Прогнозирование распространения COVID-19 в ЕС с использованием рандомизированного машинного обучения динамических моделей*

А. Ю. Попков, Ю. А. Дубнов, Ю. С. Попков

Федеральный исследовательский центр «Информатика и управление» РАН, г. Москва, Россия

Аннотация. Работа посвящена применению метода рандомизированного машинного обучения для прогнозирования развития эпидемии COVID-19, основанной на эпидемиологической модели SIR. Предлагается два варианта моделирования, первый основан на использовании модели SIR с оценкой параметров по реальным оперативным данным о случаях заболевания, второй основан на идее моделирования индикатора распространения инфекции и его прогнозирования. Сравнительное исследование предлагаемых методов и подходов базируется на сравнении со стандартным подходом, основанным на методе наименьших квадратов и проводится на наборе данных нескольких стран Европейского союза. Показана работоспособность предлагаемого подхода и его эффективность и адекватность в условиях малого количества данных с высоким уровнем неопределенности.

Ключевые слова: моделирование эпидемий; SARS-CoV-2; COVID-19; SIR; рандомизированное машинное обучение; энтропия; энтропийное оценивание; прогнозирование; рандомизированное прогнозирование.

DOI 10.14357/20718632220307

Введение

Эпидемия новой коронавирусной инфекции началась в ноябре-декабре 2019 года в Китае, и уже к февралю 2020 года Всемирная организация здравоохранения присвоила ей статус пандемии, тем самым подтвердив ее глобальное распространение. Быстрый рост числа инфицированных и высокая смертность от вызываемой вирусом SARS-CoV-2 болезни COVID-19 привели к перегрузке систем здравоохранения почти во всех странах, включая страны Европейского союза и США. К марту 2020 года правительствам всех этих стран, в том числе и России, пришлось прибегать к крайним противоэпидемическим мерам, включая тотальный локдаун, запрет на перемещения людей, закрытие границ между регионами и государствами.

К настоящему времени пандемия, очевидно, еще не завершена, несмотря на проведение в течение 2021 года массовой вакцинации во многих странах мира. Также необходимо отметить, что свойства вируса SARS-CoV-2 не изучены в полной мере, несмотря на усилия многих развитых стран, в течении 2021 года наблюдалось несколько новых вспышек заболеваемости вследствие появления новых вариантов вируса, более устойчивых к приобретенному или вакцинному иммунитету.

* Работа выполнена при поддержке Российского фонда фундаментальных исследований (проект № 20-04-60119).

В этой связи, актуальность исследований, направленных на решения эпидемиологических задач в контексте новой болезни не подлежит сомнению.

С самого начала пандемии начали проводиться научные исследования, направленные на прогнозирование развития эпидемии, основными целями которых являлась оценка масштабов эпидемии в кратко- и среднесрочной перспективе. Понимание эпидемических процессов и их прогнозирование необходимо для оценки готовности системы здравоохранения, принимаемых мер сдерживания эпидемии, состояния экономики.

Основными подходами к моделированию эпидемического процесса на сегодняшний день являются подходы, использующие динамические эпидемиологические модели [1]. Главная идея, на которой построены эти модели, состоит в разделении популяции на непересекающиеся группы (компартменты), динамику которых описывают дифференциальные уравнения с некоторым количеством параметров. К настоящему времени было разработано большое количество вариантов таких моделей с большим количеством компартментов, а также другими модификациями [2-7].

Однако следует отметить, что структурное усложнение модели не всегда приводит к ее эффективному применению, что следует из малого количества оперативных реальных данных, доступных для ее обучения. Кроме того, качество оперативных данных хоть и существенно возросло по сравнению с началом 2020 года, тем не менее, остается недостаточно высоким в статистическом смысле вследствие политики сбора первичных данных (случаев заболеваемости), качеством тест-систем и массой других факторов. Эти факторы также ограничивают применение современных методов анализа данных, основанных на нейросетевых методах.

Одними из основных задач, стоящими перед эпидемиологами и правительствами стран в контексте борьбы с эпидемией является предсказание уровня заболеваемости и смертности в будущем. Опыт последних двух лет показал, что строить долгосрочные прогнозы эффективно не получается, однако, даже грубые краткосрочные прогнозы позволяют выработать если не меры по борьбе с патогеном, то хотя бы прогнозировать меры реакции на будущий рост заболеваемости.

В работе предлагается два подхода к решению этой задачи. Первый подход состоит в прогнозировании с использованием динамической модели SIR уровня заболеваемости путем оценивания параметра модели по реальным данным. Второй подход состоит в построении модели параметра этой модели, являющегося одним из основных индикаторов эпидемии, и прогнозировании эволюции этого индикатора. Важным свойством этого подхода является то, что построение и оценка этой модели происходит с использованием реальных данных о наблюдениях за другим, связанным, процессом.

Оценивание параметров моделей производится с использованием теории *Рандомизированного машинного обучения* (РМО) [8]. Основным достоинством этого метода является независимость от реальных характеристик используемых данных. Для корректного применения метода не требуются подтверждения или предположения о нормальности данных (или иных их вероятностных свойствах), а полученные в результате обучения распределения получены в условиях максимальной энтропии (максимальной неопределенности), таким образом, отражая наиболее "плохой" сценарий развития исследуемого процесса. Эти свойства энтропийного подхода установлены в работах Больцмана [9], Джейнса [10-11], Шеннона [12]. Еще одной важной особенностью метода является получение, вместе с оптимальными распределениями параметров, энтропийно-оптимальных распределений шумов (стохастических компонент данных), содержащихся в данных. Это свойство существенно отличает метод от классических подходов, в которых делаются различные предположения о характеристиках шумов.

1. Материалы и методы

1.1. Данные

В вычислительных экспериментах используются оперативные данные по нескольким странам Европейского союза, собираемые сервисом Data Hub [13, 14]. Самыми надежными оперативными данными об эпидемии являются дневные данные о регистрируемых случа-

ях и смертях. Далее будем использовать следующие обозначения для наборов данных, полученных из первичных оперативных данных:

• Дневные абсолютные
  - Confirmed (Cd) — Количество инфицированных (случаев)
  - Deaths (Dd) — Количество умерших
  - Recovered (Rd) — Количество выздоровевших
• Дневные 7-дневные средние Cd_avg, Dd_avg, Rd_avg
• Накопленные (общие, куммулятивные) C, D, R
• Дневные относительные Cd/N, Dd/N, Rd/N и общие относительные C/N, D/N, R/N, где N — численность населения в изучаемой стране.

На Рис. 1 представлены используемые оперативные данные по Германии с начала эпидемии COVID-19. На рисунке слева представлены дневные данные с 7-дневным скользящим средним, справа — общие (куммулятивные) данные.

Модель SIR показывает динамику соответствующих частей популяции, следовательно для обучения модели необходимо использовать общие (накопленные) данные на душу населения (относительные). Использование относительных данных позволяет работать в масштабе [0,1], а также сравнивать динамику в разных регионах (странах) из-за различий в численности населения.

Общие данные по дням вычисляются по 7-дневному среднему из дневных данных и выравниваются таким образом, чтобы начало данных приходилось на понедельник (первый день недели).

Необходимо отметить, что в настоящее время эпидемиологические данные и модели как правило оперируют именно недельными показателями вследствие специфики сбора данных в течении недели: выписка пациентов происходит во многих странах в воскресенье или понедельник, тестирование более активное в конце недели, в некоторых странах пятница и суббота — выходные дни и т.д.

В обучении используются 4 точки данных (5 для обучения МНК), соответствующие началу недели, начиная с 230 дня от начала эпидемии (33 неделя).

1.2. Модель SIR

Модель SIR (Susceptible-Infected-Removed) является «золотым стандартом» современной эпидемиологии [1]. Она реализуется системой нелинейных дифференциальных уравнений, описывающих динамику соответствующих частей популяции

$$\frac{dS}{dt} = -\beta SI, \quad \frac{dI}{dt} = \beta SI - \gamma I, \quad \frac{dR}{dt} = \gamma I, \quad S(0) = 1. \tag{1}$$

Модель основана на трех группах (компартментах): уязвимых (Susceptible), инфицированных (Infected), удаленных (Removed). В группе S находятся люди, не имеющие иммунитета к инфекции, в группу I попадают заболевшие (инфицированные), в группу R — умершие и выздоровевшие.

В модели есть два параметра: β, называемый transmission rate, характеризует скорость переда-

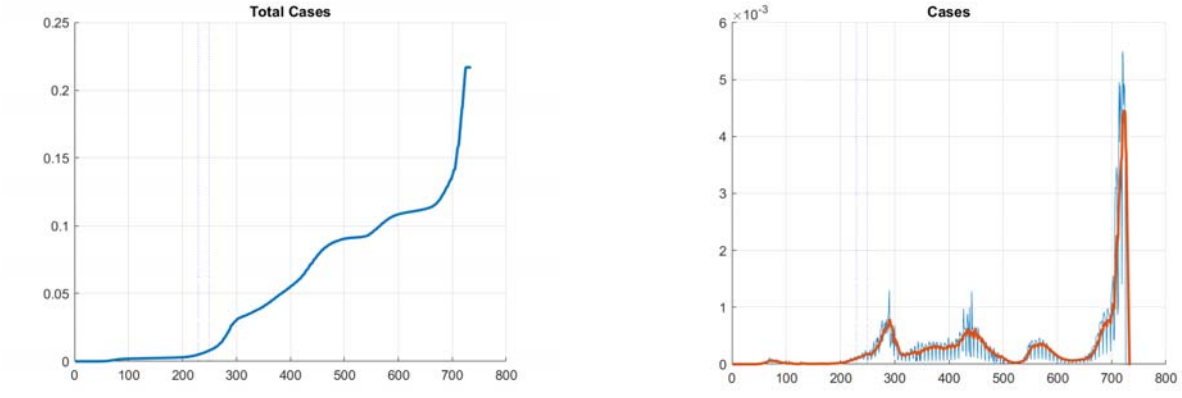


Рис. 1. Оперативные данные о случаях заболевания в Германии с начала эпидемии COVID-19

чи инфекции от одного члена популяции к другому и определяется как среднее количество контактов человека за единицу времени × вероятность передачи инфекции в результате контакта; в итоге он определяет среднее количество зараженных от одного инфицированного, и γ, называемый recovery rate, который характеризует перемещение из группы I в группу R.

На основе этих параметров определяются основные индикаторы эпидемии: $d = 1/\gamma$ — средний инфекционный период (mean infectious period), в течение которого человек может распространять инфекцию, и основное репродуктивное число (basic reproduction number) $R_0 = \beta/\gamma$. Последний показывает, сколько человек каждый инфицированный может заразить за период d.

Параметр γ является характеристикой самой инфекции, в то время как β зависит не только от свойств инфекции, но и от структуры общества, плотности населения, структуры рынка труда, потоков общественного транспорта и также большого количества иных факторов, которые составляют типичную жизнь современного городского жителя. Также этот показатель можно использовать, чтобы анализировать развитие эпидемии со временем, а также оценивать эффективность принимаемых для борьбы с ней мер.

В этой связи, далее будем рассматривать параметр γ фиксированным и равным 0.1, что соответствует инфекционному периоду продолжительностью 10 дней (этот период был установлен для первого, уханьского штамма).

1.3. Модель коэффициента распространения β

Для построения модели β произведем его оценку методом наименьших квадратов, используя модель (1) в скользящем окне шириной 4 недели. Результаты этой оценки для Германии представлены на Рис. 2 слева вместе с несколькими вариантами аппроксимации этих данных различными функциями. Из этого графика видно, что на длинном временном интервале β удерживалась на стабильном уровне и даже немного падала. По всей видимости, это явилось результатом мер борьбы с эпидемией, вводимых правительством. Однако, как видно из графика, на коротких временных интервалах (несколько недель/месяцев) наблюдалось повышение и снижение β по причине подъема заболеваемости вследствие отмены каких-то ограничений или появления другого штамма вируса.

В связи с этим задачу прогнозирования будем решать на коротком временном интервале, в частности, попытаемся применить предлагаемый здесь подход на периоде подъема заболеваемости. Один из таких периодов начался на 33 неделе, на Рис. 2 справа показано 5 точек, соответствующих неделям, и аппроксимация этих данных экспоненциальной функцией

$$\beta(x) = a\exp(bx) + c, \qquad (2)$$

где x имеет смысл времени, в рассматриваемой задаче является индексом дня или недели.

Учитывая характер данных на исследуемом интервале и качество их аппроксимации выбранной функцией, будем использовать (2) в

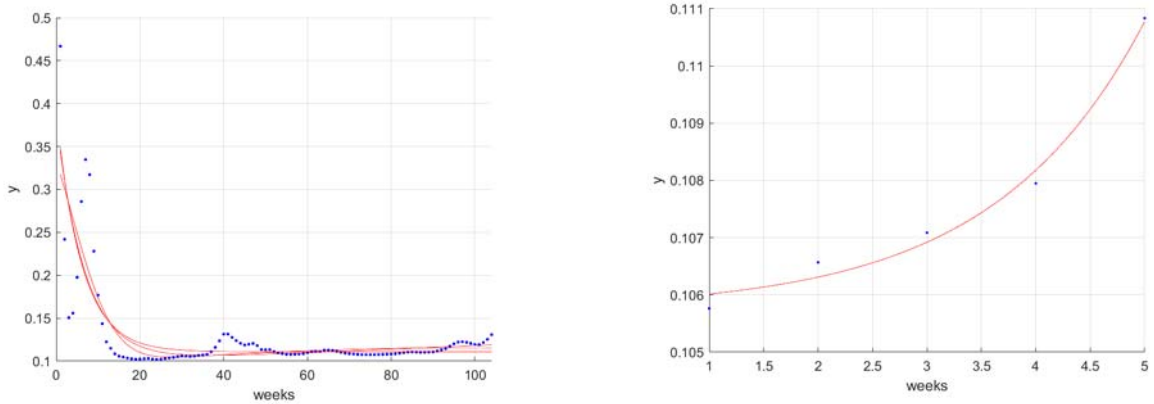


Рис. 2. Оценка β и аппроксимация по данным Германии

качестве модели β, которую требуется оценить предлагаемым в работе методом.

1.4. Прогнозирование общего количества инфицированных

Основной задачей, на решение которой ориентировано моделирование процесса распространения эпидемии с помощью описанной динамической модели, состоит в построении прогноза общего количества инфицированных в будущем. В этом контексте будем рассматривать два подхода.

Первый состоит в оценивании параметра модели β с последующим использованием этой оценки для реализации модели на интервале прогнозирования. Соответствующую модель далее будем называть M_1.

Второй подход состоит в моделировании процесса эволюции β как индикатора состояния эпидемии. Здесь необходимо построить его модель и оценить ее параметры по наблюдениям за основным процессов. Прогнозирование с использованием такой модели (далее будем называть ее M_2) состоит в прогнозировании параметра β с помощью соответствующей модели, после чего этот прогнозный показатель используется в качестве параметра основной модели для построения прогноза.

Во всех вариантах соответствующая модель оценивается (обучается) двумя методами: методом наименьших квадратов и методом рандомизированного машинного обучения, после чего строится прогноз соответствующим методом, в последнем случае — методом рандомизированного прогнозирования, состоящим в сэмплировании оценок распределений параметров и вычислении средней по ансамблю траектории.

1.5. Рандомизированное машинное обучение динамической модели SIR

От модели в виде (1) перейдем к виду «вход-выход» с помощью схемы Эйлера с шагом h, получим следующую систему разностных уравнений

$$\begin{aligned} S[k] &= S[k-1] - h\beta S[k-1]I[k-1], \\ I[k] &= I[k-1] + \\ &\quad + h(\beta S[k-1]I[k-1] - \gamma I[k-1]), \\ R[k] &= R[k-1] + h\gamma I[k-1], \end{aligned} \tag{3}$$

где k — индекс узла регулярной сетки с шагом h.

Введем обозначения

$$F(\mathbf{x},\mathbf{a}) = y[n] + h\Phi(t[n], y[n], \mathbf{a}),$$
$$\mathbf{x} = (t[n], x[n], y[n]), \quad \mathbf{y} = (S, I, R),$$

где Ф определяет выражения из правой части (3), тогда (3) может быть представлено в виде

$$\mathbf{y} = F(\mathbf{x},\mathbf{a}). \tag{4}$$

Следуя теории рандомизированного машинного обучения [8], задача оценивания параметров этой модели состоит в оценивании распределений вероятностей параметров $\mathbf{a}$ вместе с шумами измерений выхода $\boldsymbol{\xi}$, и реализуется решением оптимизационной задачи

$$H(P,Q) = -\left[\int_{\mathcal{A}} P(\mathbf{a})\ln P(\mathbf{a})d\mathbf{a} + \int_{\Xi} Q(\boldsymbol{\xi})\ln Q(\boldsymbol{\xi})d\boldsymbol{\xi}\right] \to \max_{P,Q}, \tag{5}$$

$$\int_{\mathcal{A}} P(\mathbf{a})d\mathbf{a} = 1, \quad \int_{\Xi_j} q_j(\xi)d\xi = 1, \tag{6}$$

$$\int_{\mathcal{A}} F(\mathbf{x}_j,\mathbf{a})P(\mathbf{a})d\mathbf{a} + \int_{\Xi} \xi q_j(\xi)d\xi = \hat{y}_j,\ j = \overline{1,m}, \tag{7}$$

где m — количество точек данных, в которых производится оценивание, а распределения параметров и шумов являются интервальными, причем, предполагая независимость измерений выхода, распределения шумов имеют вид

$$\mathbf{a} \sim P(\mathbf{a}), \qquad \mathbf{a} \in \mathcal{A} \in R^d,$$
$$\boldsymbol{\xi} \sim Q(\boldsymbol{\xi}) = \prod_{j=1}^{m} q_j(\xi), \qquad \boldsymbol{\xi} \in \Xi \in R^m, \tag{8}$$

где d — размерность пространства параметров, m — количество точек оценивания и, соответственно, размерность пространства шумов. Здесь и далее полужирным шрифтом будем обозначать векторные величины из соответствующих пространств.

В этих выражениях векторный параметр модели может быть реализован разными способами. Рассмотрим два варианта, соответствующие моделям M_1 и M_2. В варианте M_1 $\mathbf{a} = \beta$, в варианте M_2 предусматривается построение модели для параметра β, реализуемой функцией (2), тогда $\mathbf{a} = (a, b, c)$

Решением задачи (5)-(7) являются энтропийно-оптимальные распределения, вид которых не будем приводить здесь, способ решения этой задачи и вид распределений подробно описан в [8].

2. Результаты

Реализация всех вычислительных экспериментов была проведена на платформе

MATLAB 9.7 (2019b) с использованием пакетов Optimization и Curve Fitting соответствующих версий.

Для всех точек на интервале обучения используется шум в пределах 30% во всех точках ($\Xi_j = [-0.3,0.3]$), а интервалы рандомизированных параметров (8) устанавливаются в пределах 50% от оптимальных значений параметров соответствующих моделей, полученных при помощи МНК (далее будем использовать обозначение β_{ols} для этого значения). Интервалы представлены в Табл. 1, 2.

На представленных ниже рисунках используются следующие обозначения: реальные данные (наблюдения) с меткой real, траектории, полученные при обучении МНК с меткой ols, средняя рандомизированная траектория с меткой avg, также представлен ансамбль траекторий и область стандартного отклонения, вычисленная по нему.

Результаты обучения модели M_1 для Германии, Франции, Швеции и Финляндии приведены на Рис. 3. Результаты обучения модели M_2 для Германии, Франции, Швеции и Финляндии приведены на Рис. 4. Реальные данные, представляющие собой оценку β в скользящем окне, указаны с меткой est, траектории, полученные при обучении МНК, с меткой ols, средняя рандомизированная траектория с меткой avg, с меткой mean указана траектория, полученная от реализации модели со средним по распределению значением параметров. Вертикальные линии отмечают интервал обучения и интервал прогнозирования.

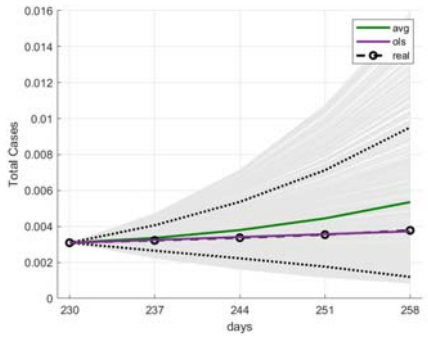
Результаты прогнозирования для Германии с использованием моделей M_1 и M_2 представлены на Рис. 5. На рисунках используются следующие обозначения: метка ols соответствует МНК, mee — модели M_1, β_{ols} — модель M_2, обученная МНК, β_{mee} — модель M_2, обученная предлагаемым в работе методом, C — реальные данные.

Табл. 1. Интервалы параметров для модели M_1.

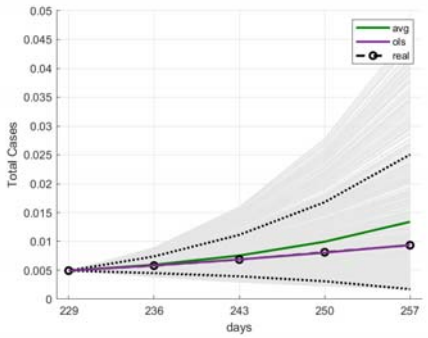
	β_{ols}	β^-	β^+
Германия	0.1070	0.0538	0.1613
Франция	0.1251	0.0625	0.1874
Италия	0.1336	0.0679	0.2038
Великобритания	0.1179	0.0600	0.1800
Испания	0.1198	0.0598	0.1793
Швеция	0.1073	0.0540	0.1620
Финляндия	0.1072	0.0542	0.1625

Табл. 2. Интервалы параметров для модели M_2.

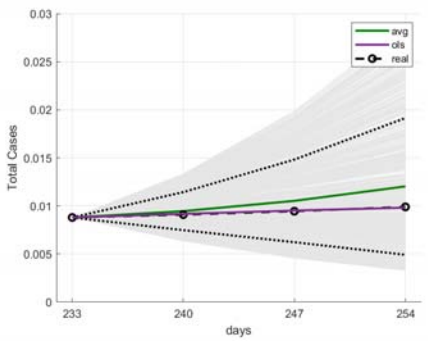
	A_1	A_2	A_3
Германия	[0.0001,0.0002]	[0.3628,1.0885]	[0.0528,0.1586]
Франция	[0.0594,0.1780]	$[-0.0003, -0.0010]$	[0.0032,0.0095]
Италия	[1.2686,3.8057]	[0.0013,0.0039]	$[-1.2115, -3.6344]$
Великобритания	[0.4031,1.2094]	[0.0030,0.0090]	$[-0.3505, -1.0514]$
Испания	[0.2965,0.8896]	$[-0.0008, -0.0025]$	$[-0.2364, -0.7091]$
Швеция	[0.0506,0.1517]	[0.0055,0.0165]	[0.0010,0.0031]
Финляндия	[0.0018,0.0055]	[0.1460,0.4379]	[0.0498,0.1494]



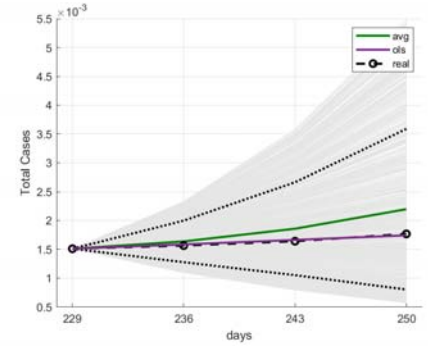
а) Германия



б) Франция

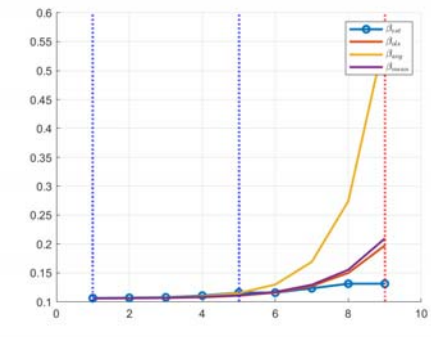


в) Швеция

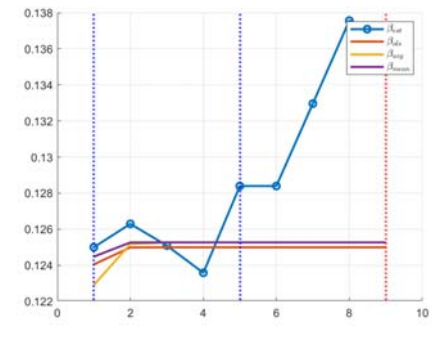


г) Финляндия

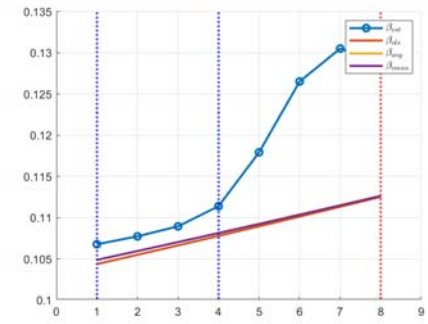
Рис. 3. Результаты обучения модели M_1



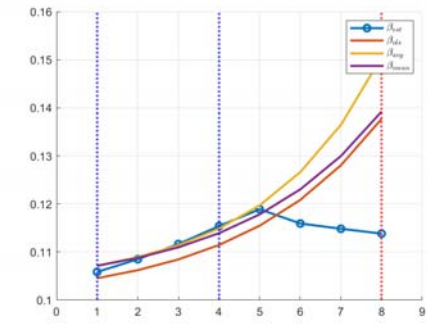
а) Германия



б) Франция



в) Швеция



г) Финляндия

Рис. 4. Результаты обучения модели M_2

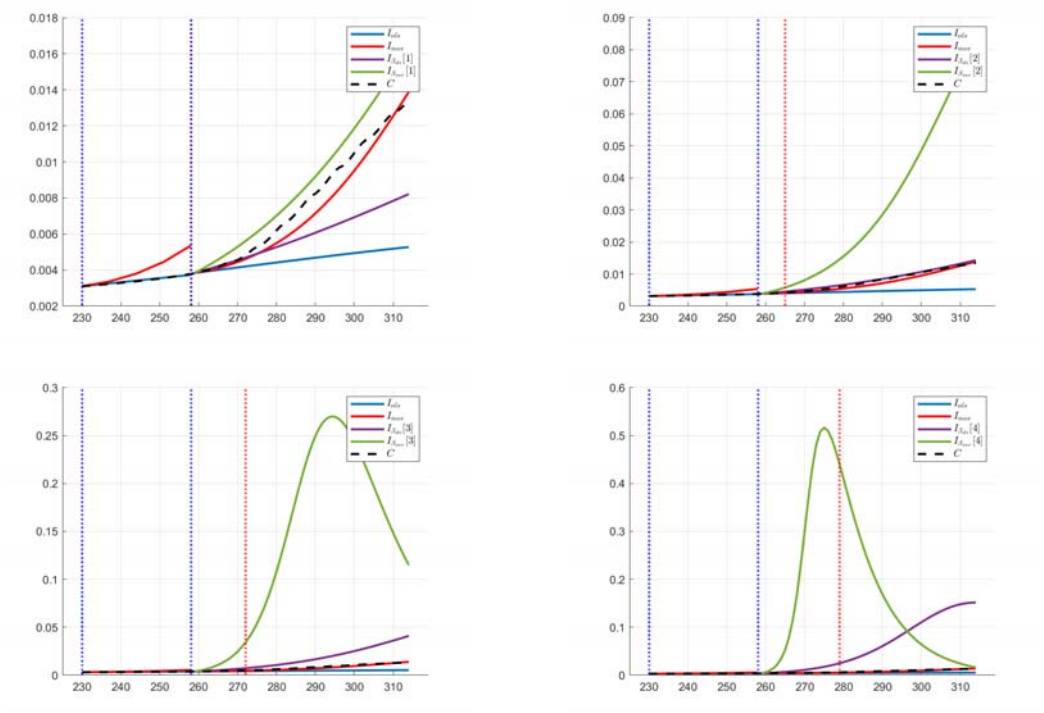


Рис. 5. Результаты прогнозирования для Германии

Табл. 3. Результаты прогнозирования модели M_1.

	β_{ols}	β_{mean}	$\delta(I_{ols}, C)$	$\delta(I_{mee}, C)$
Германия	0,1070	0,1091	0,29	**0,03**
Франция	0,1251	0,1247	0,15	0,18
Италия	0,1336	0,1342	0,09	0,40
Великобритания	0,1179	0,1178	0,20	**0,15**
Испания	0,1198	0,1194	0,06	0,23
Швеция	0,1073	0,1086	0,29	**0,06**
Финляндия	0,1072	0,1083	0,20	0,29

Для модели M_2 было рассмотрено 4 варианта прогноза, в котором β прогнозировалось на 4 недели вперед с интервалов 1 неделя. На рисунках значения, соответствующие этим прогнозным величинам, указаны в квадратных скобках.

Результаты прогнозирования модели M_1 представлены в Табл. 3. С меткой ols указаны значения параметра β и ошибок при прогнозировании, полученные МНК, с меткой mee — значения соответствующих величин, полученных методом РМО, в колонке β_{mean} указано среднее по распределению значение.

Результаты прогнозирования модели M_2 представлены в Табл. 4. В строке с индексом n указан индекс точки (соответствующей началу недели) прогнозирования, в первой группе колонок указаны прогнозные значения β для каждой точки, полученные с помощью модели, во второй группе — ошибки между прогнозной траекторией I модели (1) и реальными данными C. Для каждой страны указаны две группы показателей, в первой строке показатели, полученные методом МНК, во второй — показатели, полученные предлагаемым в работе методом РМО.

Табл. 4. Результаты прогнозирования модели M_2.

	$\beta[n]$				$\delta(I_{\beta[n]}, C)$			
n	1	2	3	4	1	2	3	4
Германия	0,1108	0,1162	0,1273	0,1502	0,39	0,20	0,13	0,49
	0,1152	0,1301	0,1693	0,2744	**0,23**	**0,19**	0,64	0,85
Франция	0,1246	0,1246	0,1245	0,1244	0,21	0,21	0,22	0,22
	0,1231	0,1230	0,1230	0,1229	0,26	0,26	0,26	0,27
Италия	0,1470	0,1537	0,1603	0,1669	0,26	0,34	0,41	0,47
	0,2112	0,2180	0,2249	0,2318	0,66	0,68	0,69	0,70
Великобритания	0,1299	0,1348	0,1399	0,1449	0,06	0,16	0,24	0,32
	0,1897	0,1949	0,2000	0,2052	0,65	0,67	0,69	0,70
Испания	0,1153	0,1143	0,1133	0,1123	0,16	0,19	0,22	0,25
	0,1346	0,1336	0,1325	0,1315	0,25	0,23	**0,22**	**0,20**
Швеция	0,1090	0,1101	0,1113	0,1125	0,47	0,42	0,38	0,33
	0,1093	0,1104	0,1115	0,1127	**0,46**	**0,41**	**0,37**	**0,33**
Финляндия	0,1155	0,1208	0,1281	0,1377	0,01	0,13	0,27	0,41
	0,1197	0,1266	0,1365	0,1505	0,10	0,24	0,40	0,54

В таблицах отмечены ошибки, значение которых меньше, чем при использовании МНК, что является индикатором потенциальных преимуществ предложенного подхода в прогнозировании по сравнению с МНК.

3. Обсуждение

Полученные в рамках эксперимента результаты в целом показывают работоспособность предлагаемого подхода к решению задачи прогнозирования развития эпидемии COVID-19. Можно отметить несколько важных наблюдений.

Первое наблюдение связано с тем, что прогноз, полученный классическим методом наименьших квадратов почти всегда приводит к недооценке выхода модели по сравнению с реальными наблюдаемыми данными, в то время, как предлагаемый в работе подход, основанный на энтропийно-рандомизированной концепции, обладает большей гибкостью и потенциально позволяет получать более адекватные реальным данным, а главное, контексту рассматриваемой задачи, прогнозы.

Второе наблюдение связано с подходом к моделированию индикатора β по наблюдениям за общим количеством случаев заболевания. Ошибки при прогнозировании с использованием данного подхода получаются существенно выше, чем при стандартном подходе, связанном с оцениванием параметра модели SIR, причем это касается как МНК, так и метода РМО. При этом, как видно из соответствующих графиков, выбор модели для β существенным образом сказывается на результатах моделирования. По всей видимости это обусловлено нелинейностью и существенной чувствительностью модели SIR от входных данных. Для некоторых стран, вероятно, требуется выбирать другую модель β и применять ее на другом периоде обучения. Продемонстрированная в работе методика с выбранной моделью β в экспоненциальном виде неплохо работает на интервале начала роста β.

Необходимо также отметить, что рассматриваемая в работе задача в целом обладает определенной спецификой, связанной как с существенной неопределенностью в данных о случаях

заболевания, так и с мерами разной эффективности, вводимых правительством с целью борьбы с эпидемией. Это приводит, в частности, к существенным расхождениям любых прогнозов с реальными данными. В контексте прогноза заболеваемости представляется более адекватным переоценивать, чем недооценивать общие показатели заболеваемости, т.к. недооценка приведет к недостаточной реакции на развитие эпидемии, чем это может потребоваться.

В целом можно отметить, что поставленный эксперимент подтверждает работоспособность и эффективность энтропийно-рандомизированного подхода к решению задач, связанных с прогнозированием распространения эпидемии, однако, оставляет большое пространство для дальнейших исследований, которые могут быть направлены на изучение других моделей, более подходящих для иных стадий развития эпидемии, иных подходов к формированию интервалов рандомизированных параметров, а также использования более сложных динамических эпидемиологических моделей, например, SEIR.

4. Заключение

В работе предложено применение метода рандомизированного машинного обучения для прогнозирования развития эпидемии COVID 19, моделирование которой реализовано с помощью динамической эпидемиологической модели SIR. Рассмотрены две постановки задачи прогнозирования, первая состоит в оценивании параметра модели SIR, вторая — в моделировании и оценивании коэффициента распространения инфекции с использованием реальных данных о ежедневных случаях заболевания. Вторая задача обладает определенной новизной в силу того, что проблема прогнозирования коэффициента распространения и связанного с ним главного репродуктивного числа R_0 к настоящему времени не имеет стандартного решения. В этом контексте в работе предложен подход, работоспособность и потенциальная эффективность которого продемонстрирована в вычислительном эксперименте на реальных данных для нескольких стран Европейского союза. Следует отметить, что в эксперименте используется малое количество данных для обучения соответствующих моделей, при этом полученные результаты обладают достаточным уровнем качества и адекватности в контексте решаемой задачи.

Литература

1. van den Driessche P. Mathematical Epidemiology / ed. by Brauer Fred, van den Driessche Pauline, and Wu Jianhong. — Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. — Vol. 1945 of Lecture Notes in Mathematics. — P. 147–157. — Access mode: https://doi.org/10.1007/978-3-540-78911-6.
2. Khan Irtesam Mahmud, Haque Ubydul, Kaisar Samiha, Rahman Mohammad Sohel. A Computational Modeling Study of COVID-19 in Bangladesh // The American Journal of Tropical Medicine and Hygiene. — 2021. — jan. — Vol. 104, no. 1. — P. 66–74.
3. Lavielle Marc, Faron Matthieu, Lefevre J´er´emie H., and Zeitoun Jean-David. Predicting the propagation of COVID-19 at an international scale: extension of an SIR model // BMJ Open. — 2021. — may. —Vol. 11, no. 5. — P. e041472.
4. Lawson Andrew B., Kim Joanne. Space-time covid-19 Bayesian SIR modeling in South Carolina // PLOS ONE. — 2021. — mar. — Vol. 16, no. 3. — P. e0242777.
5. Purkayastha Soumik, Bhattacharyya Rupam, Bhaduri Ritwik, Kundu Ritoban, Gu Xuelin, Salvatore Maxwell, Ray Debashree, Mishra Swapnil, Mukherjee Bhramar. A comparison of five epidemiological models for transmission of SARS-CoV-2 in India // BMC Infectious Diseases. — 2021. — jun. — Vol. 21, no. 1.
6. de Andres P. L., de Andres-Bragado L., and Hoessly L. Monitoring and Forecasting COVID-19: Heuristic Regression, Susceptible-Infected-Removed Model and, Spatial Stochastic // Frontiers in Applied Mathematics and Statistics. — 2021. — may. — Vol. 7.
7. Deo Vishal and Grover Vishal. A new extension of state-space SIR model to account for Underreporting – An application to the COVID-19 transmission in California and Florida // Results in Physics. — 2021. —may. — Vol. 24. — P. 104182.
8. Попков Ю. С., Попков А. Ю. и Дубнов Ю. А. Рандомизированное машинное обучение при ограниченных наборах данных: от эмпирической вероятности к энтропийной рандомизации. — Москва: ЛЕНАНД, 2019. — ISBN: 978-5-9710-5908-0.
9. Больцман Л. О связи между вторым началом механической теории теплоты и теорией вероятностей в теоремах о тепловом равновесии // Больцман Л.Э. Избранные труды. / под ред. Шлак Л. С. —М. : Наука, 1984. — Классики науки.
10. Jaynes Edwin T. Information theory and statistical mechanics // Physical review. — 1957. — Vol. 106, no. 4. — P. 620–630.
11. Jaynes Edwin T. Probability theory: the logic of science. — Cambridge university press, 2003.
12. Shannon Claude E. Communication theory of secrecy systems // Bell Labs Technical Journal. — 1949. — Vol. 28, no. 4. — P. 656–715.

13. Guidotti Emanuele and Ardia David. COVID-19 Data Hub // Journal of Open Source Software. — 2020. — Vol. 5, no. 51. — P. 2376. — Access mode: https://doi.org/10.21105/joss.02376.
14. COVID-19 Data Hub. — https://www.covid19datahub.io. — 2021. — Accessed: 2021-12-20.

Попков Алексей Юрьевич. Федеральный исследовательский центр «Информатика и управление» Российской академии наук, Москва, Россия, ведущий научный сотрудник. Кандидат технических наук. Количество печатных работ: 47. Область научных интересов: энтропийные методы, рандомизированное машинное обучение, интеллектуальный анализ данных, разработка программного обеспечения. E-mail: apopkov@isa.ru.

Дубнов Юрий Андреевич. Федеральный исследовательский центр «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН), г. Москва, Россия, Научный сотрудник. Национальный исследовательский университет Высшая школа экономики (НИУ ВШЭ), г. Москва, Россия, Старший преподаватель. Количество печатных работ: 29. Область научных интересов: байесовское оценивание, машинное обучение, принцип максимума энтропии. E-mail: yury.dubnov@phystech.edu

Попков Юрий Соломонович. Федеральный исследовательский центр «Информатика и управление» Российской академии наук, Москва, Россия, главный научный сотрудник; Академик РАН, доктор технических наук, профессор; Институт проблем управления Российской академии наук, Москва, Россия, главный научный сотрудник. Количество печатных работ: 220. Область научных интересов: энтропийные методы, макросистемы, рандомизированное машинное обучение. E-mail: popkov@isa.ru

Forecasting of COVID-19 Dynamics in EU Using Randomized Machine Learning Applied to Dynamic Models

A.Y. Popkov, Y.A. Dubnov, Y.S. Popkov

Federal Research Center «Computer Science and Control» of Russian Academy of Sciences, Moscow, Russia

Abstract. The work is devoted to application of the theory of Randomized Machine Learning to forecasting of the COVID-19 pandemic based on SIR epidemiological model. We propose two modelling variants, the first is based on estimation of SIR model using real case data, the second is based on the idea of modelling transmission coefficient and its prediction. Comparative study of proposed approach is based on a comparison with the standard least squares approach and is carried out on a dataset of several countries of the European Union. It is shown the performance of the proposed approach and its effectiveness and adequacy under conditions of small amount of data with a high level of uncertainty.

Keywords: epidemic modelling; SARS-CoV-2; COVID-19; SIR; randomized machine learning; entropy; entropy estimation; forecasting; randomized forecasting.

DOI 10.14357/20718632220307

References

1. van den Driessche P. Mathematical Epidemiology / ed. by Brauer Fred, van den Driessche Pauline, and Wu Jianhong. — Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. — Vol. 1945 of Lecture Notes in Mathematics. — P. 147–157. — Access mode: https://doi.org/10.1007/978-3-540-78911-6.
2. Khan Irtesam Mahmud, Haque Ubydul, Kaisar Samiha, Rahman Mohammad Sohel. A Computational Modeling Study of COVID-19 in Bangladesh // The American Journal of Tropical Medicine and Hygiene. — 2021. — jan. — Vol. 104, no. 1. — P. 66–74.
3. Lavielle Marc, Faron Matthieu, Lefevre J´er´emie H., and Zeitoun Jean-David. Predicting the propagation of COVID-19 at an international scale: extension of an SIR model // BMJ Open. — 2021. — may. —Vol. 11, no. 5. — P. e041472.
4. Lawson Andrew B., Kim Joanne. Space-time covid-19 Bayesian SIR modeling in South Carolina // PLOS ONE. — 2021. — mar. — Vol. 16, no. 3. — P. e0242777.
5. Purkayastha Soumik, Bhattacharyya Rupam, Bhaduri Ritwik, Kundu Ritoban, Gu Xuelin, Salvatore Maxwell,

Ray Debashree, Mishra Swapnil, Mukherjee Bhramar. A comparison of five epidemiological models for transmission of SARS-CoV-2 in India // BMC Infectious Diseases. — 2021. — jun. — Vol. 21, no. 1.
6. de Andres P. L., de Andres-Bragado L., and Hoessly L. Monitoring and Forecasting COVID-19: Heuristic Regression, Susceptible-Infected-Removed Model and, Spatial Stochastic // Frontiers in Applied Mathematics and Statistics. — 2021. — may. — Vol. 7.
7. Deo Vishal and Grover Vishal. A new extension of state-space SIR model to account for Underreporting – An application to the COVID-19 transmission in California and Florida // Results in Physics. — 2021. —may. — Vol. 24. — P. 104182.
8. Popkov Y.S., Popkov A.Y., Dubnov Y.A. Randomizirovannoe mashinnoe obuchenie: ot empiricheskoi veroyatnosti k entropiinoy randomizacii. — Moscow: LENAND, 2019. — ISBN: 978-5-9710-5908-0.
9. Boltzmann L. On connection between the second law of mechanical theory of heat and probability theory in heat equilibrium theorems / In: Boltzmann L.E. Selected proceedings, Moscow: Nauka, 1984.
10. Jaynes Edwin T. Information theory and statistical mechanics // Physical review. — 1957. — Vol. 106, no. 4. — P. 620–630.
11. Jaynes Edwin T. Probability theory: the logic of science. — Cambridge university press, 2003.
12. Shannon Claude E. Communication theory of secrecy systems // Bell Labs Technical Journal. — 1949. — Vol. 28, no. 4. — P. 656–715.
13. Guidotti Emanuele and Ardia David. COVID-19 Data Hub // Journal of Open Source Software. — 2020. — Vol. 5, no. 51. — P. 2376. — Access mode: https://doi.org/10.21105/joss.02376.
14. COVID-19 Data Hub. — https://www.covid19datahub.io. — 2021. — Accessed: 2021-12-20.

Popkov A. Yu.. Federal Reearch Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia, leading research scientist. PhD in Engineering. Number of publications: 47. Scientific area: entropy methods, randomized machine learning, data mining, software development. E-mail: apopkov@isa.ru

Dubnov Yu. A. Federal Reearch Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia, research scientist. National Research University Higher School of Economics, Moscow, Russia, chief lecturer. Number of publications: 29. Scientific area: bayesian estimation, machine learning, maximum entopy principle.
E-mail: yury.dubnov@phystech.edu

Popkov Yu. S. Federal Reearch Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia, chief research scientist.; Member of RAS, Doctor of Science in Engineering, professor; Institute of Control Sciences of Russian Academy of Sciences, Moscow, Russia, chief research scientist. Number of publications: 220. Scientific area: entropy methods, macrosystems, randomized machine learning. E-mail: popkov@isa.ru