DOI: 10.15514/ISPRAS-2024-36(2)-13

Пространственное POD-разложение эпидемиологических данных COVID-19

С.А. Елистратов, ORCID: 0000-0002-7006-6879 <*sa.elist-ratov@yandex.ru*>

Институт системного программирования им. В.П. Иванникова РАН,
109004, Россия, г. Москва, ул. А. Солженицына, д. 25,
Институт математики им. С.Л. Соболева СО РАН,
630090, Россия, г. Новосибирск, ул. Академика Коптюга, 4.

Аннотация. В работе исследуются пространственно-временные ряды основных эпидемиологических показателей COVID-19 (распространенность, смертность, показатель выздоровления) для различных регионов России. С целью выявления пространственной корреляции применено POD-разложение, выделены основные моды, получены соответствующие временные зависимости; к последним применено шумоподавление с помощью Empirical Mode Decomposition. Показано, что вследствие разного характера временных коэффициентов для исследуемых параметров совместное POD-разложение нецелесообразно. Исследована сходимость разложения к исходным данным в зависимости от числа мод разложения; выявлен экспоненциальных характер этой зависимости.

Ключевые слова: эпидемиология; COVID-19; POD; EMD.

Для цитирования: Елистратов С.А. Пространственное POD-разложение эпидемиологических данных COVID-19. Труды ИСП РАН, том 36, вып. 2, 2024 г., стр. 181–192. DOI: 10.15514/ISPRAS–2024–36(2)–13.

Благодарности: Работа выполнена при поддержке Российского научного фонда (проект №23-71-10068).

COVID-19 Epidemiological Indicators POD Spatial Decomposition

S.A. Elistratov, ORCID: 0000-0002-7006-6879 <*sa.elist-ratov@yandex.ru*>

Institute for System Programming of the Russian Academy of Sciences,
25, Alexander Solzhenitsyn st., Moscow, 109004, Russia,
Sobolev Institute of Mathematics, Siberian Branch of the Russian Academy of Sciences,
4, Acad. Koptyug avenue, Novosibirsk, 630090, Russia.

Abstract. The epidemiological indicators (prevalence, mortality, convalescence) for COVID-19 are investigated as a temporary-spatial dependencies. Proper Orthogonal Decomposition is applied for the first time to this type of data; the main modes and corresponding coefficients are obtained. Due to this method, it is shown that there are modes concentrated in the particular regions which means there are independent factors for disease spreading. Additionally, it is showed that Empirical Mode Decomposition can be successfully applied for noise-reduction and better understanding time dependencies. The exponential nature of the decomposition error decree shows the accuracy of the decomposition. Despite the ability of POD to reveal hidden dependencies, it requires rows of simultaneous data, which in fact may not be so. In the article the correction applied is discussed, however it may not be enough because of mistakes in raw data. The method is recommended to use unless the data it is applied to is inaccurate.

Keywords: epidemiology; COVID-19; POD; EMD.

For citation: Elistratov S.A. COVID-19 epidemiological indicators POD spatial decomposition. *Trudy ISP RAN/Proc. ISP RAS*, vol. 36, issue 2, 2024. pp. 181-192 (in Russian). DOI: 10.15514/ISPRAS-2024-36(2)-13.

Acknowledgements. The investigation was conducted under the help of RSF (project №23-71-10068).

1. Введение

Установление пространственной корреляции между эпидемиологическими показателями является важной задачей при оценке рисков и осуществлении мер по противодействию распространению инфекционных заболеваний. В настоящее время прогностические модели развиты слабо, поэтому важную роль в изучении этого вопроса играет обработка и интерпретация полевых данных.

Основные модели распространения инфекционных заболеваний сосредоточены [1-2] на эволюции показателей в конкретной области; существующие модели пространственной передачи инфекции основаны на клеточных автоматах и описывают небольшие модельные объекты с высокой плотностью и простой геометрией [3-5]. Интересным представляется подход [6] на основе имитационного моделирования, однако его применение для разбиения на большое число регионов представляется затруднительным.

Учесть пространственную динамику в дифференциальных уравнениях не представляется возможным вследствие бесструктурного расположения населенных пунктов, а модели сплошной среды в масштабах регионов неприменимы из-за малого размера последних по сравнению с расстояниями между ними.

Тем не менее, оказывается возможным связать показатели заболеваемости во времени и пространстве с помощью POD-разложения. Этот метод уже зарекомендовал себя в динамике сплошных сред, где используется для выделения характерных мод течений; однако область его применения не ограничивается задачами с непрерывной средой. Метод лишен необходимости использовать какие-либо сетки и воспринимает зависимости в разных точках как именованные временные ряды, что позволяет применять его для записей данных в ячейках разного размера, разделенных сложной границей, роль которых в текущем исследовании играют субъекты административного деления. Такое разложение позволяет отследить пространственную корреляцию между данными из разных регионов и выделить характерные моды эволюции эпидемиологических показателей.

2. Методы

2.1 Методы разложения

В общем виде POD-разложение позволяет представить двумерный массив данных в виде суммы слагаемых с разделенными индексами:

$$u_{ij} = \sum_k T_i^k \Phi_j^k \quad (1)$$

Для разложения используется следующая процедура: составляется матрица ковариации $C = u^T u$, собственные вектора которой являются векторами разложения Φ^k. Собственные значения называются энергиями мод и отражают вклад каждой компоненты в разлагаемую функцию; в случае разложения физического поля скорости они пропорциональны кумулятивной кинетической энергии моды. Предполагается, что собственные вектора разложения ортогональны, поэтому временные коэффициенты можно найти путем умножения исходной матрицы $[u_{ij}] \equiv U$ на транспонированную матрицу мод. Для возможности сравнения временных коэффициентов моды разложения нормируют в L^2, что в тоже время делает конкретные числовые значения коэффициентов весьма условными из-за масштабирования.

В случае больших данных поиск собственных значений может занимать длительное время; однако в силу специфики матрицы ковариации те же моды и энергии могут быть получены путем не как собственных вектора матрицы ковариации C, а как сингулярные вектора исходной матрицы U, а энергии – как квадраты сингулярных чисел. В самом деле, пусть ξ и η – правый и левый вектора сингулярного разложения U соответственно; σ – сингулярное число:

$$U\xi = \sigma\eta;\ U^T\eta = \sigma\xi$$

Умножая слева на U^T, получим:

$$U^T U\,\xi = \sigma U^T\eta = \sigma^2\xi$$

или

$$C\xi = \sigma^2\xi$$

Сингулярное разложение (SVD) быстрее поиска собственных значений, что позволяет сократить время преобразования. Тем не менее, этот алгоритм является довольно затратным в смысле оперативной памяти, и данные с большим числом признаков могут приводить к ее переполнению. Для ее решения может быть применен алгоритм с использованием машинного обучения [7], для которого требуется лишь память под необходимое число мод, однако в случае, когда число пространственных точек невелико, предпочтительным является алгоритм SVD как более точный.

В непрерывных переменных разложение (1) можно интерпретировать следующим образом: первый индекс соответствует времени, второй – некоторому признаку:

$$u(x,t) \approx \sum_k T^k(t)\,\Phi^k(x), \qquad (2)$$

где t – время, x – обобщенная координата. При этом функции $T^k(t)$ называются временными коэффициентами разложения, а $\Phi^k(x)$ – модами. Поскольку само разложение (1) никак не ограничивает выбор параметра x, это могут быть как именованные признаки, так и координаты; кроме того, возможно совместное разложение нескольких признаков по координатам (например, разных компонент скорости [8]). POD-разложение по координатам находит широкое применение в задачах механики сплошных сред [9-13] для выделения пространственных мод течения; случаи применения пространственного POD-разложения в эпидемиологии неизвестны. Поскольку разложение не накладывает ограничение на признак, для координатного разложения не требуется специальная сетка. Это преимущество позволяет выделить в качестве мод региональное распределение исследуемой функции.

2.2 Данные и предобработка

В качестве данных использовались данные о заболеваемости COVID-19 с марта 2020 по октябрь 2023 года. Эти данные содержат распределение числа заражений, выздоровлений и смертей, а также их кумулятивные показатели по регионам России. Из-за наличия в данных ошибочных записей (отрицательное число новых смертей, убывание кумулятивных показателей и т.д.) была проведена их предобработка: записи (ряд данных для определенной даты в конкретном регионе) с отрицательными дифференциальными данными удалялись, и кумулятивные данные восстанавливались по дифференциальным (был использован такой подход, потому что ошибки в кумулятивных данных ведут к накоплению ошибки, и ошибка в одной записи требует реконструкции всех последующих). Из-за того, что сделаны в разных регионах могут быть записаны не каждый день и приходятся на разные даты, был определен общий для всех период наблюдения (с 17 апреля 2020 по 24 сентября 2023 года) и проведена процедура интерполяции на посуточную сетку, поскольку для POD-разложения требуется матрица одновременных записей. Затем кумулятивные данные дифференцировались для восстановления дифференциальных величин. Полученная выборка содержит 1255 срезов по времени. Сразу оговоримся, что будем рассматривать только дифференциальные (суточные) показатели, поскольку для них временная динамика видна более наглядно.

Пространственные моды у соответствующих суточных и кумулятивных показателей одинаковы, поскольку они связаны суммированием только по времени (см. выражение (2)).

3. Результаты

3.1 Распространенность

Разложение заболеваемости (числа новых случаев заражения) оказалось нерепрезентативным, поскольку основной вклад приходится на первую моду (>85%), которая целиком сосредоточена на крупных городах (Москва и Санкт-Петербург на рис. 1). Это происходит из-за неодинаковой численности населения в разных регионах. Поэтому для POD-разложения имеет смысл рассматривать не заболеваемость, а распространенность (число новых случаев заражения на 100.000 человек); дальнейшие данные также рассматриваются нормированными по населению региона.

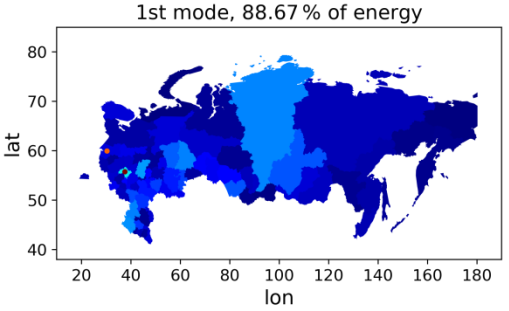

Рис. 1. Первая мода разложения заболеваемости
Fig. 1. First POD mode of incidence

На рис. 2 представлены временные коэффициенты разложения. Первая мода с наиболее высоким (64%) вкладом в разложение несмотря на нормировку имеет довольно неравномерное распределение числа новых случаев заражения (рис. 4), причем наибольшее значение наблюдаются в Ямало-Ненецком автономном округе, в то время как в Москве наблюдается довольно низкая заболеваемость. Причина такой высокой заболеваемости в ЯМАО неясна и требует отдельного исследования.

Отметим, что несмотря на неотрицательность исходного временного ряда коэффициенты POD-разложения могут принимать отрицательные значения, например, мода гашения. Такая мода уменьшает значение исследуемой функции на своем носителе по сравнению с более высокоэнергетическими компонентами.

В случае анализа числа заболеваний подобное поведение присуще моде 2 со значительным вкладом (17%), которая сосредоточена на Москве, Санкт-Петербурге и Якутии. Мода была активна в первом квартале 2022 года; в то же время наблюдался максимум временного коэффициента основной моды. Интересно, что в период подъема заболеваемости мода 2 имеет положительные значения (то есть заболеваемость в этих регионах росла быстрее), а в период спада, наоборот, отрицательные, что говорит о том, что процесс роста и спада заболеваемости происходит быстрее на фоне средней картины.

Зашумленность данных проявляется в виде высокочастотных осцилляций временных коэффициентов. Сами по себе осцилляции с характерным периодом в несколько дней представляют мало интереса, поэтому разумно будет выделить основную динамику. Осреднение в окнах не является эффективным, поскольку не осредняет должным образом функции с резкими изменениями.

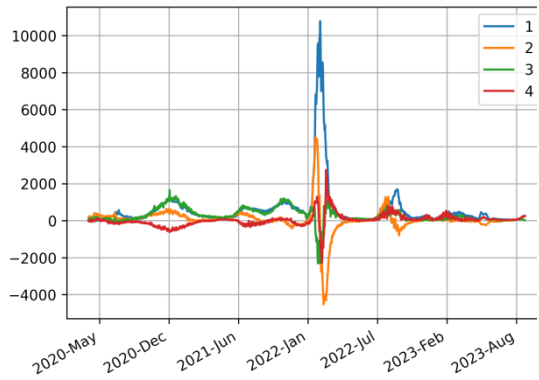

Рис. 2. Временные коэффициенты разложения для распространенности

Fig.2 Temporal POD coefficients of prevalence

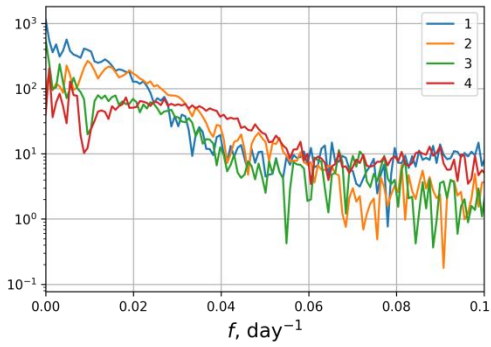

Рис. 3. Спектры временных коэффициентов

Fig.3. Prevalence POD coefficients spectra

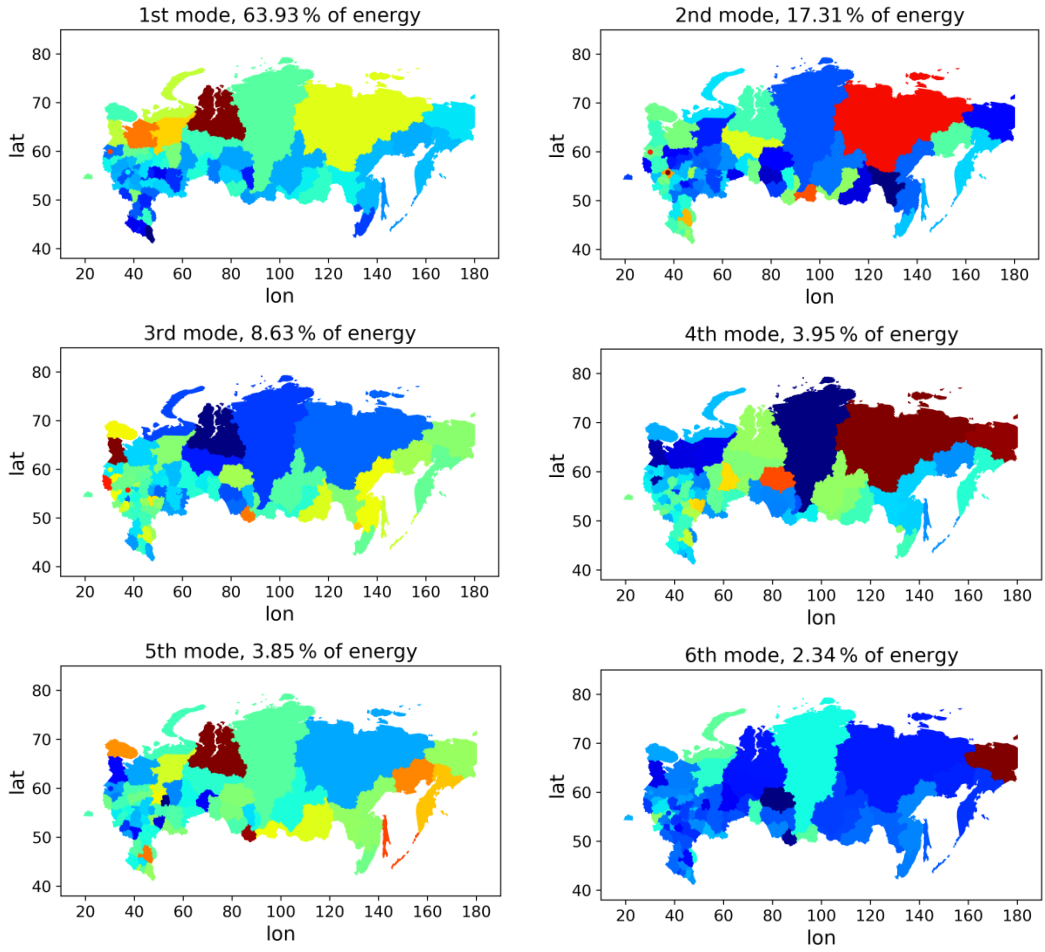

Рис. 4. Моды разложения распространенности

Fig. 4. POD modes of prevalence

Чтобы избавиться от высокочастотного шума, предлагается использовать разложение по эмпирическим модам (преобразование Гильберта-Хуанга, EMD [14]). Отдельные моды не могут быть использованы сами по себе, однако учитывая, что моды выделяются, начиная с наиболее высокочастотной компоненты, для осреднения можно вычесть первые несколько эмпирических мод из исходного временного ряда. Результат такого преобразования для временных коэффициентов разложения зависимости числа новых заболеваний представлен на рис. 5 (вычитались первые 3 моды); устранение осцилляций позволяет сделать этот график более наглядным (ср. с рис. 2). При этом такое осреднение ведет к выполаживанию спектра в области высоких частот (рис. 3, 6); тем не менее, EMD в том виде, в котором оно используется, является лишь способом улучшить репрезентативность временных коэффициентов и никак не сказывается на пространственных модах.

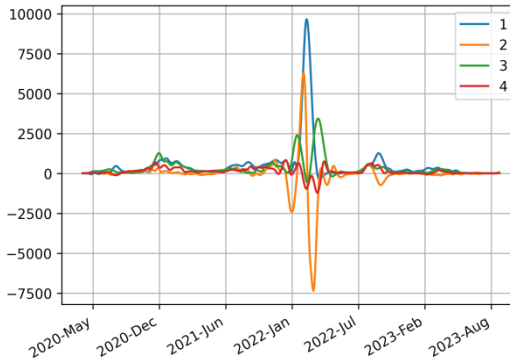

Рис. 5. Временные коэффициенты разложения для распространенности (EMD-фильтрация)

Fig.5 Temporal POD coefficients of prevalence (with EMD filtration applied)

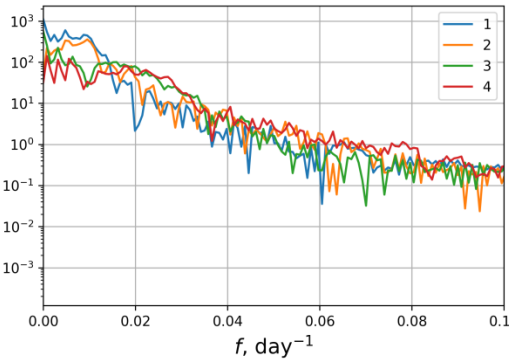

Рис. 6. Спектры временных коэффициентов (EMD-фильтрация)

Fig.6 Temporal POD coefficients of prevalence (with EMD filtration applied)

3.2 Смертность

Несмотря на то, что датасет содержит несколько эпидемиологических показателей, а POD-разложение позволяет проводить процедуру декомпозиции для нескольких разлагаемых функций, мы намеренно не делаем такое разложение на эпидемиологических данных. Дело в том, что совместное разложение имеет жесткое условие в виде общего временного коэффициента:

$$\begin{pmatrix} u(x,t) \\ v(x,t) \end{pmatrix} \approx \sum_k \begin{pmatrix} \Phi_u^k(x) \\ \Phi_v^k(x) \end{pmatrix} T^k(t) \qquad (3)$$

где u и v — две совместно разлагаемые функции. Из вида разложения (3) видно, что несмотря на разные пространственные составляющие временной коэффициент у них общий. Поэтому такое разложение, например, для числа новых случаев заражения и числа летальных исходов, вообще говоря, имеет мало смысла, поскольку даже при абсолютной корреляции между заражениями и смертями смертельный исход не наступает одновременно с заражением.

Временные коэффициенты разложения и соответствующие осредненные кривые представлены на рис. 7-8. Видно, что они не имеют такой четкой структуры, как коэффициенты разложения числа заражения, и более наглядно демонстрируют работу EMD.

Первая мода разложения (рис. 9) имеет очень неравномерный характер и максимумы совсем не в тех регионах, что заболеваемость. Вторая мода сосредоточена на Ненецком автономном округе, а характер соответствующего коэффициента указывает на то, что смертность в этом регионе в период повышенной смертности в целом по стране (с июня 2021 по март 2022 года)

значительно колебалась. Причины такого поведения в одном конкретном регионе неизвестны и требуют исследования факторов, не включенных в датасет.

Различие в пространственном распределении этих мод для заболеваемости и смертности свидетельствует о том, что на эти показатели влияют разные региональные факторы, а также подтверждает, что они должны быть исследованы отдельно друг от друга.

Третья и четвертая мода, локализованы в отдельных восточных регионах (рис. 9), что позволяет предположить наличие в них независимых факторов, влияющих на смертность.

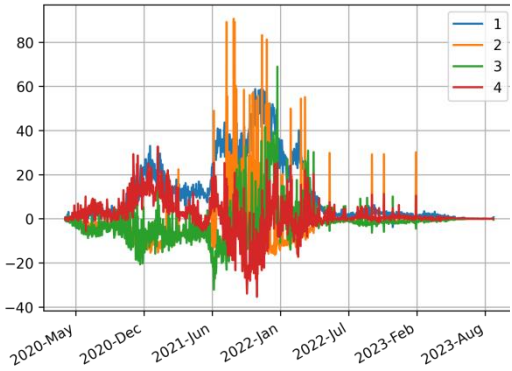

Рис. 7. Временные коэффициенты разложения смертности
Fig.7 Temporal POD coefficients of mortality

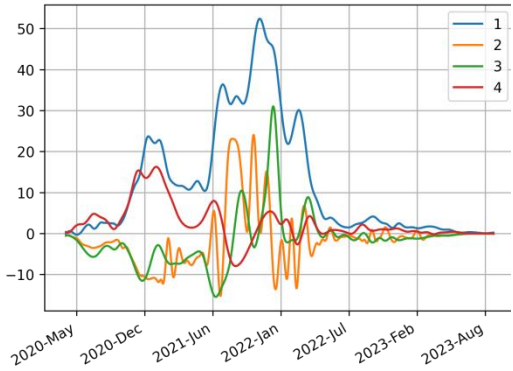

Рис. 8. Временные коэффициенты разложения смертности (EMD-фильтрация)
Fig.8 Temporal POD coefficients of mortality (with EMD filtration applied)

3.3 Показатель выздоровления

Пространственные моды числа выздоровлений (рис. 11) вкупе с их временными коэффициентами (рис. 14) обнаруживают сходство этого показателя с числом новых заболеваний и свидетельствует о пространственной корреляции между ними; корреляция между заболеваемостью/выздоровлениями и смертностью довольно слабая. Это может быть вызвано тем, что период выздоровления варьируется незначительно, а время от заболевания до смерти в значительной степени индивидуально для каждого летального пациента. Тем не менее, несмотря на схожесть мод и временных коэффициентов упомянутых эпидемиологических показателей, распределение энергий для их разложения (рис. 12-13) различается.

3.4 Сходимость разложения

Выделение первых мод дает представление о поведении решения и основных трендах его эволюции, однако если POD-декомпозиция применяется для сжатия данных (при хранении первых n мод и их временных коэффициентов) необходимо исследовать сходимость такого разложения к исходным данным. Для этого рассмотрим ошибку в виде:

$$error = \left\| \sum_{k=1}^{n} T_i^k \Phi_j^k - U_{ij} \right\|_2^2 / \left\| U_{ij} \right\|_2^2$$

где U — исследуемая совокупность признаков, как функцию от числа рассматриваемых мод n, $\|\cdot\|_2$ – квадратичная норма по времени и пространству.

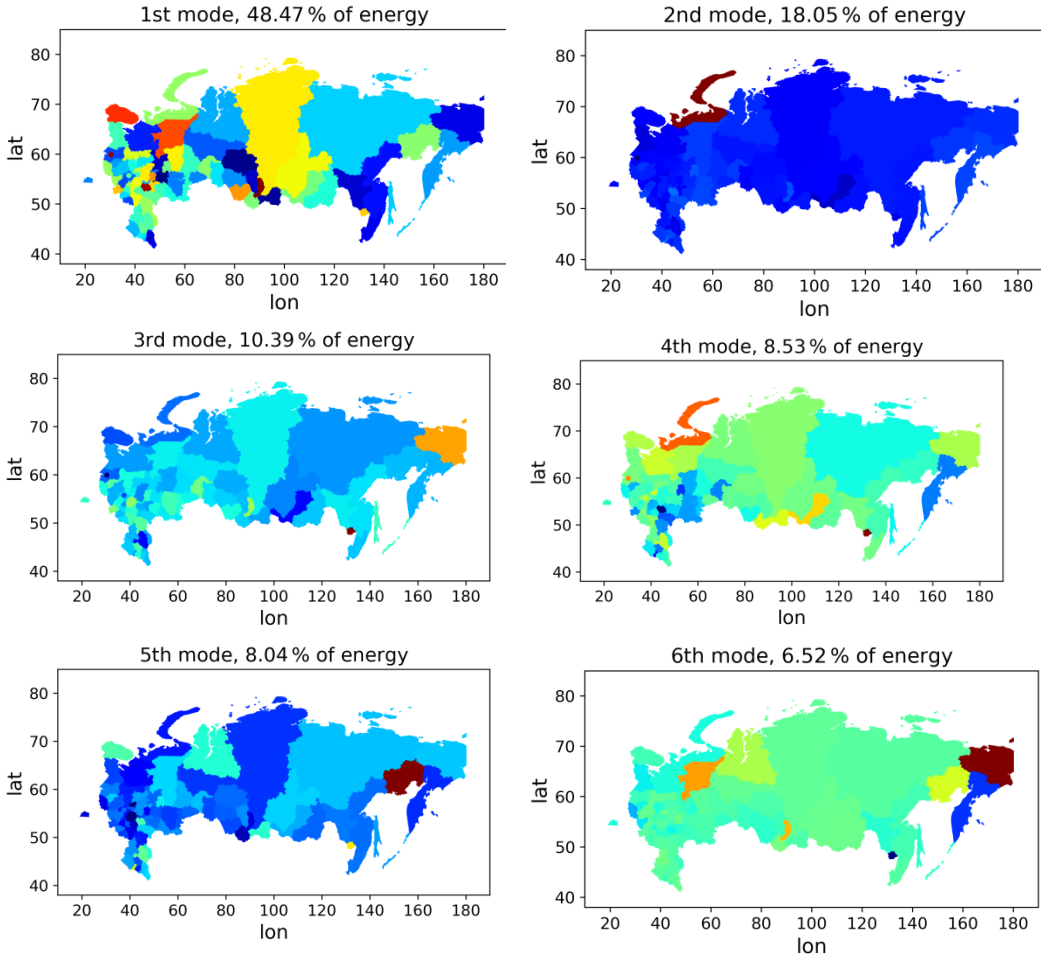

Рис. 9. Моды разложения смертности
Fig.9 POD modes of mortality

Рис. 10. Временные коэффициенты показателя выздоровления
Fig.10 Temporal POD coefficients of convalescence)

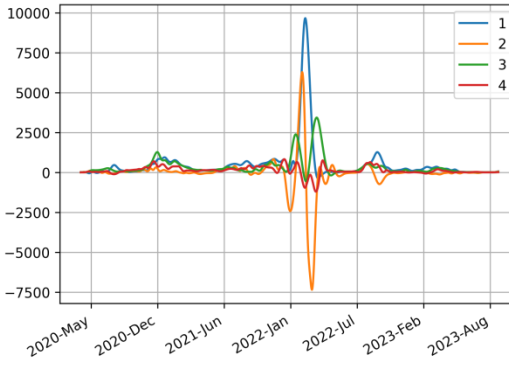

Рис. 11. Временные коэффициенты разложения показателя выздоровления (EMD-фильтрация)
Fig.11 Temporal POD coefficients of convalescence (with EMD filtration applied)

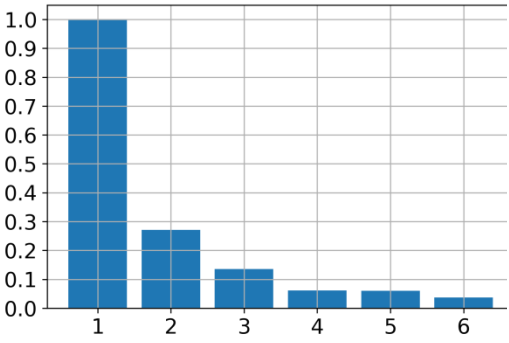

Рис. 12. Энергии мод разложения распространенности

Fig.12 Prevalence POD eigenvalues

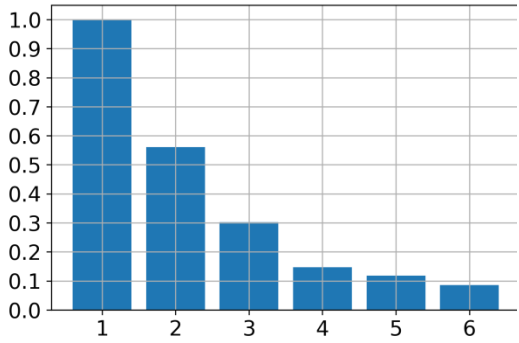

Рис. 13. Энергии мод разложения показателя выздоровления

Fig.13 Convalescence POD eigenvalues

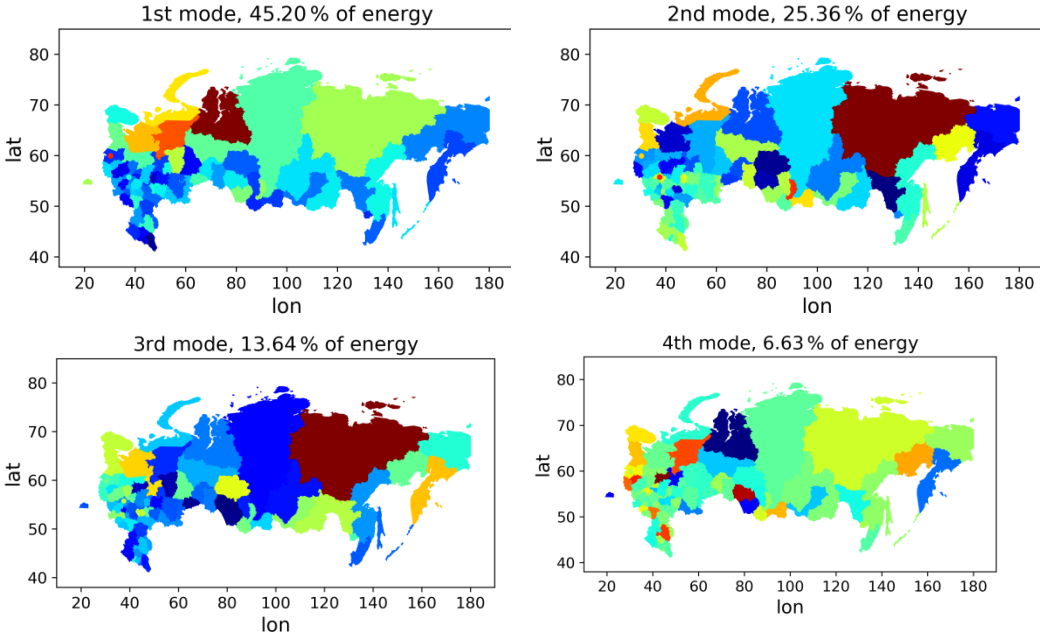

Рис. 14. Моды разложения показателя выздоровления

Fig.14. POD modes of convalescence

На рис. 15 представлены зависимости ошибки для разложений отдельных параметров (распространенность, смертность, показатель выздоровления). Рассматривать число мод >85 не имеет смысла, поскольку это полное число регионов. Видно, что на большей части интервала затухание ошибки имеет экспоненциальный характер — на графике приведены аппроксимирующие зависимости. Показатели затухания (множители в экспоненте) близки, однако неодинаковы, причем наиболее отличается показатель не для смертности, а для распространенности. Экспоненциальный характер затухания мод свидетельствует о том, что POD может эффективно применяться для компрессии этих данных. О неэффективности совместного разложения также свидетельствует рис. 16, показывающий корреляцию первых десяти мод разложений трех эпидемиологических показателей. Видно, что вне диагонали довольно слабая корреляция наблюдается между первыми модами распространенности и показателя выздоровления, что вполне логично; однако в остальном корреляция не

превышает 0.5, что можно расценивать как отсутствие линейной связи между модами различных показателей.

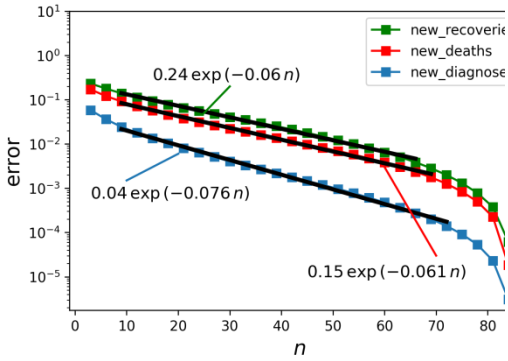

Рис. 15. Ошибка разложения
Fig.15 Decomposition error

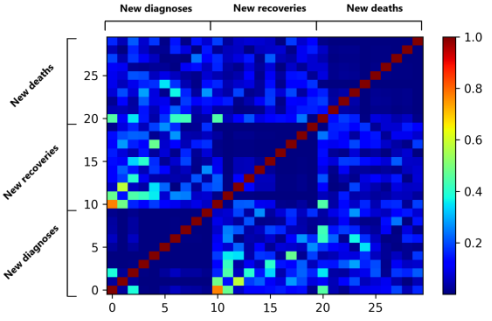

Рис. 16. Коэффициенты корреляции мод различных показателей
Fig.16 Different indicators modes correlation

4. Заключение

POD-разложение применяется в эпидемиологии довольно редко, а его пространственный вариант рассматривается впервые. Несмотря на это, POD является достаточно удобным инструментом при анализе эпидемиологических данных и позволяет проследить их связь. На примере трех рассматриваемых показателей (показателей заболеваемости, смертности и выздоровления) показано, что поведение смертности значительно отличается от поведения остальных показателей, что может быть обусловлено внутренними причинами этих явлений. Несмотря на то, что POD-разложение не может объяснить эти причины, анализ данных с его помощью поможет скорректировать направление исследований эпидемиологов для более глубокого понимания механизма распространения заболеваний.

К числу недостатков метода можно отнести, во-первых, то, что выявленные таким способом зависимости являются сугубо линейные, и в случае нелинейного характера отношений между различными показателями необходимо применять более алгоритмы. Во-вторых, несмотря на отсутствие привязки к каким-либо пространственным сеткам, для разложения необходимы ряды единовременных записей, чего на практике может не быть. Данные об эпидемиологических показателях зачастую нерегулярны, а порой содержат откровенно неприемлемые ошибки, которые приходится искусственно корректировать при обработке вследствие отсутствия альтернативных данных. Применявшаяся при исследовании корректировка исключением ошибочных записей с последующей интерполяцией может влиять на получаемое разложение, а также никак не исключает скрытые ошибки, поэтому в подобного рода исследованиях необходимо использовать как можно более аккуратные и подробные данные.

Список литературы / References

[1]. Hethcote H. The Mathematics of Infectious Diseases / SIAM Review, vol. 42(4), 2000, pp. 599–653.
[2]. Beckley R., C. Weatherspoon C., Alexander M. et al. Modeling epidemics with differential equations / Tennessee State University Internal Report.
[3]. Santos G., Alves T., Alves G., Filho A. Asynchronous SIR model on Two-Dimensional Quasiperiodic Lattices, vol 01, 2019.
[4]. Bisin A., Moro A. Learning Epidemiology by Doing: The Empirical Implications of a Spatial-SIR Model with Behavioral Responses / Journal of Urban Economics, vol. 127, 2022.

[5]. von Csefalvay C. Spatial dynamics of epidemics: Epidemics in discrete and continuous space / Computational Modeling of Infectious Disease, vol. 13, 2023, pp. 257–303.
[6]. Derevich I., Panova A. (2019). Effects of Random Migration on the Growth of the Population of a Biological System. Journal of Engineering Physics and Thermophysics. 92. 10.1007/s10891-019-02054-x.
[7]. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html.
[8]. Elistratov S.A. POD-based Hydrodynamical Structures Visualization in Flows with an Internal Wave Attractor / Scientific Visualization, vol. 15(2), 2023, pp. 125 – 133.
[9]. Yilmaz I., Aylı E., Aradag S. Investigation of the Effects of Length to Depth Ratio on Open Supersonic Cavities Using CFD and Proper Orthogonal Decomposition / The Scientific World Journal, vol. 01, 2013, — p. 810175.
[10]. Berkooz G, Holmes P, Lumley J L. The Proper Orthogonal Decomposition in the Analysis of Turbulent Flows / Annual Review of Fluid Mechanics, vol. 25(1), 1993, pp. 539–575, DOI: 10.1146/annurev.fl.25.010193.002543.
[11]. Brunton, Steven L. and Kutz, J. Nathan. 7 Data-driven methods for reduced-order modeling / Snapshot-Based Methods and Algorithms: Volume 2, edited by Peter Benner, Stefano Grivet-Talocia, Alfio Quarteroni, Gianluigi Rozza, Wil Schilders and Luís Miguel Silveira, Berlin, Boston: De Gruyter, 2021, pp. 307-344, DOI: 10.1515/9783110671490-007.
[12]. Nakamura T., Fukami K., Fukagata K. Identifying key differences between linear stochastic estimation and neural networks for fluid flow regressions / SciRep, vol. 12(3), 2020, p. 3726.
[13]. Duan Fan, Wang Jinjun. Fluid–structure–sound interaction in noise reduction of a circular cylinder with flexible splitter plate / Journal of Fluid Mechanics, vol. 920(8), 2021.
[14]. Huang N.E., Shen Z., Long S.R. et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis / Proceedings of the Royal Society of London, vol. A4, 1998.

Информация об авторах / Information about authors

Степан Алексеевич ЕЛИСТРАТОВ – сотрудник Лаборатории цифрового моделирования технических систем Института системного программирования с 2021 года. Сфера научных интересов: математическое моделирование, прогностические модели, методы пониженной размерности.

Stepan Alekseevich ELISTRATOV – employee of Technical Systems Digital Modelling Laboratory of the Institute for System Programming of the RAS since 2021. Research interests: numerical simulation, forecast models, reduced order methods.