательность данных, а первые 10 элементарных матриц объясняют 80% траекторной матрицы ряда.

Элементарные матрицы были сгруппированы по принципу совпадения частот соответствующих им элементарных рядов. На рисунке 2 изображены выделенные методом SSA основные компоненты из временного ряда по числу новых заболевших коронавирусом в г. Новосибирске за период с 2020 по 2023 гг. Так, нулевая компонента была выделена в отдельный ряд, представляющий собой общий тренд развития заболеваемости с течением пандемии. Этот тренд, соответствующий изменению скорости распространения вируса от перехода от одного штамма вируса к другому, на рисунке представлен синей линией. Первая, вторая и третья компоненты были объединены и развернуты в один ряд, который соответствует периодам повышения количества случаев обнаружения заболевания, ассоциированным с появлением нового штамма (оранжевая линия). Отметим, что полученные периоды повышения по датам совпадают с официально заявляемыми штаммами (Whuhan, Delta, Omicron, Deltacron). Оставшиеся с 4-той по 10-тую компоненты также были объединены в один ряд. Так были выбраны периоды повышения количества случаев обнаружения заболевания, не связанные со штаммами. Такие периоды могут быть ассоциированы с сезонностью заболеваемости и её хроническим протеканием (зеленая линия). Отметим, что последний выделенный ряд имеет периодичность в пределах 3-х месяцев. Выделенные ряды объясняют 80% ряда. Остальные компоненты классифицированы как шум.

3. 2. Данные по числу новых умерших и людей в критическом состоянии

Схожий анализ был выполнен для данных по числу людей, умерших от коронавируса (*new_deaths*) или находящихся в критическом состоянии (*n_critical*). На рисунках 3 и 4 представлены вклады элементарных матриц, ассоциированных с сингулярными числами траекторных матриц для умерших и критических больных, соответственно. Из рисунков видно, что максимальные сингулярные числа для таких рядов вносят преобладающий вклад (92% для умерших, 82% для критических). Это может быть объяснено тем, что изменение в уровне смертности (и числа критических больных) в основном объясняется сменой штамма вируса и развитием качества медицинской помощи и профилактических мер на протяжении борьбы с эпидемией, а не с сезонными вспышками.

Таким образом, из данных было выделено по две компоненты - тренд, соответствующий нулевой элементарной матрице, и периодическая компонента, объединяющая с 1 по 6 элементарные матрицы. Выделенные компоненты согласно рисункам 3, 4 объясняют более 97% исходных временных рядов. На рисунке 6 предложено сравнение нормированных в пределах от [0,1] выделенных трендов и периодик для обоих рядов. Из рисунка 5 видно, что для обоих временных рядов тренды идентичны, что может указывать на их линейную зависимость. Периодичность числа критических случаев, в целом совпадает с периодами повышения заболеваемости, выделенными на рисунке 2, как хроническое протекание болезни. Пики повышения уровня умерших смещены относительно критических в пределах 1-3 недель для разных штаммов.

3.3. Данные по числу госпитализированных и пациентов, находящихся на ИВЛ

И, наконец, анализ был выполнен для людей, поступивших на госпитализацию (*hospitalised*) и находящихся на искусственном вентилировании легких (*ventilation*). Для экономии текста, мы не приводим здесь вклад сингулярных чисел (и ассоциированных с ними элементарных матриц) для текущих рядов. Вклад максимальных собственных чисел составляет 75% и 73% относительно общего разложения для рядов госпитализированных людей и пациентов, находящихся на искусственной вентиляции легких, соответственно. Вклад каждого следующего по модулю сингулярного числа составляет менее 5%. Первые 10 максимальных по модулю сингулярных чисел (и ассоциированных с ними элементарных матриц) объясняют более 95% траекторной матрицы исходного ряда.

Как и в предыдущем случае, трендовые компоненты (см. рисунок 6) обоих рядов имеют одинаковое поведение. Пики периодических компонент совпадают между собой для двух исследуемых рядов для вспышки, ассоциированной со штаммом Delta, для остальных имеют смещение в пределах 3-х недель. То есть, после пика повышения числа госпитализированных в течение 3-х недель будет наблюдаться пиковая потребность в аппаратах ИВЛ. Возникновение пиков в периодических компонентах совпадает со вспышками, определенными как “сезонные” (ассоциированные с хроническим протеканием болезни) на рисунке

2.

3.4. Оценка восприимчивости населения к разным штаммам вируса Sars-Cov-2 с помощью инструментов математического моделирования

Для моделирования динамики распространения коронавируса в период 2020-2022 гг. разрабатывалось большое число математических моделей. В данной работе не будет приводиться детальное описание разных разработанных подходов, существует большое число обзорных работ,

описывающих преимущества и недостатки разных моделей. Отметим лишь обзорную работу 2022 г. [1], освещающую развитие области в период быстрого распространения эпидемии, и работу 2024 г. [6], проделанную после завершения пандемии. Отметим, что в большинстве разработанных для прогнозирования распространения Covid-19 моделей лежит принцип SIR-динамики, предложенный еще Кермаком и МакКендриком в 1920-х годах [15], где популяция разбивается на несколько эпидемиологических групп (в оригинальных работах на три группы: S - неиммунные, I - инфицированные, R - иммунные) и передача вируса определяется через взаимодействие этих групп между собой и набором параметров, описывающих характер распространения вируса (скорость инфицирования, контактное число, скорость выздоровления и т.д.). На основе такого принципа моделирования было разработано большое число более сложных моделей, например, в форме систем обыкновенных дифференциальных уравнений [26], уравнений в частных производных [22], моделей оптимального управления [5] и агентно-ориентированных моделей [4]. Основой для качества такого типа моделирования является правильное определение параметров, определяющих распространение вируса. В работе [17,19] было показано, что наиболее чувствительным параметром, влияющим на результаты моделирования, является параметр скорости распространения вируса. Вместе с тем определение этого параметра является сложной задачей, поскольку сам параметр включает в себя заразность вируса, среднее число контактов заболевшего с неинфицированной частью населения, информацию о соблюдении изоляции и противовирусных мер и т.д. В математическом моделировании для определения этого параметра используются постановки обратных задач относительно реальных данных по числу инфицированных [17].

На основе анализируемых данных для описания распространения вируса Sars-Cov-2 в г. Новосибирске была предложена дифференциальная модель SEIR-HCD [16]. Наличие большого числа измеряемых показателей по распространению пандемии совместно с эпидемиологами позволило разработать сложную математическую модель, позволяющую учесть многие аспекты исследуемого вируса. Модель представляется в виде системы дифференциальных уравнений вида

$$\begin{cases} dS / dt = -(5 - a)(\alpha_I S(t)I(t) + \alpha_E S(t)E(t)) / 5 + \omega_{imm} R(t), \\ dE / dt = (5 - a)(\alpha_I S(t)I(t) + \alpha_E S(t)E(t)) / 5 - \omega_{inc} E(t), \\ dI / dt = \omega_{inc} E(t) - \omega_{inf} I(t), \\ dR / dt = \beta \cdot \omega_{inf} I(t) + (1 - \varepsilon_{HC}) \cdot \omega_{hosp} H(t) - \omega_{imm} R(t), \\ dH / dt = (1 - \beta) \cdot \omega_{inf} I(t) + (1 - \mu) \cdot \omega_{crit} C(t) - \omega_{hosp} H(t), \\ dC / dt = \varepsilon_{HC} \omega_{hosp} H(t) - \omega_{crit} C(t), \\ dD / dt = \mu \omega_{crit} C(t) \end{cases} \quad (2)$$

с известными начальными условиями

$$\begin{aligned} S(0) = S_0, E(0) = E_0, I(0) = I_0, R(0) = R_0, \\ H(0) = H_0, C(0) = C_0, D(0) = D_0. \end{aligned} \quad (3)$$

Здесь популяция разделена на 7 групп: восприимчивые к заболеванию (S), подвергшиеся воздействию или бессимптомные (E), инфицированные или имеющие симптомы (I), иммунные или выздоровевшие (R), госпитализированные (H), критические (C) и умершие (D). В каждый момент времени $S(t) + E(t) + I(t) + R(t) + H(t) + C(t) + D(t) = 1$, где 1 означает всю популяцию. Параметры $a, \alpha_I, \alpha_E, \omega_{inc}, \omega_{inf}, \beta, \varepsilon_{HC}, \omega_{hosp}, \omega_{imm}, \mu, \omega_{crit} \in \mathbf{R}$ называются эпидемиологическими параметрами и описывают распространение вируса. Здесь a — индекс самоизоляции; α_E, α_I — параметры заразности вируса в пределах эпидемиологических групп E и I ; β — доля инфицированных без осложнений; μ — смертность; ε_{HC} — доля критических случаев от числа госпитализированных;

292 $\omega_{inc}, \omega_{inf}, \omega_{hosp}, \omega_{imm}, \mu, \omega_{crit}$ — обратные величины среднему числу дней
293 (ω_{inc}) от контакта до заражения, (ω_{inf}) продолжительности
294 симптоматического периода, (ω_{hosp}) развития тяжелых случаев в
295 критические, (ω_{crit}) критического состояния, (ω_{imm}) поддержания
296 иммунитета. Область изменения параметров для пандемии Covid-19 и их
297 более подробное описание предложены в [23]. Данная модель была
298 успешно использована для прогнозирования развития Covid19 в г.
299 Новосибирске и Москве [16]. Модель хорошо исследована [17] и на её
300 основе предлагались модели оптимального управления [21] и модели
301 среднего поля [23].

302 Для Новосибирской области параметры модели (2), (3) были
303 восстановлены с помощью решения обратной задачи по данным. Метод
304 восстановления параметров описан в работе [16]. Данные по количеству
305 людей в каждой эпидемиологической группе S, E, I, R, H, C, D были
306 получены из исследуемых в первой части работы данных и представлены
307 в файле по ссылке <https://disk.yandex.ru/i/vCTMirdmeGgJwQ>.
308 Восстановленные параметры модели (2), (3) могут быть найдены там же.

309 Скорость передачи вируса здесь определяется совокупностью
310 параметров $\alpha_i \frac{5-a}{5}$. Напомним, что параметр a здесь является индексом
311 самоизоляции, рассчитываемым во время пандемии компанией Yandex и
312 изменяющимся от 0 до 5, а значения параметра α_i восстанавливались с
313 помощью решения обратной задачи. Вместе с тем, динамика скорости
314 передачи вируса со сменой разных штаммов определяется трендом,
315 выделенным по числу новых заболевших и отображенным на рисунке 2
316 (синяя линия). Скорость, с которой динамика передачи вируса
317 развивается, представляет собой первую разность от выделенного тренда
318 и должна отображать динамику изменения скорости передачи вируса. На

рисунке 7 отображено сравнение нормированного ряда первой разности от выделенного тренда для числа инфицированных, отображенного на рисунке 2 (синяя линия), и нормированного ряда параметра скорости передачи вируса $\alpha_I \frac{5-a}{5}$.

Из рисунка 7 видно, что по данным детектированных инфицированных рост скорости распространения вируса начался сильно раньше, чем это улавливает обратная задача. Покажем, что динамика, полученная по данным (синяя линия из рисунка 7) лучше отображает распространения вируса, чем параметры, восстановленные с помощью обратной задачи (черная линия из рисунка 7). Положим, что динамика

изменения параметров $\alpha_I \frac{5-a}{5}$ определяется согласно синей линии из

рисунка 7 и подставим новые полученные значения в модель (2),(3). На рисунке 8 приведено сравнение результатов моделирования реальных данных по количеству населения в эпидемиологической группе инфицированных (I) (оранжевая линия), данных, приближенных с помощью модели (2),(3) и динамикой распространения вируса, определяемой через решение обратной задачи (черная линия), и данных, приближенных с помощью модели (2),(3) и динамикой распространения вируса, определяемой через выделенный методом SSA тренд инфицирования населения (синяя линия). На рисунке 8 представлены результаты только для короткого 100-дневного периода, поскольку на полном 2-х летнем промежутке из-за разных пиковых показателей для разных штаммов разница между кривыми визуально не различается. Средняя квадратичная разница на всем промежутке между кривой с реальными данными и кривой, посчитанной как решение модели (2),(3) с восстановленными через решение обратной задачи параметрами – 28.96 (в количестве человек). Средняя квадратичная разница на всем

промежутке между кривой с реальными данными и кривой, посчитанной как решение модели (2),(3) с восстановленными через SSA анализ – 22.86 (в количестве человек).

4. Выводы и обсуждения

В работе предложен пост-анализ данных о наблюдениях за динамикой вируса Sars-Cov-2 в г. Новосибирске за период с 2020 по 2023 гг. Анализ данных проводился с помощью метода SSA. Преимущество такого подхода состоит в том, что метод позволяет разбить исследуемый исходный временной ряд на сумму составляющих временных рядов, каждый из которых ассоциирован с некоторым набором сингулярных чисел траекторной матрицы, построенной из элементов исходного временного ряда. Такой подход позволяет строить декомпозицию временного ряда на составляющие: тренды, сезонности, циклы. Метод был применен к данным по числу ежедневно выявляемых случаев заболеваемости Covid-19, числу критических больных, новым случаям смерти от инфекции и числу госпитализированных за промежуток времени с 2020 по 2022 гг.

Были получены следующие результаты. Во-первых, метод позволил отделить периоды распространения эпидемии, ассоциированные с появлением нового штамма, от сезонного распространения инфекции. Показано, что вклад изменения штамма вируса в рост (или спад) числа новых случаев заболеваемости в 2020-2022 гг. составлял всего порядка 50%, а оставшиеся случаи приходились на “фоновое” распространение вируса с небольшими сезонными вспышками. Во-вторых, выделен тренд, который соответствовал изменению скорости распространения вируса от перехода от одного штамма вируса к другому. Выдвинута гипотеза, что выделенный тренд может быть использован для уточнения параметра контагиозности вируса, используемого в эпидемиологических моделях, которая была подтверждена на примере использования

375 дифференциальной модели SEIR-HCD. Показан наблюдаемый
376 временной лаг между госпитализацией человека и его критическим
377 состоянием.

378 *Благодарности*

379 Работа выполнена в рамках государственного научного задания
380 Институту математики им. С.Л. Соболева СО РАН, проект FWNF-2024-
381 0002.

ТАБЛИЦЫ

Таблица 1. Описание полей анализируемых данных, доступных по ссылке <https://ai-biolab.ru/data/En>

Table 1. Description of the analyzed data, available at the link <https://ai-biolab.ru/data/En>

Имя поля Feature	Описание Description
date	- текущая дата в формате YYYY-MM-DD - current date in YYYY-MM-DD format
new_diagnoses	- количество новых выявленных случаев заражения коронавирусом - the number of newly identified cases of coronavirus infection
cum_diagnoses_article	- официальное общее число выявленных случаев заражения коронавирусом за всё время пандемии - the official total number of identified cases of coronavirus infection throughout the pandemic
cum_diagnoses	- общее число выявленных случаев заражения коронавирусом в регионе за всё время пандемии - the total number of detected cases of coronavirus infection in the region during the entire pandemic
new_recoveries	- количество новых случаев выздоровления человек с начала пандемии

	- the number of new cases of recovery people since the beginning of the pandemic
cum_recoveries	- всего выздоровевших человек с начала пандемии - total number of recovered people
new_deaths	- количество новых смертей от коронавируса - number of new deaths from coronavirus
cum_deaths	- всего умерших от коронавируса с начала пандемии - total deaths number
cum_tests	- всего проведённых тестов на коронавирус в регионе - total number of coronavirus tests carried out in the region
new_tests	- количество новых проведённых тестов на коронавирус в регионе - number of new coronavirus tests conducted in the region
adults	- текущее количество взрослых пациентов с диагнозом Covid-19 и с подозрением на него в инфекционных стационарах - the current number of adult patients diagnosed with Covid-19 and suspected of having it in infectious disease hospitals
cum_children	- всего зафиксировано случаев заражения коронавирусом у детей с начала пандемии - total number of cases of coronavirus infection in children recorded since the start of the pandemic

new_children	- новое количество инфицированных детей - new number of infected children
cur_children	- текущее число госпитализированных детей с подозрением на коронавирус и подтвержденным диагнозом - the current number of children hospitalized with suspected and confirmed coronavirus
n_critical	- текущее количество человек, находящихся в отделениях реанимации - the current number of people in intensive care units
ventilation	- текущее количество человек, находящихся под аппаратом ИВЛ - current number of people on ventilators
hospitalised	- текущее количество госпитализированных человек - current number of hospitalized people
1vac	- текущее количество человек, получивших одну дозу вакцины Спутник-V - the current number of people who have received one dose of the Sputnik-V vaccine
2vac	- текущее количество человек, получивших две дозы вакцины Спутник-V - the current number of people who have received two doses of the Sputnik-V vaccine

positive_percent	- доля инфицированных, имеющая антитела IgG к SARS-CoV-2 - proportion of infected people with IgG antibodies to SARS-CoV-2
yandex_index	- индекс самоизоляции (от компании Yandex) - Self-isolation index (proposed by Yandex)
asympt_percent	- текущий процент бессимптомных выявленных случаев - current percentage of asymptomatic cases detected

РИСУНКИ

Рисунок 1. Процент объяснительной возможности первых 30-ти элементарных матриц для траекторной для числа новых детектированных инфицированных

Figure 1. Percentage of explanatory power of the first 30 elementary matrices for the trajectory for the number of newly detected infected

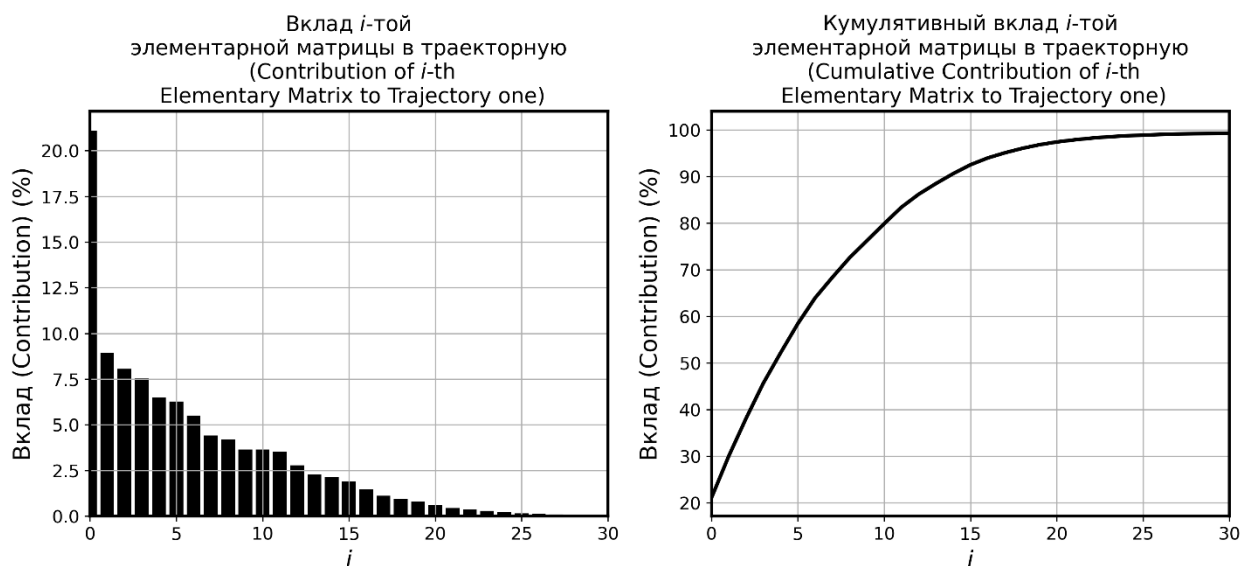


Рисунок 2. SSA анализ временного ряда по числу новых заболевших коронавирусом в г. Новосибирске за период с 2020 по 2023 гг

Figure 2. Caption: SSA analysis of time series of new coronavirus cases in Novosibirsk for the period from 2020 to 2023

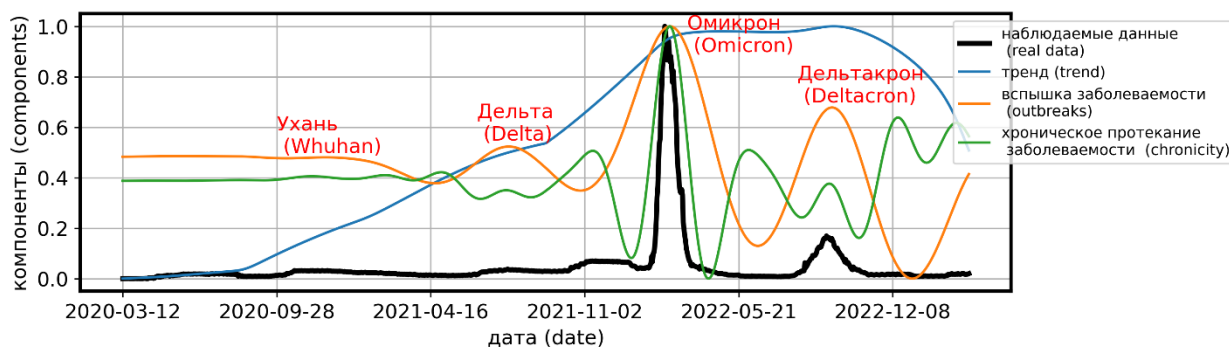


Рисунок 3. Процент объяснительной возможности первых 10-ти элементарных матриц для траекторной для числа новых детектированных умерших от коронавируса

Figure 3. Percentage of explanatory power of the first 10 elementary matrices for the trajectory for the number of newly detected deaths from coronavirus

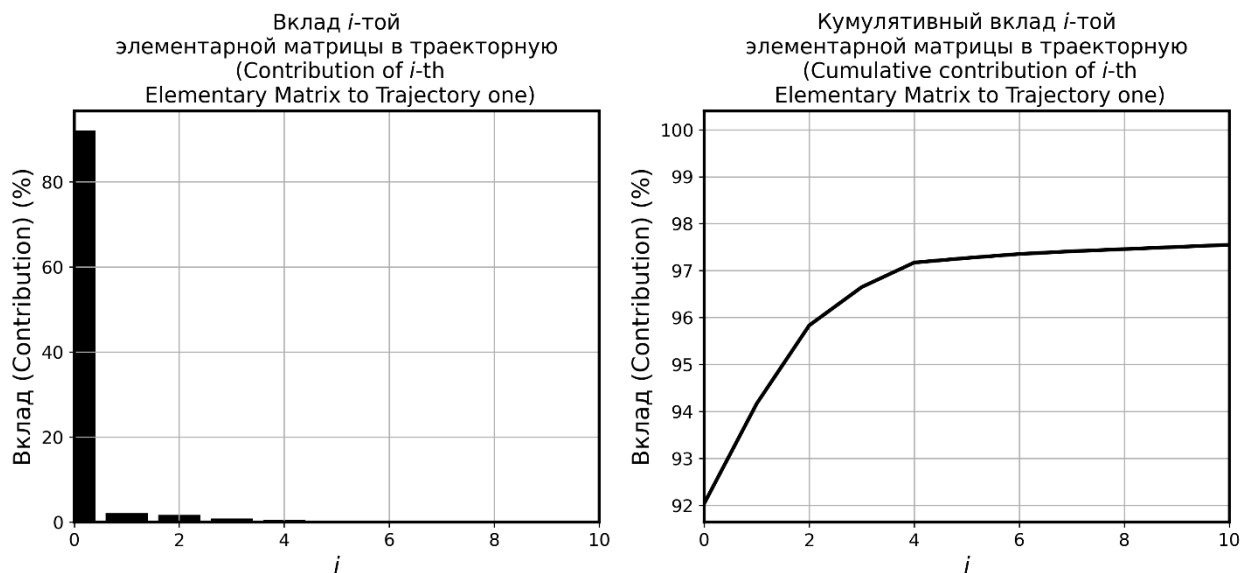


Рисунок 4. Процент объяснительной возможности первых 10-ти элементарных матриц для траекторной для числа новых детектированных критических больных

Figure 4. Percentage of explanatory power of the first 10 elementary matrices for the trajectory for the number of newly detected critically ill patients

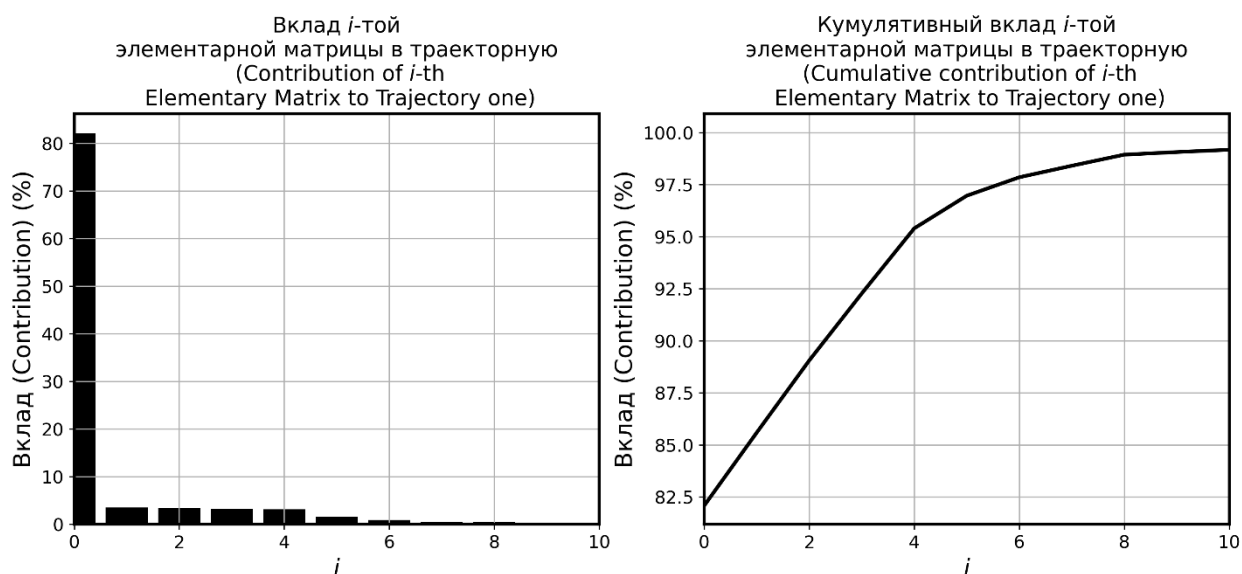


Рисунок 5. Сравнение выделенных трендовых и сезонных компонент для числа критических и умерших больных

Figure 5. Comparison of the obtained trend and seasonal components for the number of critical and deceased patients

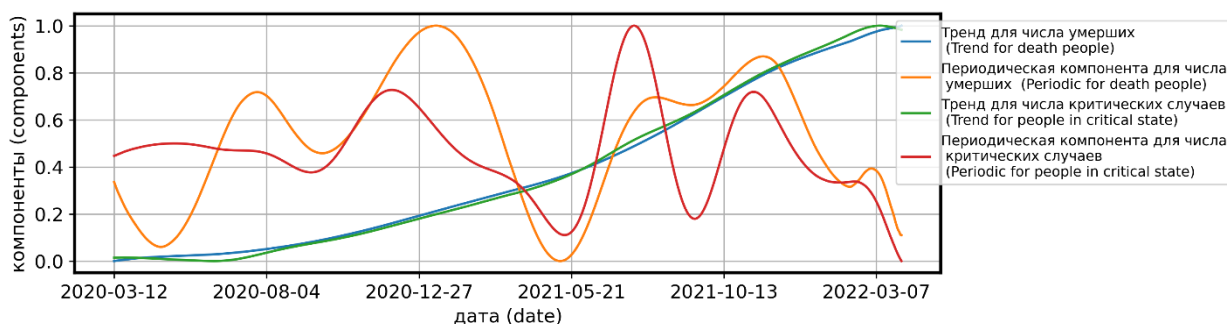


Рисунок 6. Сравнение выделенных трендовых и периодических компонент для числа госпитализированных больных и больных, находящихся на ИВЛ

Figure 6. Comparison of the obtained trend and periodic components for the number of hospitalized patients and patients on mechanical ventilation

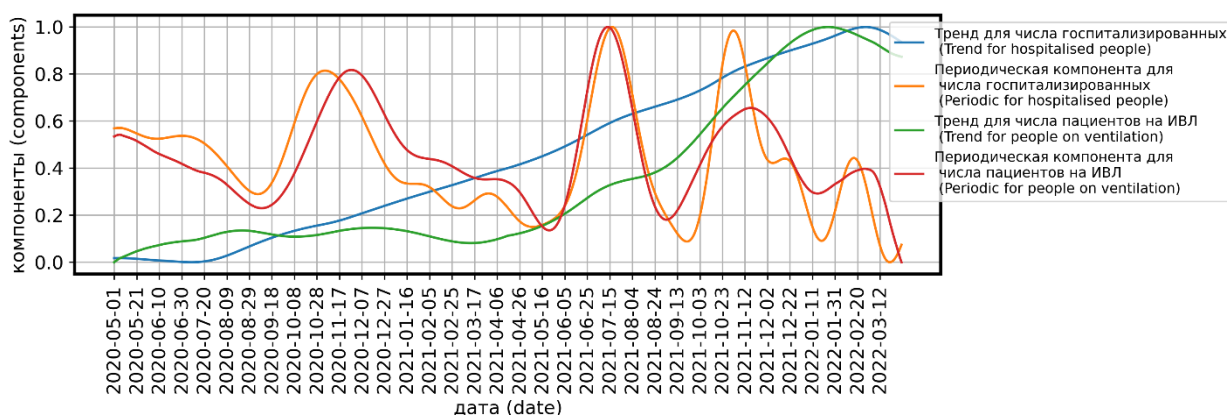


Рисунок 7. Сравнение восстановленной по данным через решение обратной

задачи для модели (2),(3) совокупности параметров $\alpha_I \frac{5-a}{5}$ (скорость передачи вируса, черная линия) и первой разности для ряда, отображающего тренд изменения количества инфицированных в зависимости от развития штамма (синяя линия)

Figure 7. Comparison of the parameter $\alpha_I \frac{5-a}{5}$ (virus transmission rate, black line) reconstructed from the data by solving the inverse problem for model (2), (3) and the first difference for a series displaying the trend in the number of infected people depending on the development of the strain (blue line)

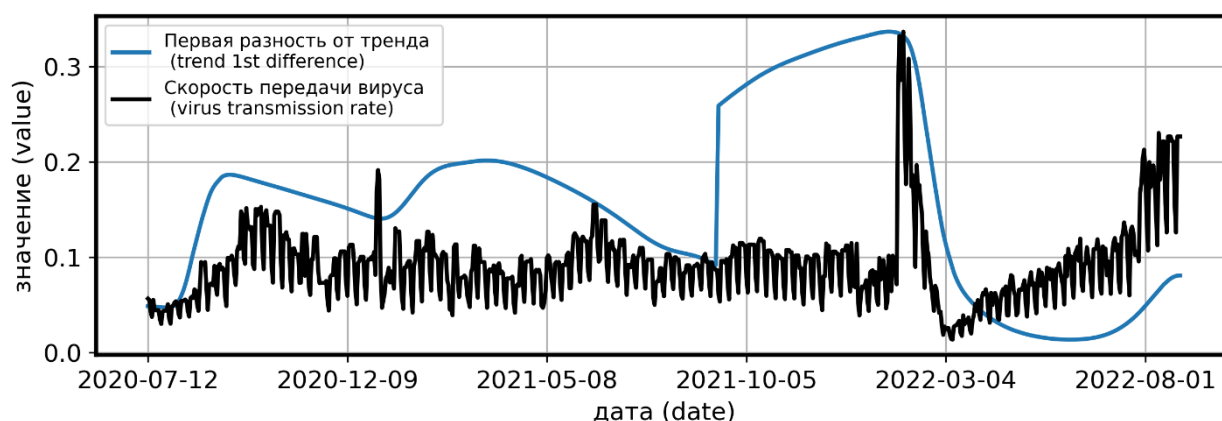
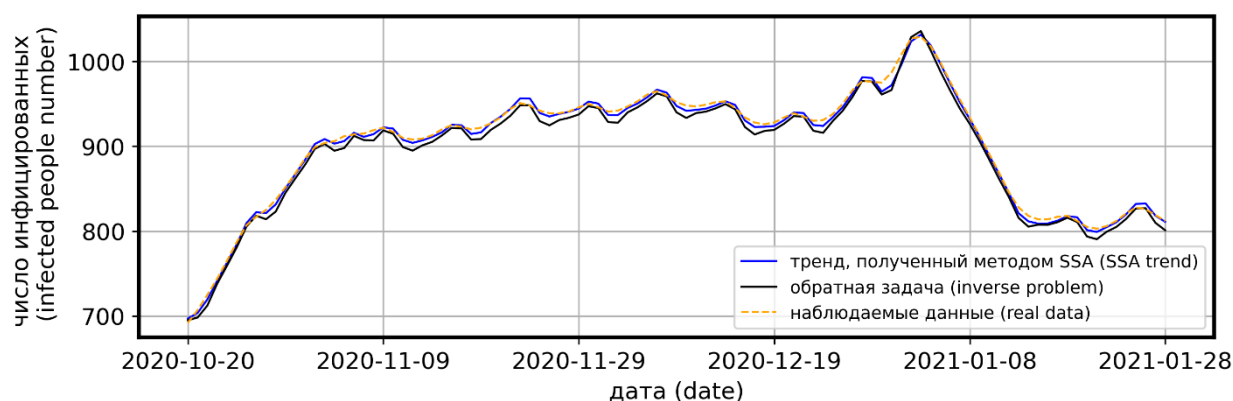


Рисунок 8. Сравнение реальных данных по числу инфицированных с результатами моделирования динамики распространения вируса по модели (2),(3), где скорость распространения вируса восстановлена (а) по тренду, выделенному методом SSA (синяя линия), (б) через решение обратной задачи

Figure 8. Comparison of real data on the number of infected with the results of modeling the dynamics of the spread of the virus using model (2), (3), where the rate of spread of the virus is reconstructed (a) by the trend identified by the SSA method (blue line), (b) by solving the inverse problem



ТИТУЛЬНЫЙ ЛИСТ_МЕТАДАННЫЕ

Блок 1. Информация об авторе ответственном за переписку

Петракова Виктория Сергеевна, кандидат физико-математических наук, старший научный сотрудник Института вычислительного моделирования СО РАН, научный сотрудник Института математики им. Соболева СО РАН;

адрес: 660036, Красноярск, Академгородок, 50/44, ИВМ СО РАН;

факс: +7(391)290–74–76

телефон: +79232673748;

e-mail: vika-svetlakova@yandex.ru

Petrakova Viktoriya Sergeevna, p.h.D, Senior Researcher at the Institute of Computational Modeling SB RAS, Researcher at the Sobolev Institute of Mathematics SB RAS;

address: 660036, Krasnoyarsk, Akademgorodok, 50/44, ICM SB RAS;

fax: +7(391)290–74–76

telephone: +79232673748;

e-mail: vika-svetlakova@yandex.ru

Блок 2. Метаданные статьи

ПОСТАНАЛИЗ ДАННЫХ ПО ЗАБОЛЕВАЕМОСТИ COVID-19 МЕТОДОМ
SSA НА ПРИМЕРЕ НОВОСИБИРСКОЙ ОБЛАСТИ

SSA POST-ANALYSIS OF COVID-19 INCIDENCE NOVOSIBIRSK REGION
DATA USING THE SSA METHOD

Сокращенное название статьи для верхнего колонтитула:

ПОСТАНАЛИЗ ДАННЫХ COVID-19
POST-ANALYSIS OF COVID-19 DATA

Ключевые слова: данные по заболеваемости Covid-19, пост-анализ данных, SSA метод, SEIR-HCD модель, эпидемиологические модели, анализ временных рядов.

Keywords: Covid-19 incidence data, post-analysis of data, SSA method, SEIR-HCD model, epidemiological models, time series analysis.

Оригинальные статьи.

Количество страниц текста – 15,

количество таблиц – 1,

количество рисунков – 8.

16.07.2024

СПИСОК ЛИТЕРАТУРЫ

Порядк овый номер ссылки	Авторы, название публикации и источника, где она опубликована, выходные данные	ФИО, название публикации и источника на английском	Полный интернет-адрес (URL) цитируемой статьи и/или DOI
1	Afzal A. et.al. Merits and limitations of mathematical modeling and computational simulations in mitigation of COVID19 pandemic: A comprehensive review. Archives of Computational Methods in Engineering, 2022, vol. 29, no. 2, pp. 1311-1337.	-	https://doi.org/10.1007/s11831-021-09634-2
2	Alanqary A., Alomar A., & Shah D. Change point detection via multivariate singular spectrum analysis. Advances in Neural Information Processing Systems, 2021, vol. 34, art. 23218-23230.	-	https://proceedings.neurips.cc/paper_files/paper/2021/hash/c348616cd8a86ee661c7c98800678fad-Abstract.html

3	Alharbi N. Forecasting the COVID-19 Pandemic in Saudi Arabia using a modified singular spectrum analysis approach: model development and data analysis. Jmirx med, 2021, vol. 2, no. 1, art. e21044.	-	https://xmed.jmir.org/2021/1/e21044/
4	Alsaeed N. I., Eman Y., and Muazzam A. An Agent-based Simulation of the SIRD model of COVID19 Spread. International Journal of Biology and Biomedical Engineering, 2020, vol. 14, pp. 210-217.	-	https://doi.org/10.46300/91011.2020.14.28
5	Balderrama R., et.al. Optimal control for a SIR epidemic model with limited quarantine. Scientific Reports, 2022, vol. 12, no. 1, art. 12583.	-	https://doi.org/10.1038/s41598-022-16619-z
6	Cao L., Qing L. COVID-19 modeling: A review. ACM Computing Surveys, 2024, vol. 57, no. 1, pp. 1-42.	-	https://doi.org/10.1145/3686150

7	Chu J. A statistical analysis of the novel coronavirus (COVID-19) in Italy and Spain. PloS one, 2021, vol. 16, no. 3, art. e0249037.	-	https://doi.org/10.1371/journal.pone.0249037
8	de Souza F.S.H., et.al. On the analysis of mortality risk factors for hospitalized COVID-19 patients: A data-driven study using the major Brazilian database. PloS one, 2021, vol. 16, no. 3, art. e0248580.	-	https://doi.org/10.1371/journal.pone.0248580
9	Desson Z., Weller E., McMeekin P., Ammi M. An analysis of the policy responses to the COVID-19 pandemic in France, Belgium, and Canada. Health Policy and Technology, 2020, vol. 9, no. 4, pp. 430-446.	-	https://doi.org/10.1016/j.hlpt.2020.09.002
10	Elsner J. B., Tsonis A. A. Singular spectrum analysis: a new tool in time series analysis. Springer Science & Business Media, 2013.	-	https://link.springer.com/book/10.1007/978-1-4757-2514-8

11	Er A. G., et.al. Multimodal data fusion using sparse canonical correlation analysis and cooperative learning: a COVID-19 cohort study. NPJ Digital Medicine, 2024, vol. 7, no. 1, art. 117.	-	https://doi.org/10.1038/s41746-024-01128-2
12	Fanelli D., & Piazza F. Analysis and forecast of COVID19 spreading in China, Italy and France. Chaos, Solitons & Fractals, 2020, vol. 134, art. 109761.	-	https://doi.org/10.1016/j.chaos.2020.109761
13	Golyandina N. Particularities and commonalities of singular spectrum analysis as a method of time series analysis and signal processing. Wiley Interdisciplinary Reviews: Computational Statistics, 2020, vol. 12, no. 4, art. e1487.	-	https://doi.org/10.1002/wics.1487
14	Kalantari M. Forecasting COVID-19 pandemic using optimal singular spectrum analysis. Chaos, Solitons & Fractals, 2021, vol. 142, art. 110547.	-	https://doi.org/10.1016/j.chaos.2020.110547

15	Kermack W.O., McKendrick A.G. A contribution to the mathematical theory of epidemics. Proceedings of the royal society of London. Series A, 1927, vol. 115, no. 772, pp. 700-721.	-	https://doi.org/10.1098/rspa.1927.0118
16	Krivorotko O., et al. Mathematical modeling and prediction of COVID-19 in Moscow city and Novosibirsk region, arXiv preprint, 2020, arXiv:2006.12619.	-	https://arxiv.org/abs/2006.12619
17	Krivorotko O., Zvonareva T., Neverov A. Identifiability of the spatial SEIR-HCD model of COVID-19 propagation. arXiv preprint, 2024, arXiv:2412.18858.	-	https://arxiv.org/abs/2412.18858
18	Li W.T., et.al. Using machine learning of clinical data to diagnose COVID-19: a systematic review and meta-analysis. BMC medical informatics and decision making, 2020, vol. 20, pp. 1-13.	-	https://doi.org/10.1186/s12911-020-01266-z
19	Ma Ch., et.al. Understanding dynamics of pandemic models to support predictions of COVID-19 transmission: parameter	-	https://doi.org/10.1109/JBHI.2022.3168825

	sensitivity analysis of SIR-type models. IEEE Journal of Biomedical and Health Informatics, 2022, vol. 26, no. 6, pp. 2458-2468.		
20	Nichita D. R., Dima M., Boboc L., & Hancean M.G. Data analysis evidence beyond correlation of a possible causal impact of weather on the COVID-19 spread, mediated by human mobility. Scientific Reports, 2024, vol. 14, no. 1, art. 17782.	-	https://doi.org/10.1038/s41598-024-67918-6
21	Petrakova V. S. Planner model for estimating the dynamic of epidemic spread under limited resources. Eurasian journal of mathematical and computer applications, 2025, vol. 13, no. 1, pp. 89-99.	-	https://doi.org/10.32523/2306-6172-2025-13-1-89-99
22	Petrakova V., Krivorotko O. Comparison of Two Mean Field Approaches to Modeling an Epidemic Spread. Journal of Optimization Theory and Applications, 2025, vol. 204, no. 3, art. 39.	-	https://doi.org/10.1007/s10957-024-02604-1

23	Petrakova V., Krivorotko O. Sensitivity of MFG SEIR-HCD epidemiological model. Lobachevskii Journal of Mathematics, 2023, vol. 44, no. 7, pp. 2856–2869.	-	https://doi.org/10.1134/S199508022307034X
24	Prakash K.B., et al. Analysis, prediction and evaluation of Covid-19 datasets using machine learning algorithms. International Journal., 2020, vol. 8, no. 5, pp. 2199-2204.	-	https://doi.org/10.30534/ijeter/2020/117852020
25	Qorib M., Oladunni T., Denis M., Ososanya E., & Cotaе, P. Covid-19 vaccine hesitancy: Text mining, sentiment analysis and machine learning on COVID-19 vaccination Twitter dataset. Expert Systems with Applications, 2023, vol. 212, art. 118715.	-	https://doi.org/10.1016/j.eswa.2022.118715
26	Ram V., Schaposnik L.P. A modified agestructured SIR model for COVID-19 type viruses. Scientific reports, 2021, vol. 11, no. 1, art. 15194.	-	https://doi.org/10.1038/s41598-021-94609-3

27	Riedmann U., et.al. Estimates of SARS-CoV-2 infections and population immunity after the COVID-19 pandemic in Austria: Analysis of national wastewater data. The Journal of Infectious Diseases, 2025, vol. 231, no. 5, art. e921-e928.	-	https://doi.org/10.1093/infdis/jiaf054
28	Shuja J., Alanazi E., Alasmary W., & Alashaikh, A. COVID-19 open source data sets: a comprehensive survey. Applied Intelligence, 2021, vol. 51, no. 3, pp. 1296-1325.	-	https://doi.org/10.1007/s10489-020-01862-6
29	Sulandari W., et.al. SSA-based hybrid forecasting models and applications. Bulletin of Electrical Engineering and Informatics, 2020, vol. 9, no. 5, art. 2178-2188.	-	https://doi.org/10.11591/eei.v9i5.1950
30	Taylan O., Alkabaa A.S., & Yılmaz M. T. Impact of COVID-19 on G20 countries: analysis of economic recession using data mining approaches. Financial Innovation, 2022, vol. 8, no. 1, art. 81.	-	https://doi.org/10.1186/s40854-022-00385-y

31	Xu B., et.al. Epidemiological data from the COVID-19 outbreak, real-time case information. Scientific data, 2020, vol. 7, no. 1, art. 106.	-	https://doi.org/10.1038/s41597-020-0448-0
32	Yi S., Sneeuw N. Filling the data gaps within GRACE missions using singular spectrum analysis. Journal of Geophysical Research: Solid Earth, 2021, vol. 126, no. 5, art. e2020JB021227.	-	https://doi.org/10.1029/2020JB021227