https://doi.org/10.31649/1997-9266-2023-170-5-50-59

УДК 004.9+578.834.1

А. В. Лосенко[1]

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ПРОГНОЗУВАННЯ ЧАСОВОГО РЯДУ КІЛЬКОСТІ ХВОРИХ НА КОРОНАВІРУС НА ОСНОВІ МОДЕЛІ FACEBOOK PROPHET

[1]Вінницький національний технічний університет

Розглянуто результати розроблення інформаційної технології прогнозування часового ряду кількості хворих на коронавірус COVID-19. Здійснено огляд напрацювань за цією тематикою та обґрунтований вибір бібліотеки Facebook Prophet як основи для розробленої ІТ. Запропоновано підвищити точність прогнозування кількості нових хворих на коронавірус у короткостроковій перспективі за допомогою вибраної моделі.

Запропоновано математичне обґрунтування наявних паралельно-послідовних ітеративних методів ідентифікації параметрів моделі Facebook Prophet для врахування значної волатильності ряду значень нових хворих. Методи дозволяють визначати гіперпараметри тренду ряду, його сезонних складових та аномалій, що впливають на значення цього ряду. Формалізовано підхід щодо порівняльного аналізу впливу основних трендів захворюваності регіонів (або країн) без урахування впливу аномалій та сезонних складових шляхом створення картограм, які дозволяють аналізувати тенденції поширення хвороби в заданому регіоні.

Розроблена архітектура запропонованої інформаційної технології та описано її складові. Створено програмне забезпечення на Python на базі платформи Kaggle, яке реалізує цю технологію, порівняно результати його застосування з моделлю SEIR-U, розробленою науковцями НАН України за матеріалами звітів Робочої групи з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 при НАН України за даними 2020—2022 рр. Порівняння довело ефективність запропонованої інформаційної технології.

Ключові слова: інформаційна технологія, COVID-19, прогнозування часових рядів, Prophet, ряд Фур'є, Python, Kaggle.

Вступ

Пандемія коронавірусу COVID-19, викликаного інфекцією SARS-CoV-2, зробила як ніколи актуальною задачу аналізу, моделювання та прогнозування її поширення. Вже існують методи прогнозування часових рядів, що використовуються для розв'язання такої типової задачі, а також моделі машинного навчання, розроблені на їхній основі, проте, у зв'язку з великою кількістю чинників, що впливають на розповсюдження цієї хвороби та хаотичність їхніх значень, ці моделі потребують застосування нових підходів до ідентифікації оптимальних значень параметрів таких моделей. Також, не менш важливим завданням, є аналіз потенційного інфекційного впливу сусідніх регіонів або країн, тому що це безпосередньо впливає на об'єкт дослідження. Зазначені вище завдання та проблеми потребують глибшого аналізу вже наявних моделей прогнозування часових рядів, а також об'єднання таких моделей і методів, алгоритмів та програм їхньої ідентифікації в інформаційну технологію, що повинна застосовуватись для прогнозування і візуалізації кількості нових хворих на коронавірус в заданому регіоні чи країні, що, також дозволить ефективно оцінювати тенденції епідеміологічної ситуації та приймати завчасні рішення з мінімізації негативного впливу цього захворювання чи його поширення.

Упродовж 2020—2023 рр. опубліковано десятки тисяч статей (на момент написання цієї статті запит «COVID-19 Forecast» на веб-платформі Google Scholar видає 149 тисяч результатів), розглянемо, які моделі машинного навчання були використані у них найчастіше:

1. Модель SIR (та її варіації — SEIR та ін., що враховують здорових (S), інфікованих (I) осіб та тих, що одужали (R), а також (E) — хворих в інкубаційному періоді, коли вони ще не є заразними) використовує системи диференціальних рівнянь. Застосовувалась в Україні Робочою групою з мате-

матичного моделювання проблем, пов'язаних з пандемією коронавірусу SARS-CoV-2 в Україні, при НАН України [1]. Ця модель є дуже чутливою до якості вхідних даних та потребує велику їхню кількість для якісного прогнозування в довгостроковій перспективі. Також, робота цієї моделі ґрунтується лише на даних про захворюваність, та не враховує інші чинники, що можуть вплинути на розповсюдження (наприклад, аномальні дати, такі як святкові дні) [2];

2. Статистичні моделі часових рядів, передусім на базі ARIMA та Prophet [3]—[5], є одними з найпопулярніших моделей для короткострокового прогнозування часових рядів, оскільки вони дозволяють враховувати різні фактори впливу за умов зашумлення даних. Враховуючи їхню порівняно кращу ефективність, ніж у вищезгаданої моделі, вони однаково потребують детального аналізу факторів, що враховуються у прогнозуванні, а самі значення параметрів моделі вимагають їхнього оптимального визначення для отримання найточніших результатів. Крім цього, в процесі прогнозування нестаціонарних часових рядів модель ARIMA потребує додаткової оптимізації її параметрів, що набагато ускладнює роботу з цією моделлю [6]. До того ж, модель Prophet дає можливість гнучкого налаштування параметрів моделі, що зручно у прогнозуванні явища з багатьма хаотичними чинниками, такими як захворюваність на коронавірус [7].

3. Гібридні моделі на основі ARIMA або Prophet, які поєднуються з нейронним мережами для підвищення загальної точності. Наявний досвід [8], [9] показує, що такий підхід в деяких випадках може давати значно меншу похибку прогнозування, але такі моделі мають високий ризик оверфітингу (перенавчання), у разі рядів даних зі значною волатильністю.

Проведений вище аналіз вказує на актуальність вирішення завдання ефективної ідентифікації параметрів саме моделі Prophet, оскільки серед аналогічних моделей вона є найдоречнішою для розв'язання задачі прогнозування захворюваності на коронавірус.

Мета дослідження — підвищити точність прогнозування кількості нових хворих на коронавірус на базі моделі Facebook Prophet у короткостроковій перспективі шляхом створення відповідної інформаційної технології.

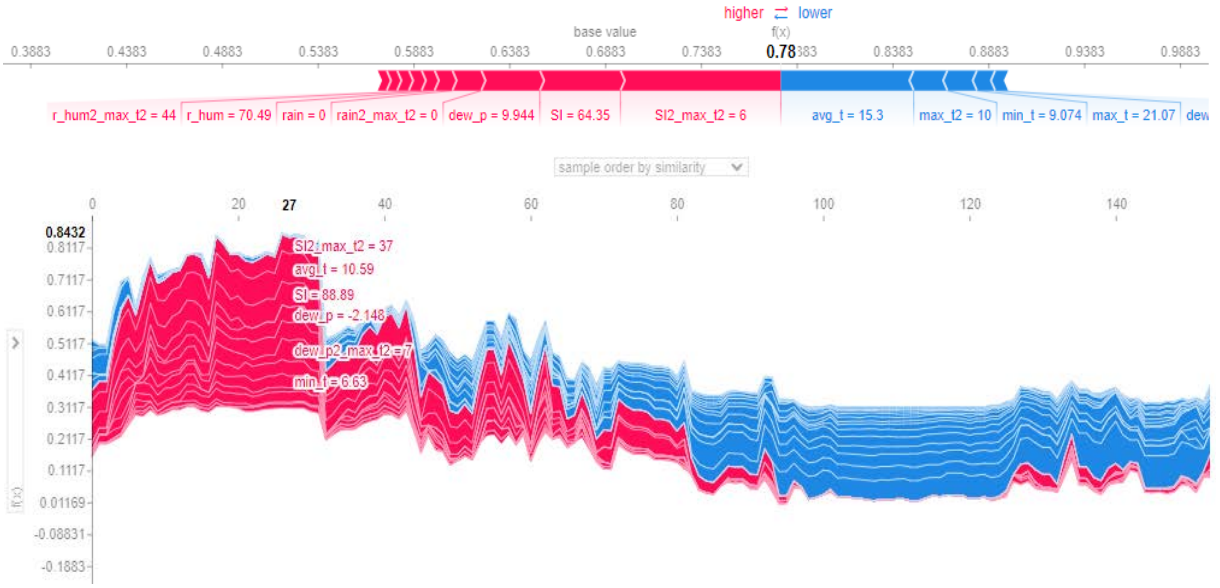
Стаття узагальнює результати досліджень, проведених протягом 2020—2022 рр. під час роботи над прогнозуванням кількості нових хворих на коронавірус в Україні та інших країнах світу, частина з яких увійшла у фахові статті та тези [10]—[13], а також у звіти [14] Робочої групи з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні, створеної Розпорядженням Президії НАН України від 3 квітня 2020 р. № 198.

Розвідувальний аналіз даних кількості нових хворих на коронавірус в Україні

Для розвідувального аналізу, моделювання та прогнозування використано дані Центру Джона Хопкінса [15] та дані Оксфордської лабораторії («Oxford COVID-19 government response tracker» — Оксфордський трекер (SI — Stringency-індекс) протикоронавірусної діяльності урядів країн світу [16]).

Проведено аналіз ступеня впливу даних про пересування людей з мобільними телефонами (Google-тренди), зміни метеоданих і даних урядового трекера датасетів [17] та деяких агрегованіших ознак, обчислених на їхній основі шляхом, наприклад, переходу від одиниць до десятків у значеннях, на прогнозування кількості $y(t)$ нових хворих на коронавірус в Україні у 2020 р. (рис. 1) [18].

Як видно на рис. 1, у різний час різні фактори мають різний вплив, зазвичай, стрибкоподібний, а отже, їх варто враховувати, але краще на їхній основі сформувати нові ознаки. У статті [10] запропоновано враховувати дати їхнього найбільшого впливу як дати аномалій: дати зміни сумарного значення урядового трекера SI (дати введення карантинних обмежень, локдаунів тощо), дати аномально теплих та сухих днів, які, зазвичай, супроводжувались збільшенням пересувань людей та підвищенням їхньої контагіозності. Також у статті [10], як дати аномалій, пропонувалось враховувати дати офіційних свят, але, як обґрунтовано у цій статті, усі дати аномалій слід враховувати з певним зсувом, оскільки через особливості перебігу коронавірусної хвороби наслідки впливу проявляються пізніше — через 4…7 діб [19], [11]. Тривалість такого зсуву Δt слід використовувати як гіперпараметр під час навчання моделі (гіперпараметром у машинному навчанні називають параметр, який має ідентифікувати дослідник за своїм спеціальним алгоритмом чи з використанням певного свого методу, на відміну від інших параметрів, які автоматично ідентифікуються типовими методами програмних бібліотек).



а)

б)

Рис. 1. Ступінь впливу даних про пересування людей з мобільними телефонами (Google-тренди), зміни метеоданих та даних урядового трекера Оксфордської лабораторії на прогнозування кількості нових хворих на коронавірус в Україні у 2020 р.: *а* — Force_plot-діаграма для заданої дати 02.10.2020 р.; *б* — компілятивна TreeExplainer-діаграма для усіх ознак разом за квітень—жовтень 2020 р. [18]

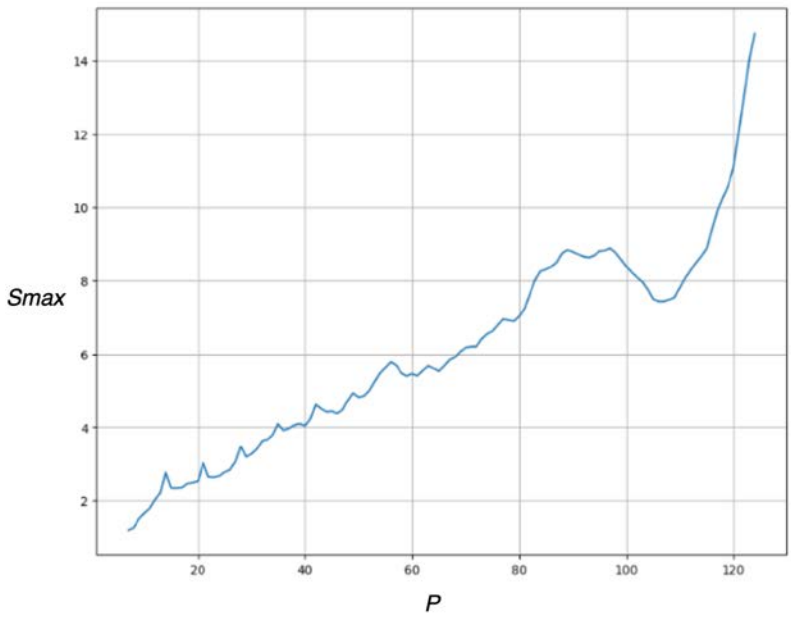


Рис. 2. Частка сезонної складової Smax, залежно від періоду *P* у добах, до амплітуди основного ряду кількості нових хворих на коронавірус в Україні у 2020 р., зі згладжуванням у 7 днів [19]

Проведено дослідження сезонності ряду. Здійснювалась його декомпозиція на тренд, сезонну складову з різним періодом та залишки. Експерименти проводились з різними періодом сезонності. Щоразу визначалось яку саме частку має амплітуда Smax сезонної складової від амплітуди (по суті, максимального значення, оскільки кількість хворих не є від'ємною) основного ряду (рис. 2).

З рис. 2 випливає, що частка сезонної складової Smax очікувано монотонно зростає зі збільшенням періоду *P* (асимптотично наближаючись до 100 %, тобто до самого ряду); перший стрибок має місце після перших 7 діб, що доволі очевидно, бо це — цикл роботи лабораторій, які проводять тестування; в подальшому теж мають місце стрибки, але вони нечітко виражені і це питання потребує окремого дослідження; суттєві нелінійності кривої свідчать імовірніше про наявність декількох видів сезонностей одночасно.

Формалізація моделі Facebook Prophet для прогнозування кількості хворих на коронавірус

Основні аспекти моделі Facebook Prophet (далі просто — Prophet) описані у статті [7] та у документації відповідної Python-бібліотеки. Математично модель Prophet для моделювання та прогнозування значень ряду $y(t)$ в залежності від часу t, записується таким чином [7]:

– для адитивного випадку $$y(t) = g(t) + s(t) + h(t) + \epsilon_t; \quad (1)$$

– для мультиплікативного випадку $$y(t) = g(t)s(t)h(t)\epsilon_t, \quad (2)$$

де $g(t)$ — тренд ряду (логістична або шматково-лінійна апроксимація даних); $s(t)$ — сезонна складова, апроксимована рядом Фур'є; $h(t)$ — складова, яка враховує вплив свят чи інших аномалій, які відбуваються нерегулярно протягом одного або кількох днів і діють з певним «вікном», тобто в діапазоні певних дат; $\epsilon(t)$ — похибка («шум») з нульовим середнім, розподілена за нормальним законом.

Велика кількість досліджень із застосування моделі (1) до прогнозування кількості нових хворих на коронавірус в Україні за даними 2020—2021 рр. показала, що краще використовувати шматково-лінійну апроксимацію тренду і варіант (2) моделі, а не (1), чи модель з логарифмічним трендом [12].

Проте, значна волатильність процесу появи нових хворих на коронавірус вимагає налаштування багатьох параметрів моделі, а тому для задач цієї статті пропонується таке удосконалення моделі (2):

$$y(t) = g(R_g, t) \cdot \prod_{i=1}^{\Phi} s\left(R_s i, P_i, n_i, t\right) \cdot h\left(\bigcup_{j=1}^{\Psi} H_j\left(R_{h\,j}, \Delta t_j, t_{0\,j}, t_{1\,j}, t\right)\right) \epsilon_t, \tag{3}$$

де R_g — ступінь регуляризації тренду (як правило, від 0,01 до 100), тобто наскільки гнучко тренд має припасовуватись до різких змін значень ряду; Ф — кількість сезонних складових, які одночасно враховуються; для кожної i-ї сезонної складової окремо задаються такі параметри: R_{si} — ступінь регуляризації, P_i — період у добах, n_i — порядок ряду Фур'є, яким описується ця складова; Ψ — кількість видів аномалій, дати яких враховуються як «свята»; для кожної j-ї аномалії («свята») окремо задаються такі параметри: R_{hj} — ступінь регуляризації, Δt_j — зсув дати відносно справжнього значення для урахування запізнення впливу, через особливості протікання коронавірусної хвороби, у добах, $[t_{0j}, t_{1j}]$ — «вікно» впливу аномалії, де t_{0j} — на яку кількість діб до дати аномалії вже починає проявлятись її вплив (залежить від типу аномалій), t_{1j} — через яку кількість діб після дати аномалії її вплив вже практично не проявляється; $H_j(\cdot)$ — матриця з датами аномалій та їхніми параметрами.

Важливо, що одну аномалію в (3) можна враховувати двома способами одночасно, наприклад, офіційні свята враховуються двічі: і як аномалія у роботі лабораторій, коли більшість лабораторій не працює — це аномалія, результат якої проявиться одразу, тобто — з параметром $\Delta t_j = 0$, і як аномалія, через скупчення людей на свята та збільшення їхньої контагіозності, результат якого проявиться через $\Omega = 4 \ldots 7$ діб.

У загальному випадку, кількість N_k гіперпараметрів моделі (3), тобто параметрів, які потрібно ідентифікувати, становить

$$N_k = 1 + 3\Phi + 4\Psi. \tag{4}$$

У разі ідентифікації параметрів методом повного перебору, коли для кожного гіперпараметра вибирається один з m варіантів значень, то кількість N_m можливих моделей, які потрібно ідентифікувати, визначиться так:

$$N_M = m^{N_K}. \tag{5}$$

Наприклад, для випадку Ф = 3 видів сезонності та Ψ = 5 видів аномалій $N_k = 30$. Якщо кожний параметр буде вибраний хоча б з $m = 4$ варіантів значень, тоді це становитиме $N_M = 1{,}15 \cdot 10^{18}$ комбінацій, що — забагато. А тому слід шукати оптимізованіші підходи.

Для ефективного застосування моделі (3) її гіперпараметрами є такі:

$$K = \left[R_g, R_{s1} \ldots R_{s\Phi}, n_1 \ldots n_\Phi, P_1 \ldots P_\Phi, R_{h1} \ldots R_{h\Psi}, \Delta t_1 \ldots \Delta t_\Psi, t_0 \ldots t_{0\Psi}, t_1 \ldots t_{1\Psi} \right]. \tag{6}$$

У разі застосування методу повного перебору, наприклад, з 4 варіантами, за формулою (5) отримаємо забагато можливих комбінацій. Звичайно, можна застосувати байєсівські методи та Python-бібліотеки HyperOpt, але такий підхід дає менш достовірні оцінки параметрів, аніж повний перебір варіантів.

Ідентифікація структури та параметрів моделі Facebook Prophet для прогнозування кількості хворих на коронавірус в Україні

Для пришвидшення ідентифікації параметрів (6) моделі (3) у статті [10] запропоновано двоетапний паралельно-послідовний метод ідентифікації, оснований на гіпотезі про те, що характер впливу тренду і періодичних складових усього ряду суттєво відрізняється від характеру впливу пооди-

ноких аномалій у певні моменти часу, що дозволяє їхні параметри ідентифікувати окремо у 2 етапи, до прикладу для випадку $\Phi = 3,\ \Psi = 4$.

$$\text{Stage 1: } \begin{cases} R_g = R_{g0},\ R_{s1} = R_{s10},\ n_1 = n_{10},\ R_{s2} = R_{s20}, \\ n_2 = n_{20},\ R_{s3} = R_{s30},\ P_3 = P_{30},\ n_3 = n_{30}, \\ R_h \to R_{hopt},\ t_0 \to t_{0opt},\ t_1 \to t_{1opt},\ \Omega \to \Omega_{opt}, M \to M_{opt}; \end{cases} \tag{7}$$

$$\text{Stage 2: } \begin{cases} R_h = R_{hopt},\ t_0 = t_{0opt},\ t_1 = t_{1opt},\ \Omega = \Omega_{opt},\ M = M_{opt}, \\ R_g \to R_{gopt},\ R_{s1} \to R_{s1opt},\ n_1 \to n_{1opt},\ R_{s2} \to R_{s2opt},\ n_2 \to n_{2opt}, \\ R_{s3} \to R_{s3opt},\ P_3 \to P_{3opt},\ n_3 \to n_{3opt}, \end{cases}$$

де M — тип моделі (адитивний чи мультиплікативний — модель (1) або (2)), змінні з індексом «0» — це початкове наближення їхніх значень, отримане на основі розвідувального аналізу, а змінні з індексом «opt» — це оптимальні значення, обчислені на відповідному етапі ідентифікації.

Окремим питанням є врахування багатохвильової природи ряду кількості нових хворих в Україні, який не є класичним періодичним рядом з одним типом сезонності. Вочевидь, ряд має довготривалий період (більше року), але мультиплікативне врахування складових згідно з (3) суттєво ускладнює можливості щодо ідентифікації цього періоду. Для цього розроблено окремий метод, описаний у статті [11]. У ній проведено дослідження того, як змінюється форма графіка ряду Фур'є, в залежності від його порядку, зокрема в який бік і яким чином зсувається найбільший пік хвилі. Зазначено, що в загальному випадку, насамперед, на графіку кількості хворих на коронавірус немає чітко вираженої періодичності, по-друге, майже немає нульових значень, тобто чітко можна знайти тільки максимуми хвиль. А тому запропоновано співвідношення, яким чином можна оцінювати порядок n ряду Фур'є за 10 % верхньої частини однієї хвилі і по ¼ періоду (тобто аналізувати достатньо тільки верхівку однієї половини хвилі, яку, зазвичай, завжди можна виокремити на графіку), а період P визначається за умовним початком і кінцем хвилі, які екстраполюються на основі мінімального значення кількості нових хворих на коронавірус $y_{\min}$ початку хвилі, максимального значення хвилі на піку $y_{\max}$ моментів часу для цих точок $t_{y_{\min}}$ та $t_{y_{\max}}$ відповідно.

$$n_{opt} = \varphi_n\left(y_{0,9\max},\ t_{y_{0,9\max}}\right),\ y_{0,9\max} = 0,9\,y_{\max}; \tag{8}$$

$$P = \varphi_p\left(y_{\min}, t_{y_{\min}}, y_{\max}, t_{y_{\max}}\right), \tag{9}$$

де $t_{y_{0,9\max}}$ — дата, коли крива хвилі кількості нових хворих на коронавірус вперше перетинає рівень $y_{0,9\max}$, що дорівнює 90 % значення від максимального значення піку $y_{\max}$ аналізованої хвилі, яка у лівій (зростальній) частині першої чверті її періоду; φ_n та φ_p — деякі функції, детально описані у статті [11].

Вираз (8) не завжди забезпечує гарні результати, якщо хвилі мають надто різний період. У таких випадках вирази (8), (9) задають початкові наближення, які далі потрібно уточнювати як гіперпараметри.

Здійснено низку таких удосконалень і спрощень, що за умови використання 4-х допустимих варіантів для кожного параметра у 2 етапи становить лише 128 варіантів, і працює відносно швидко [20].

Порівняльний аналіз трендів прогнозу сусідніх регіонів

Важливо не тільки максимально точно порахувати прогноз — треба ще й візуалізувати результат у такий спосіб, щоб можна було дослідити нові закономірності та щоб це було зручно для подальшого ухвалення рішень. Наприклад, у задачі прогнозування за багато тижнів сусідніх країн чи областей, можна спробувати знайти певні просторово-часові закономірності щодо поширення хвиль у різних напрямках. Наприклад, що надмірне зростання кількості хворих у центральній та східній Європі через певну кількість тижнів дозволить прогнозувати зростання кількості хворих в Україні. Традиційно для такої задачі створюють лінійний прогноз та картують нахил відповідних ліній. Однак, у роботі [12] запропоновано відокремлювати у прогнозі періодичні складові і вплив дат аномалій, що дозволяє порівнювати вплив саме основного тренду, ігноруючи місцеві особливості (свята та локдауни країн чи регіонів, унікальний графік роботи лабораторій та лікарень то-

що), що також дозволяє точніше аналізувати основні закономірності динаміки процесу. Математично, цей підхід виглядає таким чином. У прогнозі моделі FB Prophet з оптимальними параметрами визначаються усі складові виразу (3) $g(t)$, $s_i(t)$, $h(t)$. А тоді пропонується брати до уваги тільки складову $g(t)$ та апроксимувати її шматково-лінійною функцією з кроком в 1 тиждень у вигляді

$$g\left(R_{gopt}, t_q\right) = b_g + k_g t_q, \tag{10}$$

де t_q — час (зазвичай, номер тижня), b_g, k_g — числові коефіцієнти.

З (10) знаходимо

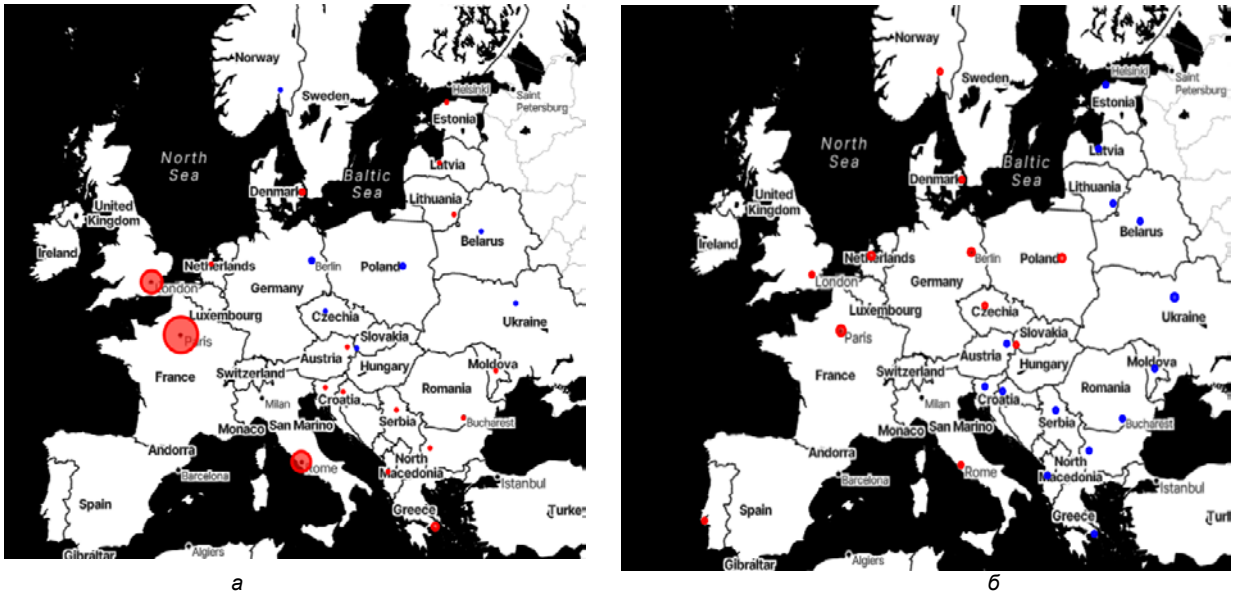
$$k_g = \frac{g\left(R_{gopt}, t_q\right)}{t_q}. \tag{11}$$

Далі відображаємо цей нахил колами на карті. Наприклад, для країн світу (рис. 3) [12]:

– центр кола збігається з координатами столиці країни;

– визначається колір кола: червоний, якщо коефіцієнт (11) є додатним (наростання кількості хворих на коронавірус) або дорівнює 1, чи — синій, якщо коефіцієнт є від'ємним;

– радіус кола r_g лінійно пропорційний значенню k_g.



а *б*

Рис. 3. Картограма прогнозу 10.01.2022 — 16.01.2022 за даними датасету [21] з різними радіусами: *а* — за формулою $r_g = 0{,}75 + 0{,}005\, k_g$; *б* — за формулою $r_g = 2 + 0{,}004\, k_g$ [22]

Розроблення інформаційної технології

Розроблено архітектуру інформаційної технології для реалізації вищенаведених авторських методів, яка показана схематично на рис. 4.

Вирішено розділити інформаційну технологію на 3 основі модулі, а саме:

– *модуль оброблення первинних* даних, який здійснює зчитування даних з датасету первинних даних захворюваності на коронавірус, їхнє попереднє очищення та виконує розвідувальний аналіз даних (Exploratory Data Analysis — EDA), для визначення характеристик вибраного набору даних.

– *модуль ідентифікації параметрів та прогнозування*, який зі свого боку, з урахуванням результатів EDA, визначає початкові наближення параметрів та області допустимих значень і виконує ідентифікацію параметрів моделі Prophet, з циклічним тренуванням моделі. Оптимальна модель з найменшою відносною похибкою далі використовується для циклічного генерування прогнозу кількості нових хворих у заданому регіоні. Після цього отриманий прогноз зображується у вигляді графіка порівняння відносно попередніх прогнозів. Цей модуль запускається багато разів для кожного регіону окремо. Результати зберігаються в датасет згенерованих прогнозів.

– *модуль створення картограми інфекційного впливу* використовує результати прогнозу з модуля, описаного вище, виконує апроксимацію та визначення нахилу тренду інфекційного впливу,

після чого ці тренди зображуються у вигляді графіка. Результати апроксимації зберігаються у датасет тенденцій та відображаються у вигляді інтерактивної картограми інфекційного впливу в масштабі карти Європи або карти світу, в залежності від обсягу дослідження.

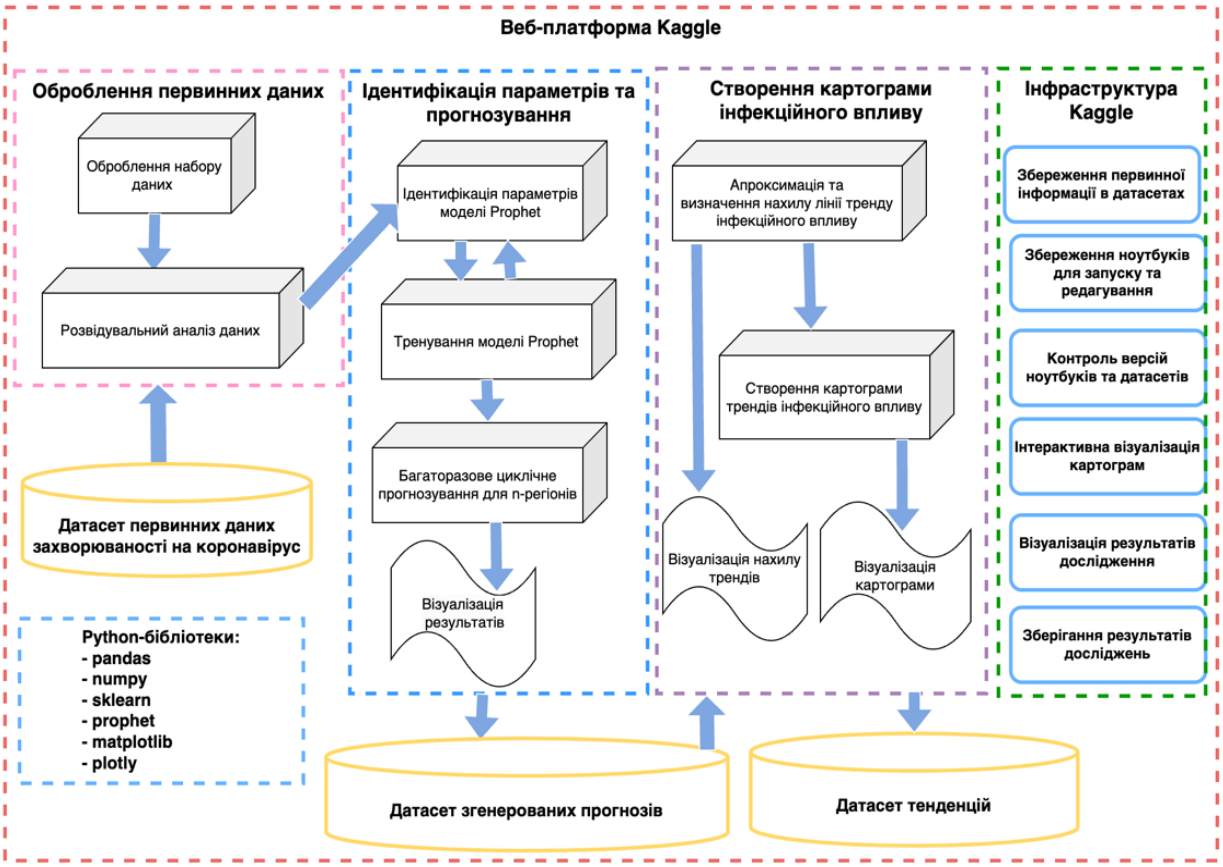


Рис. 4. Архітектура розробленої інформаційної технології

Для розроблення та застосування інформаційної технології вибрано популярну веб-платформу Kaggle, яка дозволяє розроблення ІТ в межах програмного середовища Jupyter Notebook, адаптованого для використання в онлайн-режимі. Цей веб-ресурс зберігає програмний код у форматі інтерактивних «ноутбуків», при цьому кожна нова версія ноутбука зберігається окремо, і за потреби можна обрати конкретну його версію. Також Kaggle надає хмарні ресурси для запуску програмного коду, що дає можливість ефективно тестувати та тренувати модель в основі цієї ІТ. До того ж, ця веб-платформа надає доступ до різноманітних наборів даних у зручному форматі, що спрощує задачу їхнього зчитування. В цілому, Kaggle надає весь спектр інструментів, потрібних для продуктивної розробки інформаційних технологій у сфері машинного навчання та аналізу даних.

Застосування розробленої інформаційної технології для прогнозування кількості нових хворих на коронавірус в Україні

Як зазначено раніше, результати роботи розробленої інформаційної технології використовувались в аналітичних звітах Робочої групи з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні (далі — РГ), тому доречно використати ці напрацювання для оцінювання ефективності застосування розробленої ІТ. Отже, на прикладі прогнозів, наведених у 25 звітах РГ, порівняємо похибки прогнозів за моделлю Prophet з ідентифікованими параметрами, використовуючи запропоновану ІТ та модель SEIR-U науковців з НАН України РГ. Таке порівняння дає можливість оцінити динаміку точності прогнозування захворюваності протягом листопада 2020 р. — лютого 2022 року. В таблицю внесемо такі значення:

– номер звіту, дати прогнозу;

– точність δ_{valP} прогнозу оптимальної моделі Prophet на валідаційному датасеті (зазвичай, 2 останні тижні ряду, які не використовувались для тренування моделі);

– значення відносної похибки прогнозу δ_{testP} на тестових даних за моделлю Prophet;

– відносна похибка δ_{testS} на тестових даних за моделлю SEIR–U;

– версія моделі: M_o для штаму Omicron, M_δ для штаму Delta, та M_α для штаму Alpha;

– характер кривої: D_{-1} відповідає спаданню хвилі захворюваності, D_1 зростанню, а D_0 характеризує криву між хвилями захворюваності.

Результати прогнозування часового ряду кількості хворих на коронавірус в Україні протягом 2020—2022 рр.

№ Звіту	Дати прогнозу	δ_{valP}, %	δ_{testP} для Prophet, %	δ_{testS} для SEIR-U, %	Версія моделі	Характер кривої
РГ-62 (22.02.2022)	23.02.2022—08.03.2022	4,67	87,14	131,2	M_o	D_{-1}
РГ-61 (08.02.2022)	09.02.2022—22.02.2022	1,81	36,39	29,93	M_o	D_1
РГ-60 (26.01.2022)	27.01.2022—08.02.2022	0,98	36,22	32,95	M_o	D_1
РГ-59 (11.01.2022)	12.01.2022—25.01.2022	16,53	63,60	49,61	M_o	D_0
РГ-58 (21.12.2021)	22.12.2021—04.01.2022	7,56	24,5	34,09	M_o	D_{-1}
РГ-57 (07.12.2021)	08.12.2021—21.12.2021	5,07	28,54	25,55	M_o	D_{-1}
РГ-56 (23.11.2022)	24.11.2021—07.12.2021	16,9	19,51	23,21	M_o	D_{-1}
РГ-55 (09.11.2021)	10.11.2021—23.11.2021	9,3	26,57	24,28	M_o	D_1
РГ-54 (26.10.2021)	27.10.2021—9.11.2021	7,28	21,95	23,43	M_o	D_1
РГ-53 (12.10.2021)	13.10.2021—26.10.2021	7,75	21,86	24,01	M_o	D_1
РГ-52 (29.09.2021)	29.09.2021—12.10.2021	5,41	27,67	21,61	M_o	D_1
РГ-51 (14.09.2021)	15.09.2021—27.09.2021	25,68	33,52	27,81	M_o	D_1
РГ-42 (20.04.2021)	21.04.2021—03.05.2021	8,76	24,17	32,63	M_δ	D_{-1}
РГ-41 (06.04.2021)	07.04.2021—19.04.2021	9,3	43,39	24,1	M_δ	D_1
РГ-40 (23.03.2021)	24.03.2021—05.04.2021	7,2	30,93	25,61	M_δ	D_1
РГ-39 (10.03.2021)	11.03.2021—22.03.2021	5,4	23,28	22,22	M_δ	D_1
РГ-38 (22.02.2021)	23.02.2021—1.03.2021	20,15	18,41	17,75	M_δ	D_0
РГ-37 (08.02.2021)	09.02.2021—15.02.2021	12,7	19,9	18,1	M_δ	D_0
РГ-36 (25.01.2021)	26.01.2021—01.02.2021	19,9	18,95	13,73	M_δ	D_{-1}
РГ-35 (11.01.2021)	12.01.2021—18.01.2021	18,4	21,85	15,11	M_δ	D_{-1}
РГ-34 (28.12.2020)	29.12.2020—11.01.2021	7,48	30,75	19,61	M_δ	D_{-1}
РГ-32 (14.12.2020)	15.12.2020—28.12.2020	3,5	22,12	11,65	M_α	D_{-1}
РГ-31 (07.12.2020)	08.12.2020—21.12.2020	6,4	9,21	17,87	M_α	D_1
РГ-30 (30.11.2020)	01.12.2020—13.12.2020	3,2	32,47	30,81	M_α	D_1
РГ-29 (23.11.2020)	24.11.2020—06.12.2020	2,2	15,27	21,29	M_α	D_1

За даними таблиці, можна зазначити, що запропонована технологія на основі моделі Prophet в порівнянні з моделлю SEIR-U іноді точніша у 1,1…2 рази ніж модель SEIR-U, причому у 2021 р. кількість таких прогнозів збільшилась у порівнянні з 2020 р., що свідчить про ефективність запропонованої ІТ для ідентифікації моделі Prophet. Суттєва відмінність похибки для валідаційного та тестового датасету у 2022 р. свідчить про збільшення волатильності ряду, через вплив вакцинації населення, мутацій різних штамів та інші нові фактори, яке вочевидь потребує або зміни структури моделі, або збільшення ітерацій ідентифікації.

Висновки

Здійснено огляд відомих моделей для прогнозування поширення коронавірусу: на основі диференціальних рівнянь SIR, SEIR, SEIR-U; моделі часових рядів ARIMA; моделі часових рядів FB Prophet та гібридні моделі машинного навчання на основі ARIMA/Prophet та ін.

Запропоновано математичний опис розробленого раніше методу паралельно-послідовної багатопараметрової ідентифікації Prophet-моделі для короткострокового прогнозування часового ряду кількості хворих на коронавірус в заданому регіоні: сили та розміру вікна впливу дат аномалій, ступеня регуляризації та типу моделі (адитивна чи мультиплікативна), кількості параметрів Фур'є і ступеня регуляризації 3-х різних періодичних складових, які враховують добові, внутрішньотижневі та багатоденні закономірності, характерні для коронавірусу, що дозволяє підвищити точність

прогнозів та глибше дослідити закономірності, які впливають на цей часовий ряд. Параметри ряду Фур'є оцінюються з використанням розробленого раніше методу на основі певних емпіричних співвідношень.

Запропоновано підхід до аналізу можливого інфекційного впливу сусідніх регіонів за нахилом елементів шматково-лінійної апроксимації кривих прогнозів зміни тижневого тренду, який виділяє запропонована Prophet-модель та відокремлює від нього періодичні складові і вплив дат аномалій з подальшим генеруванням картограм.

Продемонстрована можливість поєднання усіх методів і підходів в єдиній інформаційній технології прогнозування часових рядів кількості хворих на коронавірус методами машинного навчання (ІТ) з єдиною нотацією усіх змінних. Розроблена архітектура цієї ІТ та подано опис її складових. Обґрунтовано доцільність використання веб-платформи Kaggle для реалізації запропонованої ІТ.

Проведено порівняння відносної похибки точності прогнозування за даними 2020—2022 рр. за моделлю Prophet, що використовується в розробленій інформаційній технології, та за моделлю SEIR-U, розробленої науковцями НАН України у складі Робочої групи з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні. За результатами цього порівняння можна зазначити, що запропонована технологія на основі моделі Prophet дає іноді точність, у 1,1—2 рази вищу ніж модель SEIR-U, причому з часом кількість таких прогнозів збільшилась, що свідчить про ефективність удосконалення розробленої інформаційної технології протягом цих років.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] І. О. Бровченко, «Розробка математичної моделі поширення епідемії COVID-19 в Україні,» *Світогляд*, № 2 (82), с. 2-14, 2020.

[2] P. Furtado, "Epidemiology SIR with Regression, Arima, and Prophet in Forecasting Covid-19," *Engineering Proceedings*, no. 5(1), 52, July, 2021. https://doi.org/10.3390/engproc2021005052 .

[3] R. Ospina, J.A.M. Gondim, V. Leiva, and C. Castro, "An Overview of Forecast Analysis with ARIMA Models during the COVID-19 Pandemic: Methodology and Case Study in Brazil," *Mathematics*, 11(14):3069, May, 2023. https://doi.org/10.3390/math11143069 .

[4] A. Hernandez-Matamoros, H. Fujita, T. Hayashi, and H. Perez-Meana, "Forecasting of COVID19 per regions using ARIMA models and polynomial functions," *Applied Soft Computing*, vol. 96, 106610, ISSN 1568-4946, November, 2020. https://doi.org/10.1016/j.asoc.2020.106610 .

[5] G. Perone, "Using the SARIMA Model to Forecast the Fourth Global Wave of Cumulative Deaths from COVID-19: Evidence from 12 Hard-Hit Big Countries," *Econometrics*, 10(2):18. January, 2022. https://doi.org/10.3390/econometrics10020018 .

[6] P. Harjule, V. Tiwari, and A. Kumar, "Mathematical models to predict COVID-19 outbreak : An interim review," *Journal of Interdisciplinary Mathematics*, no. 24, pp. 1-26, 2021. https://doi.org/10.1080/09720502.2020.1848316 .

[7] S. Taylor, and B. Letham, "Forecasting at Scale," *The American Statistician*, 72, 2017. https://doi.org/10.1080/00031305.2017.1380080 .

[8] D. Borges, and M. C. V. Nascimento, "COVID-19 ICU demand forecasting: A two-stage Prophet-LSTM approach," *Applied Soft Computing*, vol. 125, 2022. https://doi.org/10.1016/j.asoc.2022.109181 .

[9] G. Perone, "Comparison of ARIMA, ETS, NNAR, TBATS and hybrid models to forecast the second wave of COVID-19 hospitalizations in Italy," *The European Journal of Health Economics*, no. 23, pp. 917-940, 2022. https://doi.org/10.1007/s10198-021-01347-4 .

[10] В. Б. Мокін, А. В. Лосенко, і А. Р. Ящолт, «Інформаційна технологія аналізу та прогнозування кількості нових випадків хвороби на коронавірус SARS-COV-2 в Україні на основі моделі Prophet,» *Вісник Вінницького політехнічного інституту*, № 5, с. 71-83, Лист. 2020. https://doi.org/10.31649/1997-9266-2020-152-5-71-83 .

[11] В. Б. Мокін, А. В. Лосенко, і А. Р. Ящолт, «Інформаційна технологія аналізу та прогнозування багатохвильової кількості нових випадків захворювань на коронавірус COVID-19 на основі моделі Prophet,» *Вісник Вінницького політехнічного інституту*, № 6, с. 65-75, Груд. 2020. https://doi.org/10.31649/1997-9266-2020-153-6-65-75 .

[12] В. Б. Мокін, і А. В. Лосенко, «Картування тренду тижневих прогнозів за моделлю Facebook Prophet зміни кількості нових хворих на коронавірус у країнах Європи протягом січня-березня 2021 року,» на *Науково-технічна конференція підрозділів ВНТУ L*, Вінниця, 10-12 березня, 2021 р.

[13] В. Б. Мокін, М. В. Дратований, А. В. Лосенко, і С. О. Жуков, «Прогнозування хвиль коронавірусу на основі відновленої когнітивної карти міжрегіонального впливу,» *Інформаційні технології та комп'ютерна інженерія*, т. 52, вип. 3, с. 86-94, Груд. 2021.

[14] Робоча група з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні, *Прогноз розвитку епідемії COVID-19 в Україні на 14–28 грудня 2020 року* («*Прогноз РГ-32*»), базова установа — Інститут проблем математичних машин і систем НАН України, створена Розпорядженням Президії НАН України від 3 квітня 2020 р. № 198. [Електронний ресурс]. Режим доступу: http://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=7277 Дата звернення: 14.10.2023 .

[15] Anthony Goldbloom, *COVID-19 data from John Hopkins University*. Kaggle. [Електронний ресурс]. Режим доступу: https://www.kaggle.com/datasets/antgoldbloom/covid19-data-from-john-hopkins-university . Дата звернення: 13.10.2023

[16] T. Hale, et al., "A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker)," *Nature Human Behaviour*. 2021. https://doi.org/10.1038/s41562-021-01079-8 .

[17] Vitalii Mokin, and Arsen Losenko, "COVID-19-UA: Regression with Google mobility". 2022. [Electronic resource].

Available: https://www.kaggle.com/code/vbmokin/covid-19-ua-regression-with-google-mobility . Accessed: 13.10.2023.

[18] Vitalii Mokin, and Arsen Losenko, "COVID-19 in Ukraine: Explanation of patterns". [Electronic resource]. Available: https://www.kaggle.com/vbmokin/covid-19-in-ukraine-explanation-of-patterns . Accessed: 13.10.2023.

[19] Робоча група з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні, *Прогноз розвитку епідемії COVID-19 в Україні на 14–28 грудня 2020 року* («Прогноз РГ-32»), базова установа — Інститут проблем математичних машин і систем НАН України, створена Розпорядженням Президії НАН України від 3 квітня 2020 р., № 198. [Електронний ресурс]. Режим доступу: http://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=7277 . Дата звернення: 12.10.2023.

[20] Vitalii Mokin, and Arsen Losenko, "COVID in UA: Prophet with 4, Nd seasonality," *Kaggle Notebook*. [Electronic resource]. Available: https://www.kaggle.com/code/vbmokin/covid-in-ua-prophet-with-4-nd-seasonality . Accessed: 12.10.2023.

[21] Vitalii Mokin, and Arsen Losenko, "COVID-19: Forecast trends for the many countries," Kaggle Notebook. [Electronic resource]. Available: https://www.kaggle.com/datasets/vbmokin/covid19-forecast-trends-for-the-many-countries/ . Accessed: 12.10.2023.

[22] Vitalii Mokin, and Arsen Losenko, "COVID-19: Week trends 70 countries mapping," *Kaggle Notebook*. [Electronic resource]. Available: https://www.kaggle.com/code/vbmokin/covid-19-week-trends-70-countries-mapping/notebook . Accessed: 12.10.2023 .

Рекомендована кафедрою системного аналізу та інформаційних технологій ВНТУ

Стаття надійшла до редакції 19.10.2023

Лосенко Арсен Володимирович — аспірант кафедри системного аналізу та інформаційних технологій, e-mail: arsenlosenko@gmail.com

A. V. Losenko[1]

Information Technology of Time Series Forecasting of New COVID-19 Disease Patients Based on Prophet Model

[1]Vinnytsia National Technical University

The article describes the results of the development of information technology for forecasting the time series of the number of COVID-19 patients. A review of the developments on this subject was carried out and the choice of the Facebook Prophet library as the basis for the developed IT was substantiated. It is proposed to increase the accuracy of the forecasting of the number of new coronavirus patients in the short-term using the selected model.

Mathematical substantiation of the available parallel-serial iterative methods for identifying the parameters of the Facebook Prophet model to take into acco0unt the significant volatility of the number of new patients was proposed. The methods make it possible to determine the hyperparameters of the trend of the series, its seasonal components and anomalies affecting the value of this series. The approach to comparative analysis of the influence of the main incidence trends of regions (or countries) without considering the influence of anomalies and seasonal components has been formalized by creating cartograms that enable to analyze trends in the spread of the disease in a given region.

The architecture of the proposed information technology is developed, and its components are described. A Python software based on the Kaggle platform was created, it implements this technology, the results of its application is compared with the SEIR-U model, developed by the scientists of the National Academy of Sciences of Ukraine, based on the reports of the Working Group on Mathematical Modeling of Problems Related to the SARS-CoV-2 Coronavirus Epidemic, at the National Academy of Sciences of Ukraine according to data, obtained in the period 2020—2022. The comparison, carried out, proved the effectiveness of the proposed information technology.

Keywords: information technology, COVID-19, time series forecasting, Prophet, Fourier series, artificial intelligence, forecasting development scenarios.

Losenko Arsen V. — Post-Graduate Student of the Chair of System Analysis and Information Technologies, e-mail: arsenlosenko@gmail.com