ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА КОМП'ЮТЕРНА ТЕХНІКА

https://doi.org/10.31649/1997-9266-2020-153-6-65-75

УДК 004.9+578.834.1

В. Б. Мокін[1]
А. В. Лосенко[1]
А. Р. Яцолт[1]

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ БАГАТОХВИЛЬОВОЇ КІЛЬКОСТІ НОВИХ ВИПАДКІВ ЗАХВОРЮВАНЬ НА КОРОНАВІРУС COVID-19 НА ОСНОВІ МОДЕЛІ PROPHET

[1]Вінницький національний технічний університет

Удосконалено раніше розроблену авторами інформаційну технологію аналізу та прогнозування кількості нових підтверджених випадків захворювань на коронавірус COVID-19, викликаною інфекцією SARS-CoV-2, на прикладі щодобових сумарних по Україні даних поточної «хвилі» з урахуванням різних свят і псевдосвят, які можуть мати аномальний вплив. Створена раніше технологія була працездатною лише для ділянки невпинного зростання значень однієї хвилі, а удосконалена вже може застосовуватись для аналізу та прогнозування даних протягом усього періоду, оскільки враховує багатохвильову природу цього явища. Запропоновано алгоритм ідентифікації параметрів кожної хвилі. Розроблено низку математичних співвідношень, які дозволяють у першому наближенні оцінити дату початку, завершення та період кожної хвилі, навіть за випадку, коли одна хвиля переходить в іншу.

Запропоновані нові емпіричні співвідношення для оцінювання порядку ряду Фур'є для опису коливального процесу кожної хвилі лише по 10 % її значень у верхівці, оскільки, зазвичай, такі дані є в явному вигляді, інакше дані не будуть розпізнані як окрема хвиля. Співвідношення виведені окремо для випадку лише додатних коефіцієнтів, коли пік розташований ліворуч від середини напівперіоду, і окремо — для випадку знакозмінного ряду, коли пік розташований праворуч від неї. Однак, ці наближені оцінки рекомендовано уточнювати у певному діапазоні можливих значень, оскільки, в загальному випадку різних варіантів значень амплітуди гармонік запропоновані співвідношення можуть давати занижені оцінки.

Запропоновано застосовувати ідентифіковану за цією технологією модель для прогнозування найпесимістичнішого та найоптимістичнішого сценаріїв розвитку явища, тобто зміни кількості нових підтверджених випадків захворювань на коронавірус COVID-19 у майбутньому у заданій країні.

Створено програмне забезпечення на Python на базі платформи Kaggle, яке застосовано, як для України, так і ще для 69 країн світу. За допомогою ідентифікованих моделей отримано низку важливих висновків щодо розуміння закономірностей поширення коронавірусу як в Україні, так і в інших 69 країнах світу. Результати передано у Робочу групу з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні.

Ключові слова: *інформаційна технологія, COVID-19, прогнозування часових рядів, Prophet, ряд Фур'є, штучний інтелект, прогнозування сценаріїв розвитку*

Вступ

У нашій статті [1] розроблено інформаційну технологію аналізу та прогнозування кількості нових підтверджених випадків захворювань на коронавірус COVID-19 у заданому регіоні (країні, області, населеному пункті), викликаної інфекцією SARS-CoV-2, та випробували її на прикладі щодобових сумарних по Україні даних щодо однієї найбільшої «хвилі» (з 6 липня 2020 р.) з урахуванням різних свят і псевдосвят, які можуть мати аномальний вплив. Побудовані за цією технологією моделі показали гарні результати і для України, і для декількох десятків країн світу за наявними даними спостережень. Але подальші спостереження показали неспроможність цієї моделі до

прогнозування, оскільки ділянка сталого наростання завершилась, крива досягла чергового локального максимуму і почалось зменшення кількості нових хворих. Тому, модель, побудована, виключно на ділянці зростання, перестала бути адекватною. Не могла вона й моделювати весь ряд даних у ретроспективі — протягом року, де було ще дві хвилі зі значно меншими локальними максимумами. Для цього потрібні інші моделі, які враховують багатохвильову природу процесу. Подібними дослідженнями займаються й в інших країнах [2]—[6], але точність їх моделювання не достатньо висока, передусім, через не дуже високу адекватність і точність базової моделі, яка використовується для описування основного тренду, який накладається на ідентифіковані хвилі. Як показав проведений аналіз, більшість авторів моделей використовує для описування хвиль відомі ряди Фур'є [5], [6] — саме цими рядами моделює різні сезонні складові і модель Prophet. Отже, доцільно удосконалити створену нами раніше інформаційну технологію, описану у статті [1], виконавши адекватніший опис багатохвильової природи часового ряду.

Мета дослідження — підвищити точність прогнозування кількості нових підтверджених випадків захворювань на коронавірус COVID-19 у регіоні протягом тривалого часу з урахуванням їх багатохвильової природи, удосконаливши авторську інформаційну технологію на основі моделі Prophet з урахуванням різних свят і псевдосвят.

Опис створеної раніше інформаційної технології

Нагадуємо, що описана у статті [1] авторська інформаційна технологія аналізу та прогнозування кількості нових підтверджених випадків захворювань на коронавірус COVID-19 побудована на основі моделі Prophet, в якій адаптивно ідентифікуються такі параметри:

– розмір вікна, сила впливу (масштаб), режим (мультиплікативний чи адитивний) урахування та ступінь регуляризації значень в аномальні дати (свята і псевдосвята);

– мультиплікативність чи адитивність урахування, ступінь регуляризації та кількість коефіцієнтів ряду Фур'є для опису тижневої (7-денної) сезонності;

– мультиплікативність чи адитивність врахування, ступінь регуляризації та кількість коефіцієнтів ряду Фур'є для опису іншої сезонності з періодом у n днів (аналіз показав, що оптимальним, наприклад, для України в осінній період $n = 4$ днів).

Як свята і псевдосвята, які враховані у цій моделі з різними параметрами вікна впливу (деякі з нульовим вікном, тобто в той же день і вплив на сусідні дати відсутній, а деякі — із запізненням на від 6 до 10 днів) виділялись такі:

1. Державні свята.

2. Дати, коли одночасно було дуже тепло і без опадів, коли різко збільшувалась кількість людей у місцях відпочинку — «метеопаттерни».

3. Дні послаблення карантину за Stringency-індексом (за даними «Oxford COVID-19 government response tracker» — Оксфордського трекеру коронавірусної діяльності урядів країн світу: https://www.bsg.ox.ac.uk/research/research-projects/oxford-covid-19-government-response-tracker), які містяться у відомому датасеті Google-платформи «COVID-19 Open Data» (https://github.com/GoogleCloudPlatform/covid-19-open-data), котрий відображає усі послаблення карантину, згідно з рішеннями уряду України, за 17 критеріями. Дати, коли ця сума зменшувалась, формалізуються як дати послаблення карантину.

4. Дні свят для врахування аномально малої кількості тестувань на свята.

Детальніше графіки цих даних, аналіз їх значень, як саме вони припасовувались та ідентифікувались параметри вікна впливу див. у статті [1].

Варто зазначити, що до дат п. 3 можуть додаватись так звані дати «карантину вихідного дня», якими в Україні були 14, 15, 21, 22, 28, 29 листопада 2020 р. (у моделі статті [1] вони ще не враховувались, оскільки на момент подачі статті у журнал, ще не було достатньо даних спостережень для цього). Їх урахування згодом дозволило зменшити відносну похибку з 4,34 % до 3,95 %, що зазначено у розділі 3 звіту [7], співавторами якого є автори цієї статті В. Б. Мокін і А. В. Лосенко.

Модель випробовувалась за відкритими офіційними даними РНБО України, які оновлюються щодня і є доступними по API з веб-порталу (https://covid19.rnbo.gov.ua/). Саме у такий спосіб їх збирають усі світові веб-сервіси. Для інших країн використовувався вищезгаданий датасет «COVID-19 Open Data», але дані по Україні там, на жаль, не усі бувають достовірними.

Удосконалення інформаційної технології

Як відомо, ряд Фур'є демонструє високу ефективність для опису періодичних кривих складної форми [6], [8]—[11]. Ним можна описувати неперіодичні криві, але тоді тільки в окремі інтервали часу. Саме такий вигляд має часовий ряд кількості нових підтверджених випадків захворювань на коронавірус COVID-19 у 2020 р. за даними РНБО України (рис. 1).

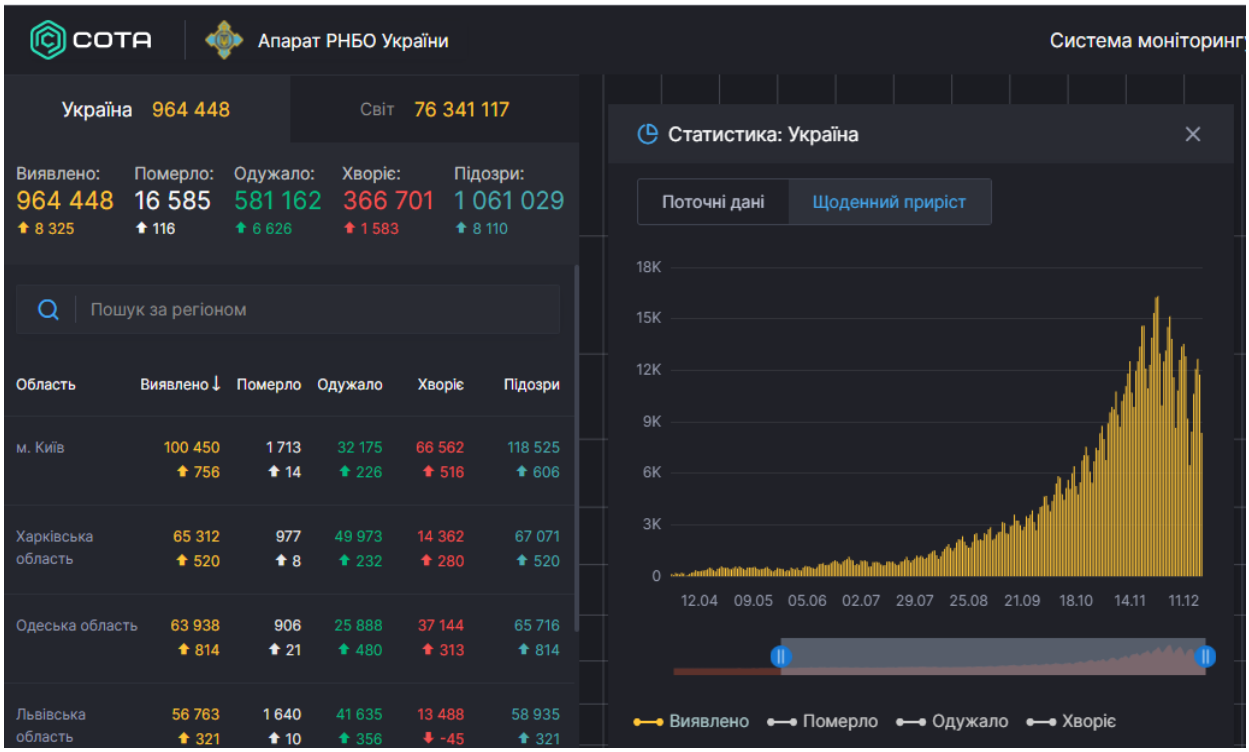


Рис. 1. Кількість нових підтверджених випадків захворювань на коронавірус COVID-19 у 2020 р. за даними РНБО України (https://covid19.rnbo.gov.ua/)

На рис. 1 (графік праворуч), якщо відфільтрувати тижневу періодичність, обумовлену графіком роботи лабораторій (особливість збирання даних тестування), добре видно дві порівняно невеличкі хвилі і одну значно більшу за максимальним значенням і за періодом.

Для врахування довільного виду сезонності у моделі Prophet у вигляді ряду Фур'є слід задати такі головні параметри [12]:

– «period» (T) — період сезонності у днях;

– «fourier_order» (n) — порядок ряду Фур'є (натуральні числа: 1, 2, 3, …);

– «prior_scale» — показник, який відповідає за ступінь регуляризації, тобто на скільки буде допускатись можливий розкид значень навколо заданого часового ряду (значення від 0 до 1, більші значення зменшують похибку, але погіршують прогнозні можливості моделі (це явище у технологіях штучного інтелекту називається «overfitting» — з англ. «перенавчання»), як правило задається не більшим 0,3);

– інтервал часу $[t_0, t_1]$ (дати днів), в межах якого має місце ця сезонність (за замовчуванням — це весь період часу моделювання).

Між цими параметрами є очевидний зв'язок:

$$T = 2(t_1 - t_0). \tag{1}$$

Ще слід задавати режим «mode» (адитивний чи мультиплікативний), але, враховуючи суттєву нелінійність процесу майже в усіх країнах і різну висоту таких хвиль, встановлюємо режим як мультиплікативний, за замовчуванням.

Отже, поставлена задача зводиться до визначення параметрів T, n, t_0, t_1 для кожної хвилі за даними часового ряду на інтервалі $[\theta_0, \theta_1]$. Передусім, слід визначити параметри t_0, t_1. Пропонуємо чотири підходи для їх знаходження, основані на різних ситуаціях:

Підхід 1. Весь ряд спостережень — це одна велика хвиля, яка наростала весь час, але завершиться в наступний тиждень (якщо не завершиться, то вона просто не є об'єктом цієї статті, оскільки там ще не можна ідентифікувати жодної хвилі), тоді $t_0 = \theta_0$, $t_1 = \theta_1$.

Підхід 2. Хвилі є чітко вираженими, а значення між ними, а значення між є меншими від певної межі 1…5 % від максимального значення відповідної хвилі; тоді кожна хвиля ідентифікується по точках (датах), в яких значення стають меншими цієї межі.

Підхід 3. Хвилі переходять одна в іншу, але мінімальні значення між ними не перевищують 50 % від максимуму кожної з них, тоді, у першому наближенні, можна вважати, що нова хвиля починається у дату, коли попередня досягає свого максимуму і починає зменшення значень.

Підхід 4. Найзагальніший випадок. Розглянемо його детальніше.

Як відомо, ряд Фур'є має вигляд [8]—[11]

$$y(t) = \frac{a_0}{2} + \sum_{i=1}^{n} a_i \cos\left(\frac{2\pi}{T} it\right) + \sum_{i=1}^{n} b_i \sin\left(\frac{2\pi}{T} it\right), \tag{2}$$

де коефіцієнти Фур'є знаходяться за виразами [8]—[11]:

$$a_0 = \frac{1}{T}\int_0^T y(t)dt;\ a_i = \frac{2}{T}\int_0^T y(t)\cos\left(\frac{2\pi}{T} it\right)dt;\ b_i = \frac{2}{T}\int_0^T y(t)\sin\left(\frac{2\pi}{T} it\right)dt,\ i = \overline{1,n}, \tag{3}$$

де, коли ряд починається не з нуля, а з деякого часу t_0, то час t замінюється на $(t - t_0)$.

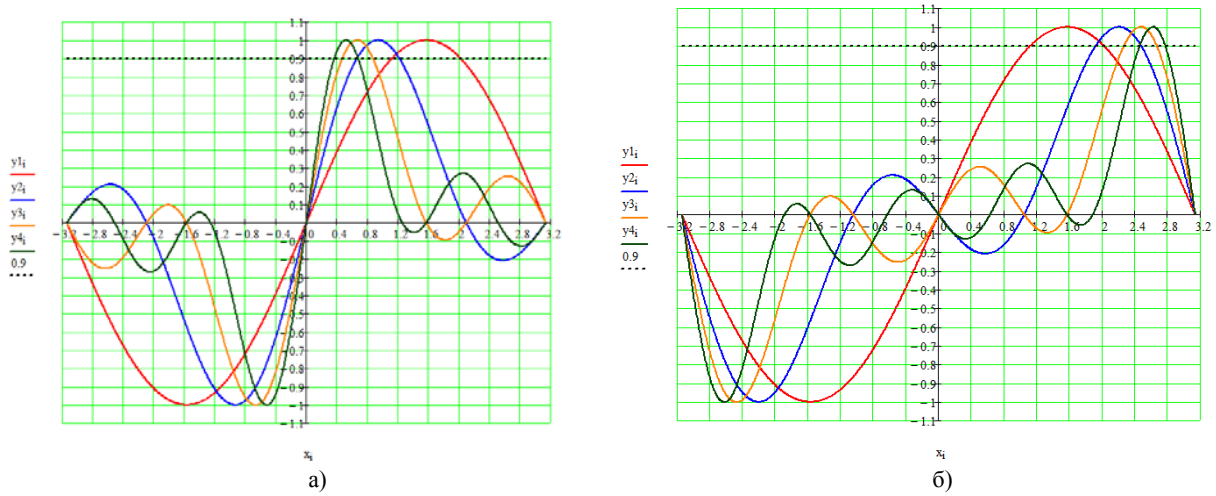
Як видно з виразів (2), (3), перші гармоніки мають найбільший період коливання T (коли $i = 1$), а усі інші — менший у 2, 3, … n разів.

Основна проблема застосування відомих методів для знаходження параметрів ряду Фур'є у вигляді (3) полягає в тому, що у нас немає окремих значень часового ряду, до яких можна застосувати ці методи. У нас є часовий ряд, в якому є і артефакти у вигляді впливу аномальних дат, і модуляція у вигляді тижневої сезонності, обумовленої графіком роботи лабораторій тестування, і складна природа самого процесу, через яку усі складові припасовуються мультиплікативно, а не адитивно. Тому ми пропонуємо певним шляхом оцінити можливі значення параметрів T і n, а потім запропонувати вбудованому алгоритму ідентифікації моделі Prophet вибрати з них найкращий варіант (якраз там і реалізовані оті відомі методи визначення параметрів ряду Фур'є). Інтервал часу теж можна оцінити приблизно, особливо, коли хвилі не мають чітких контурів і одна хвиля переходить в іншу, як це має місце в Україні.

Нагадуємо, що ряд Фур'є (рис. 2) є сукупністю так званих «гармонік» (синусоїд і косинусоїд певної амплітуди і частоти) [8]. Побудуємо графіки таких рядів Фур'є:

$$yn_i = \sum_{n} b_i \sin\left(nx_i\right), \tag{4}$$

де коефіцієнти b_i підбираються таким чином, щоб максимумом функції було значення 1. Графіки таких функцій для $n = 1, 2, 3$ показано на рис. 2.

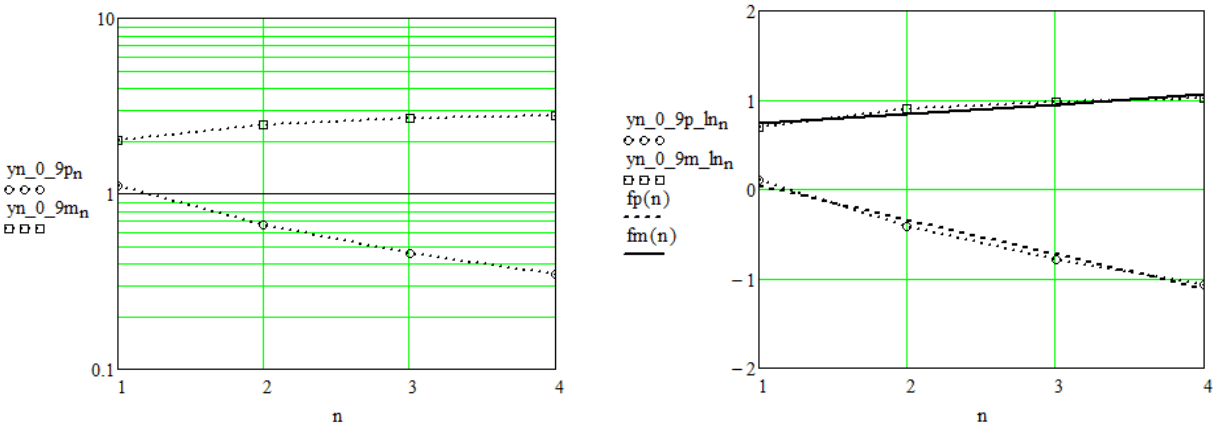


а) б)

Рис. 2. Сума різної кількості n гармонік (лише синусоїд виду (4)) ряду Фур'є yn від нормованого часу $t = 2\pi/T$: а — усі коефіцієнти Фур'є є додатними; б — коефіцієнти Фур'є мають множник $(-1)^{n-1}$

Як видно з рис. 2, зі збільшенням кількості гармонік n, пік (максимальне значення) ряду зсувається далі від піку ряду, побудованого за $n = 1$, тобто у вигляді однієї синусоїди. Причому, у випадку додатних коефіцієнтів пік зсувається ліворуч, а у випадку знакозмінних — праворуч.

Основним способом ідентифікації порядку n рядів Фур'є є перебір варіантів значень допоки похибка апроксимації буде зменшуватись. Але є й додаткова умова про те, що коефіцієнти Фур'є довільного сигналу спадають у порядку, обернено пропорційному своєму номеру i [9]. Однак, застосування цього методу, зазвичай основаного на застосуванні методу найменших квадратів для припасовування ряду Фур'є, дещо ускладнюється, по-перше, значною зашумленістю ряду, по-друге, нелінійним впливом інших складових (аномальних свят і псевдосвят, тижневої сезонності), а по-третє, що часто, уся крива відсутня і висновок про порядок ряду Фур'є слід зробити тільки за частиною хвилі (див. другу хвилю на рис. 1), тому пропонується інший підхід. Зазвичай, висновок про наявність хвилі робиться за наявності її піка і хоча б 10 % значень верхівки кривої (див. рис. 1). Проведемо на рис. 2 лінію на рівні 90 % від максимального значення (нагадуємо, що усі криві там нормовані і їх пік знаходиться на рівні 1). Один напівперіод кожної кривої перетинає цей рівень двічі, але, як видно з рис. 2а, відстань першої точки перетину кожною кривою цієї лінії до середини напівперіоду $\pi/2$ прямо пропорційна значенню n: чим більшим є це значення n, тим більшою є відстань. Для графіка на рис. 2б залежність така сама, але не до першої, а до другої точки перетину. Що цікаво, принаймні для перших трьох значень n ці залежності у логарифмічних координатах схожі на прямі (рис. 3). Отже, можна припустити, що за координатою однієї цієї точки можна оцінити порядок ряду Фур'є.



а) б)

Рис. 3. Залежність від порядку ряду Фур'є n значень нормованого часу $t = 2\pi/T$ для точок 90 % від максимального для рядів з усіма додатними коефіцієнтами (yn_0_9_p) і з від'ємними коефіцієнтами у парних гармоніках (yn_0_9_m): а — у логарифмічних координатах по осі абсцис; б — для логарифму від цих залежностей ln(yn_0_9_p), ln(yn_0_9_m) та їх апроксимація прямими fp і fm, відповідно

Апроксимація прямими залежностей на рис. 3 дала такий результат:

$$n_p = floor\left(\frac{0{,}424 - \ln(\text{yn\_0\_9\_p})}{0{,}385}\right); \quad n_m = floor\left(\frac{\ln(\text{yn\_0\_9\_m}) - 0{,}644}{0{,}104}\right), \tag{5}$$

де *floor*() — це функція округлення до цілого значення у менший бік (на Python).

Звичайно, формули (5) можна використовувати тільки у першому наближенні, тому що вони виведені, по-перше, тільки для $n = 1, 2, 3$, а по-друге, тільки для ряду у вигляді (4). За інших коефіцієнтів Фур'є вона буде іншою, але виведення її у загальнішому вигляді — це тема окремого дослідження.

Для оцінювання довжини і періоду кожної хвилі, тобто параметрів T, t_0, t_1, пропонуємо застосовувати алгоритм, показаний на рис. 4.

1. Згладжуємо дані і знаходимо дати піків (максимальних значень) усіх хвиль: спочатку знаходимо їх для згладженого ряду, а потім в їх околі уточнюємо дати для оригінального ряду. Як правило, варто задати якийсь мінімум (наприклад, 1 % чи 0,1 % від максимального значення найбільшої хвилі) для того, щоб відрізняти випадкові флуктуації від справжньої хвилі, неврахування якої спричинить в подальшому погіршення прогнозування.

2. Згладжуємо дані і знаходимо дати мінімальних значень усіх хвиль — початок і кінець кожної хвилі, пік якої було знайдено у п.1: спочатку знаходимо їх для згладженого ряду, а потім в їх околі уточнюємо дати для оригінального ряду. Важливо, знайти точку, де хвиля переходить у режим випадкових флуктуацій і втрачає схожість з коливальним процесом.

3. Уточнюємо можливі координати (дату дня) початку і кінця кожної хвилі, тобто кількість днів t_0 до дати, з якої кількість хворих у цій хвилі буде дорівнювати нулю. Аналіз даних показав, що у більшості країн у світі таке число знайти точно неможливо, оскільки дані весь час є більшими за нуль. Тоді пропонуємо оцінювати t_0, виходячи з синусоїдальної природи зміни значень, коли відстань від максимального значення до початку хвилі складає чверть періоду:

$$\frac{y_{\min 1}}{y_{\max}} = \sin\left(\left(t_{\min 1} - t_0\right) \frac{2\pi}{4\left(t_{\max} - t_0\right)} \right), \tag{6}$$

де $y_{\min 1}$ — мінімальне значення кількості нових хворих на початку хвилі у дату $t_{\min 1}$, знайдене у п. 2 алгоритму, $y_{\max}$ — максимальне значення кількості нових хворих на піку хвилі у дату $t_{\max}$, знайдене у п.1 алгоритму.

Нескладно показати, що:

$$t_0 = \frac{t_{\min} - a\, t_{\max}}{1 - a}, \quad a = \frac{2}{\pi} \arcsin\left(\frac{y_{\min}}{y_{\max}} \right). \tag{7}$$

Аналогічно знаходимо кінцеве значення хвилі, тобто кількість днів t_1 до дати, з якої кількість хворих у цій хвилі буде дорівнювати нулю після проходження піку:

$$\frac{y_{\min 1}}{y_{\max}} = \sin\left(\left(t_{\min 2} - \left(t_1 - \frac{T}{2} \right) \right) \frac{2\pi}{T} \right); \; T = 4\left(t_1 - t_{\max}\right), \tag{8}$$

де $y_{\min 2}$ — мінімальне значення кількості нових хворих в кінці хвилі у дату $t_{\min 2}$, знайдене у п.2 алгоритму.

Нескладно показати, що:

$$t_1 = \frac{t_{\min 2} - (a-2) t_{\max}}{1 - a}; \quad a = \frac{2}{\pi} \arcsin\left(\frac{y_{\min}}{y_{\max}} \right). \tag{9}$$

Як альтернативний варіант, значення дат [t_0, t_1] можна оцінювати і, виходячи, з обвідної у вигляді прямої лінії, що з'єднує точку мінімуму і максимуму хвиль, а не у вигляді синусоїди.

1. Знаходимо період хвилі за формулою (1).

2. Для діапазонів дат [t_0, t_1] усіх виявлених хвиль налаштовуємо окремі моделі Prophet з 7-денною сезонністю (для врахування графіка лабораторій), модулем врахування дат свят і псевдосвят, та додатковою сезонністю з ідентифікованим у пп. 1—4 початковим наближенням значень параметрами T і n, для яких задаємо діапазони змін значень. Вибираємо такі параметри, які задають мінімум критерію для валідаційного датасету.

Як критерій для валідаційного датасету може бути вибраний будь-який із загальноприйнятих для часових рядів, наприклад [12]: MAE — середня абсолютна похибка (з англ. «Mean Average Error») чи MSE — середньоквадратична похибка (з англ. «Mean Square Error»). Або можна брати, як у статті [1]: WAPE — сумарна відносна похибка за усі дати валідаційного датасету (з англ. «Weighted Mean Average Percentage Error»).

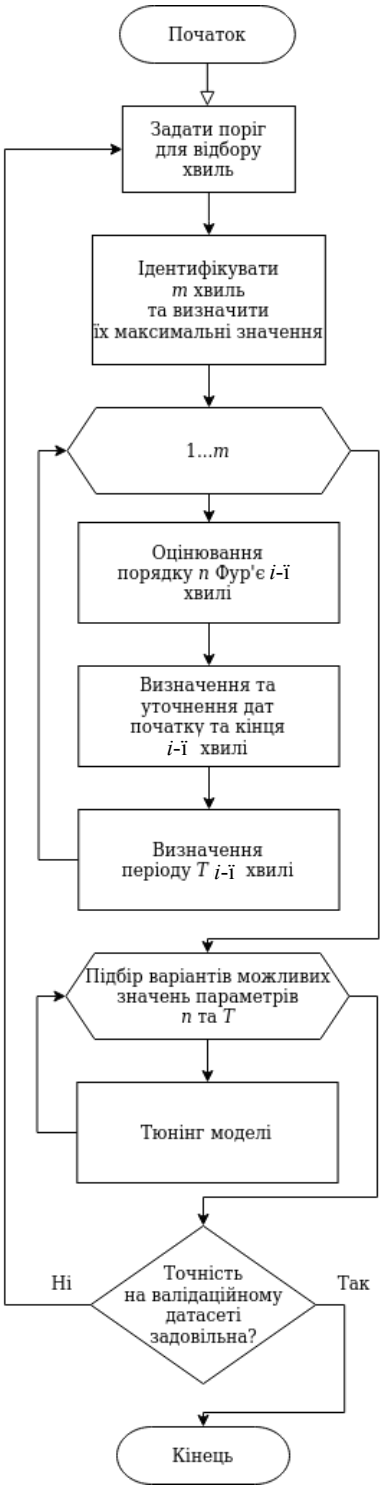


Рис. 4. Алгоритм удосконаленої інформаційної технології

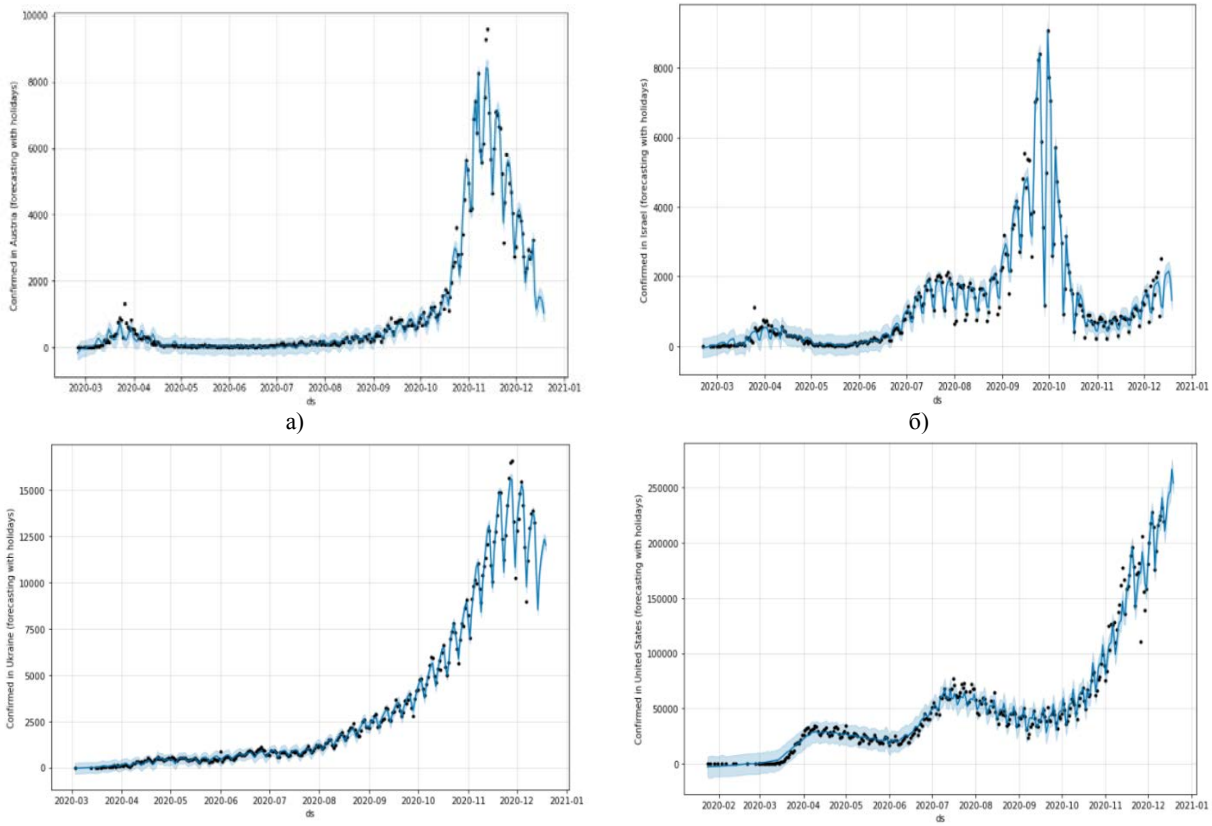
Валідаційний датасет вибирається в залежності від кінцевої мети дослідження. Якщо метою є дослідження закономірностей у регіоні протягом усього часу,

тоді валідаційний датасет слід формувати з випадково вибраних значень з усього ряду, а якщо метою є прогнозування, тоді варто брати останні декілька тижнів, як це робилось у статті [1]. Загалом-то, якщо формування ряду проходило у відносно схожих умовах, тоді модель, відібрана у перший спосіб, детальніше враховує особливості регіону і тому, теоретично, має кращі прогнозні можливості для довгострокового прогнозування. Модель же, сформована у другий спосіб — краща для короткострокового прогнозування на 1—3 тижнів вперед.

Варто зазначити, що, якщо основною метою є короткострокове прогнозування і остання зі спостережуваних хвиль є найбільшою, як це мало місце у середині грудня 2020 р. в Україні, тоді, у першому наближенні, можна весь ряд описувати моделлю для цієї найбільшої хвилі, але тоді порядок Фур'є n для неї має вибиратись таким чином, щоб період менших хвиль T_i хоча б приблизно був кратним періоду найбільшої хвилі $T_{\max}$: $T_i = T_{\max}/n$, $n = 1, 2, \ldots$. Як правило, варто брати $n \geq 3$.

Застосування удосконаленої інформаційної технології для країн світу

Запропонована удосконалена технологія застосована для низки країн світу (рис. 5) на основі даних згаданого вище датасету «COVID-19 Open Data». За даними робився прогноз на 2 тижні вперед, тому бралась до уваги відносна похибка тільки за останні 2 тижні. Як видно, має місце гарний збіг даних спостережень з модельними даними на кожній хвилі, що підтверджує ефективність запропонованої технології.



а) б)

г) д)

Рис. 5. Результат застосування удосконаленої інформаційної технології до країн світу з багатохвильовою динамікою кількості нових підтверджених випадків захворювань на коронавірус COVID-19:
а — Австрія; б — Ізраїль; в — Україна; г — США

Підходи щодо програвання сценаріїв розвитку захворюваності за допомогою удосконаленої інформаційної технології

Низка зроблених гіпотез та особливість вибору алгоритму ідентифікації параметрів моделі дещо ускладнює її застосування для задач прогнозування, адже саме такою є мета статті. Для усунення цієї проблеми пропонуємо брати до уваги не тільки варіант моделі з оптимальними параметрами на останні тижні спостережень, а й два граничні випадки: найпесимістичніший і найоптимістичніший сценарії. А тоді варто прогнозування робити не тільки з урахуванням довірчого інтервалу, а й у діа-

пазоні між песимістичним і оптимістичним варіантами. Пропонуємо 2 підходи до їх моделювання.

Підхід 1. Налаштування моделі під штучно змінені дані валідаційного датасету:

– песимістичний варіант — це коли дані на наступний тиждень будуть прогнозуватись як більші в *K* разів, ніж в останній тиждень;

– оптимістичний варіант — це коли дані на наступний тиждень будуть прогнозуватись як менші в *K* разів, ніж – в останній тиждень.

Для початку взяти, до прикладу, $K = 2$. Порахувати. У разі, якщо моделювання дасть малу похибку (менше 10 %), тоді *K* слід збільшити. Отриманий прогноз покаже які, теоретично можливі, варіанти може спрогнозувати модель за наявними даними. Далі варто застосувати цей підхід до тих дат, коли були сумніви в адекватності моделі (коли вона дала прогноз низької точності) і перевірити чи були взагалі можливості точнішого прогнозування, якщо брати один з таких сценаріїв прогнозування.

Значення *K* можна вибирати ще й на основі аналізу ряду спостережень — порівняти з періодичністю у 7 днів усі значення і знайти у скільки максимально разів вони відрізнялись протягом року. Хоча, в цьому випадку, можливі й аномалії, коли, наприклад, лікарі роблять переоблік і в один день дають результати тестувань за останні пару місяців (чи днів), як часто буває, на жаль, в деяких країнах, зокрема і в Україні, щодо летальних випадків. Насправді, й дані щодо нових хворих теж надходять з різними затримками по різних областях [7], тому висока точність прогнозу часто має дещо рандомізований характер.

Підхід 2. Імітація початку зростання (нової хвилі) та початку спаду (проходження піку чи його частина різкого зменшення значень) на основі ідентифікованих для певної країни параметрів хвиль:

– песимістичний варіант — імітується ситуація, що в наступний тиждень почнеться нова хвиля з типовими для певної країни параметрами і почнеться різке зростання значень;

– оптимістичний варіант — імітується ситуація, що в наступний тиждень почнеться різкий спад кількості нових хворих з типовими для певної країни параметрами і почнеться різке зменшення значень.

Саме для моделювання сценаріїв за другим підходом запропоноване удосконалення технології для ідентифікації типових особливостей багатохвильової динаміки процесу є найбільш ефективним.

Висновки

Удосконалено раніше розроблену авторами інформаційну технологію аналізу та прогнозування кількості нових підтверджених випадків захворювань на коронавірус COVID-19, спричиненою інфекцією SARS-CoV-2, на прикладі щодобових сумарних по Україні даних з урахуванням різних свят і псевдосвят, які можуть мати аномальний вплив. Створена раніше технологія працездатна лише для ділянки невпинного зростання значень однієї хвилі, а удосконалена — вже може застосовуватись для аналізу та прогнозування даних протягом усього ряду спостережень з початку минулого року, оскільки враховує багатохвильову природу цього явища. Запропоновано алгоритм ідентифікації параметрів кожної хвилі. Запропоновано декілька підходів та розроблено необхідні математичні співвідношення, які дозволяють у першому наближенні оцінити дату початку, завершення та період кожної хвилі, навіть у випадку, коли одна хвиля переходить в іншу.

Проведене дослідження показало, що дійсно можна запропонувати нові емпіричні співвідношення для оцінювання порядку ряду Фур'є для опису коливального процесу кожної хвилі лише за 10 % її значень у верхівці, оскільки найчастіше такі дані завжди доступні, інакше окремі хвилі не розпізнати. Співвідношення виведені окремо для випадку лише додатних коефіцієнтів, коли пік розташований ліворуч від середини напівперіоду, і окремо — для випадку знакозмінного ряду, коли пік розташований праворуч від неї. Однак, ці наближені оцінки рекомендовано уточнювати у певному діапазоні можливих значень, оскільки, в загальному випадку різних варіантів значень амплітуди гармонік запропоновані співвідношення можуть давати занижені оцінки.

Запропоновано застосовувати ідентифіковану за цією технологією модель для прогнозування найпесимістичнішого та найоптимістичнішого сценаріїв розвитку явища, тобто зміни кількості нових підтверджених випадків захворювань на коронавірус COVID-19 у заданій країні.

Створено програмне забезпечення на Python на базі платформи Kaggle, яке застосовано, як для України, так і ще для 69 країн світу. За допомогою ідентифікованих моделей отримано низку важливих висновків щодо розуміння закономірностей поширення коронавірусу, якщо вважати побудовані моделі достатньо адекватними:

1. Врахування багатохвильової природи процесу дозволило зменшити похибку до 3,5 %, але на-

явного досить короткого ряду спостережень не достатньо для перевірки адекватності побудованої моделі, а отже, її поки що не можна використовувати для довгострокового прогнозування.

2. Як показало проведене моделювання на базі дещо спрощеної моделі, має місце значний вплив свят та інших аномальних дат в деяких країнах світу: США, Литві, Іспанії, Індонезії, Австрії, де модель з урахуванням свят і псевдосвят має похибку майже у 1,5—2 рази меншу, ніж модель без урахування таких аномальних дат, і при цьому така похибка становить в останній тиждень менше від 5…20 % (для США — 5 %). В Україні ж спрощена модель з урахуванням аномальних дат дає похибку 8,74 % (ефективніша модель — 3,95 %), а без їх урахування — 34,07 %.

3. Модель для України з точністю 3,95 % отримано з урахуванням впливу карантину вихідного дня (14, 15, 21, 22, 28, 29 листопада) зі зсувом на 7 днів для врахування часу до появи симптомів, затримки на тестування й оприлюднення їх результатів; ця ж модель без урахування карантину вихідного дня дає точність лише 4,34 %, що свідчить про те, що карантин вихідного дня дійсно впливає на зменшення кількості захворювань.

4. Стрімке зниження прогнозованих значень кількості нових захворювань, на жаль, не дає впевненості у тому, що така динаміка збережеться ще декілька тижнів, оскільки ряд спостережень є надто малим для таких прогнозів, тому рекомендується використовувати зроблений прогноз обережно.

Той факт, що в моделях не здійснювався повний перебір усіх можливих значень параметрів, не враховувалась явно динаміка інших факторів (наприклад, зміна щодобової кількості тестувань чи кількості ліжкомісць), на жаль, не дає впевненості в тому, що ці моделі можна використовувати для довгострокового прогнозування та в тому, що отримані результати дають остаточні відповіді на поставлені питання у довгостроковій перспективі.

Результати роботи передано в Робочу групу з математичного моделювання проблем, пов’язаних з епідемією коронавірусу SARS-CoV-2 в Україні, членом якої є один зі співавторів статті — д-р техн. наук, професор В. Б. Мокін. Ця група готує аналітику загальнодержавного рівня та передає її в РНБО, Кабінет Міністрів України та ін. Усі звіти цієї Робочої групи публікуються на її сторінці на сайті Національної академії наук України: http://www.nas.gov.ua/UA/Activity/covid/Pages/wg.aspx .

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] В. Б. Мокін, А. В. Лосенко, і А. Р. Ящолт, «Інформаційна технологія аналізу та прогнозування кількості нових випадків захворювань на коронавірус SARS-COV-2 в Україні на основі моделі Prophet», *Вісник Вінницького політехнічного інституту*, № 5, с. 71-83, 2020.

[2] Tommaso Banelli,·and Marco Vuano, Federico Fogolari, Andrea Fusiello, Gennaro Esposito, and Alessandra Corazza, “Automation of peak-tracking analysis of stepwise perturbed NMR spectra,” *Journal of Biomolecular NMR*, vol. 67, pp. 121-134, 2017.

[3] C. Peng, S. W. Unger, F. V. Filipp, M. Sattler, and S. Szalma, “Automated evaluation of chemical shift perturbation spectra: New approaches to quantitative analysis of receptor-ligand interaction NMR spectra,” *J. Biomol NMR*, vol. 29, pp. 491-504, 2004.

[4] Peipei Wanga, and Xinqi Zheng, “Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics,” *Chaos, Solitons & Fractals*, vol. 139, October 2020. https://doi.org/10.1016/j.chaos.2020.110058 .

[5] M. Indhuja and P. P. Sindhuja, “Prediction of covid-19 cases in India using prophet,” *International Journal of Statistics and Applied Mathematics*, no. 5 (4), pp.103-106, 2020.

[6] Dr. Shikha Gaur, “Global Forecasting of COVID-19 Using Arima Based FB-PROPHET,” *International Journal of Engineering Applied Sciences and Technology*, vol. 5, issue 2, pp. 463-467, 2020.

[7] 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: грудень 14, 2020.

[8] Б. І. Мокін, В. Б. Мокін, і О. Б. Мокін, *Практикум для самостійної роботи студентів з навчальної дисципліни «Методологія та організація наукових досліджень». Частина 1: від постановки задачі до синтезу та ідентифікації математичної моделі*. Вінниця, Україна: ВНТУ, 2018, 179 с.

[9] Б. І. Мокін, В. Б. Мокін, і О. Б. Мокін, *Математичні методи ідентифікації динамічних систем*, навч. посіб. Вінниця, Україна: ВНТУ, 2010, 260 с.

[10] В. М. Дубовой, Р. Н. Квєтний, О. І. Михальов, і А. В. Усов, *Моделювання та оптимізація систем*, підруч. Вінниця Україна: ПП «ТД «Едельвейс», 2017, 804 с.

[11] В. П. Легеза, *Математичний аналіз*, підруч. 4-х томах. Т. 1. Київ, Україна: Політехніка, 2019, 336 с.

[12] Saturating Forecasts Forecasting Growth [Електронний ресурс]. Режим доступу:
https://facebook.github.io/prophet/docs/saturating_forecasts.html#forecasting-growth . Дата звернення: грудень 14, 2020.

Рекомендовано кафедрою системного аналізу та інформаційних технологій ВНТУ

Стаття надійшла до редакції 18.12.2020

Мокін Віталій Борисович — д-р техн. наук, професор, завідувач кафедри системного аналізу та інформаційних технологій, e-mail: vbmokin@gmail.com ;

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

Яцолт Андрій Русланович — канд. техн. наук, доцент, доцент кафедри системного аналізу та інформаційних технологій, e-mail: yasholt@gmail.com .

V. B. Mokin[1]
A. V. Losenko[1]
A. R. Yascholt[1]

Information Technology Analysis and Predicting a Multiwave Number of New COVID-19 Disease Based on Prophet Model

[1]Vinnytsia National Technical University

The article is devoted to the improvement of the information technology previously developed by the authors for the analysis and forecasting of the number of new confirmed cases of the disease for the coronavirus COVID-19 caused by the SARS-CoV-2 infection, using the example of the daily total data of the current "wave" in Ukraine, taking into account various holidays and pseudo-holydays, which may have an abnormal effect. The previously created technology was operable only for the area of continuous growth of the values of one wave, and the improved one can already be used to analyze and predict data throughout the entire period, since it takes into account the multi-wave nature of this phenomenon. An algorithm for identifying the parameters of each wave is proposed. A number of mathematical relationships have been developed that allow, in a first approximation, to estimate the start, end and period of each wave, even in the case when one wave passes into another.

A new empirical relationships is proposed to estimate the order of the Fourier series for describing the time process of each wave for only 10 % of its values at the top, since, as a rule, such data are available in an explicit form, otherwise the data will not be recognized as a separate wave. The ratios are derived separately for the case of only positive coefficients, when the peak is located to the left of the middle of the half-period, and separately — for the case of an alternating series, when the peak is located to the right of it. However, these approximate estimates are recommended to be refined within a certain range of possible values, since in the general case of different variants of the harmonic amplitude values, the proposed ratios can give underestimates.

It is proposed to apply the model identified by this technology to predict the most pessimistic and most optimistic scenarios for the development of the phenomenon, that is, changes in the number of new confirmed cases of the disease for the coronavirus COVID-19 in the future in a given country.

Python software was created based on the Kaggle platform, which is used both for Ukraine and for 69 other countries. Using the identified models, a number of important conclusions were obtained regarding understanding the patterns of the spread of coronavirus both in Ukraine and in other 69 countries of the world. The results were submitted to the Working Group on Mathematical Modeling of Problems Associated with the SARS-CoV-2 Coronavirus Epidemic in Ukraine.

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

Mokin Vitalii B. — Dr. Sc. (Eng.), Professor, Head of the Chair of System Analysis and Information Technologies, e-mail: vbmokin@gmail.com ;

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

Yascholt Andriy R. — Cand. Sc. (Eng.), Associate Professor, Associate Professor of the Chair of System Analysis and Information Technologies, e-mail: yasholt@gmail.com

В. Б. Мокин[1]
А. В. Лосенко[1]
А. Р. Ящолт[1]

Информационная технология анализа и прогнозирования многоволнового количества новых случаев заболеваний коронавирусом COVID-19 на основе модели Prophet

[1]Винницкий национальный технический университет

Усовершенствована ранее разработанная авторами информационная технология анализа и прогнозирования количества новых подтвержденных случаев заболеваний коронавирусом COVID-19, вызванных инфекцией SARS-CoV-2, на примере ежесуточных суммарных по Украине данных текущей «волны» с учетом различных праздников и псевдопраздников, которые могут иметь аномальное влияние. Созданная ранее технология была работоспособной только для участка непрерывного роста значений одной волны, а усовершенствованная уже может использоваться для анализа и прогнозирования данных в течение всего периода, поскольку учитывает многоволновую природу этого явления. Предложен алгоритм идентификации параметров каждой волны. Разработан ряд математических соотношений, которые позволяют в первом приближении оценить дату начала, завершения и период каждой волны, даже в случае, когда одна волна переходит в другую.

Предложены новые эмпирические соотношения для оценки порядка ряда Фурье для описания колебательного процесса каждой волны лишь по 10 % ее значений в вершине, поскольку, как правило, такие данные есть в явном виде, иначе данные не будут распознаны как отдельная волна. Соотношения выведены отдельно для случая только положительных коэффициентов, когда пик расположен слева от середины полупериода, и отдельно — для случая знакопеременного ряда, когда пик расположен справа от нее. Однако, эти приблизительные оценки рекомендовано уточнять в определенном диапазоне возможных значений, поскольку в общем случае различных вариантов значений амплитуды гармоник предложенные соотношения могут давать заниженные оценки.

Предложено применить идентифицированную по этой технологии модель для прогнозирования наиболее пессимистичного и оптимистичного сценариев развития явления, то есть изменения количества новых подтвержденных случаев заболеваний коронавирусом COVID-19 в будущем в заданной стране.

Создано программное обеспечение на Python на базе платформы Kaggle, которое применено как для Украины, так и еще для 69 стран мира. С помощью идентифицированных моделей получен ряд важных выводов относительно понимания закономерностей распространения коронавируса как в Украине, так и в других 69 странах мира. Результаты переданы в Рабочую группу по математическому моделированию проблем, связанных с эпидемией коронавируса SARS-CoV-2 в Украине.

Ключевые слова: информационные технологии, COVID-19, прогнозирование временных рядов, Prophet, ряд Фурье, искусственный интеллект, прогнозирование сценариев развития

Мокин Виталий Борисович — д-р техн. наук, профессор, заведующий кафедрой системного анализа и информационных технологий, e-mail: vbmokin@gmail.com ;

Лосенко Арсен Владимирович — аспирант кафедры системного анализа и информационных технологий, e-mail: arsenlosenko@gmail.com ;

Ящолт Андрей Русланович — канд. техн. наук, доцент, доцент кафедры системного анализа и информационных технологий, e-mail: yasholt@gmail.com