https://doi.org/10.31649/1997-9266-2020-152-5-71-83

УДК 004.9+578.834.1

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

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ АНАЛІЗУ ТА ПРОГНОЗУВАННЯ КІЛЬКОСТІ НОВИХ ВИПАДКІВ ЗАХВОРЮВАНЬ НА КОРОНАВІРУС SARS-COV-2 В УКРАЇНІ НА ОСНОВІ МОДЕЛІ PROPHET

[1]Вінницький національний технічний університет

Розроблено інформаційну технологію аналізу та прогнозування кількості нових підтверджених випадків захворювань на коронавірус «COVID-19», викликаних інфекцією SARS-CoV-2, на прикладі щодобових сумарних по Україні даних поточної «хвилі» з урахуванням різних свят і псевдосвят, які можуть мати аномальний вплив. Проведено огляд відомих моделей для врахування таких аномалій та обґрунтовано, що за сучасних коротких рядів даних спостережень та інших умов оптимальною для розв'язання цієї задачі є модель Facebook Prophet. Охарактеризовано наявні дані щодо можливих часових аномалій по Україні у відомому датасеті Google-платформи «COVID-19 Open Data» та запропоновано яким чином можна адаптивно враховувати такі аномалії, як: державні свята, дати, коли за даними NOAA було дуже тепло і не було опадів та дати послаблення карантину за інформацією з «Oxford COVID-19 government response tracker». Розроблено алгоритм застосування запропонованої інформаційної технології з двоетапною ідентифікацією параметрів та окремим валідаційним датасетом для ідентифікації оптимальної структури моделі на кожному з цих етапів. Створено програмне забезпечення на Python на базі платформи Kaggle, яке застосовано, як для України, так і ще для 69 країн світу. Для прискорення роботи, по-перше, розроблено спрощену версію моделі лише з одним етапом її ідентифікації, а по-друге, створено новий датасет «COVID-19: Holidays of countries» з інформацією про свята 70 країн світу, адаптований до потреб цієї технології та розміщений у Kaggle у форматі відкритих даних. За допомогою ідентифікованих моделей отримано низку важливих висновків щодо розуміння закономірностей поширення коронавірусу як в Україні, так і в інших 69 країнах світу. Побудовано модель, яка забезпечує моделювання кількості нових підтверджених випадків захворювань на коронавірус в Україні на 2 тижні вперед з похибкою 2,2 % та зроблено прогноз на наступні 2 тижні, який передано у Робочу групу з математичного моделювання проблем, пов'язаних з епідемією коронавірусу SARS-CoV-2 в Україні.

Ключові слова: інформаційна технологія, SARS-CoV-2, COVID-19, прогнозування часових рядів, Prophet, штучний інтелект.

Вступ

Пандемія коронавірусу COVID-19, викликана інфекцією SARS-CoV-2 (далі — «коронавірусу»), у світі стала серйозним випробуванням для людства. Люди, що приймають рішення, намагаються вживати заходи, спрямовані на боротьбу з факторами, що сприяють поширенню захворювання, тому що боротись з наслідками захворювань набагато складніше, ніж з причинами. За цих умов, важлива роль відводиться аналітикам, які повинні навчитись моделювати процес поширення, зараження та іншого впливу на людей цієї хвороби. Задача суттєво ускладнюється мінливими умовами (рішення уряду про зміни карантинних обмежень, різні погодні умови, різною кількістю щоденних тестувань, неоднорідністю поширення у просторі, людським фактором та ін.), малим розміром вибірки, великою кількістю впливових факторів, які важко піддаються вимірюванню тощо.

Основне завдання — це навчитись моделювати і прогнозувати щоденну кількість нових підтверджених випадків хвороби (за тестами, основаними на полімеразній ланцюговій реакції (ПЛР)), оскільки решта показників (кількість госпіталізованих, кількість тих, хто видужав, кількість смертельних випадків, кількість позитивних експрес-тестів та ін.), вже є наслідками розвитку хвороби

у тих, в кого вона підтвердилась. Другим важливим завданням є аналіз ідентифікованої математичної моделі та виявлення закономірностей зміни цієї кількості у кожній країні, які допомогли б краще зрозуміти ситуацію і вчасно вжити необхідних заходів з її покращення.

Вчені всіх країн у 2020 році зосередились на вирішенні таких проблем. Так за даними Google Scholar у 2020 році вже є більше 100 тисяч статей, що містять слово «COVID-19». Найпопулярнішими у світі є такі моделі:

– 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 [1], [2]), основана на системі диференціальних рівнянь, яка якраз використовується і в Україні на офіційному рівні для формування прогнозів і прийняття на їх основі, управлінських рішень в РНБО (Ради з національної безпеки та оборони), Міністерстві охорони здоров'я (МОЗ) України та ін. Її основні положення описані у статті [1], але ця модель дуже чутлива до якості вхідних даних, тому, як видно з «Прогноз РГ-27» на 6-13.11.2020 р. [3] та «Прогноз РГ-28» на 14-20.11.2020 р. [4], відносна похибка середнього значення прогнозних даних відносно визначених пізніше склала 6,17 % та 7,14 %, відповідно;

– різноманітні регресійні моделі на основі дерев рішень, нейронних мереж та ін. Багато моделей опубліковано на базі платформи Kaggle у відповідних датасетах зі словом «COVID-19» у назві, а також, вони є у змаганнях Kaggle, перелік яких можна знайти у [5], але досвід показує, що для ефективної роботи таких моделей замало даних, як правило, вони або недонавчаються, або, що частіше, перенавчаються;

– моделі часових рядів, які найкраще підходять для моделювання окремого показника у часі, в першу чергу, ARIMA та Prophet [6]—[9]. Саме ці моделі здаються перспективнішими для моделювання щоденної кількості нових хворих на певному етапі, без урахування інших факторів та за умов, поки не буде накопичена достатня статистика для запуску складніших моделей штучного інтелекту на основі ансамблів дерев рішень, нейронних мереж та ін.

Варто зазначити, що більшість аналітиків прогнозує накопичувальну криву, тобто сумарне за усі попередні дати значення [6], [7]. Такий прийом дозволяє штучно зменшити похибку, адже вона ділиться на значно більші числа, ніж у випадку, коли брати тільки щоденні прирости. Але прийняття рішень про карантинні обмеження робиться виключно за щоденними приростами, тобто скільки випадків є лише за одну добу, отже, ціннішим є прогнозування саме їх. Так само, через це, не варто використовувати інший популярний прийом, коли аналізують і прогнозують ковзне середнє показників, взяте з семиденним вікном, що теж робить дані суттєво відмінними від вихідних.

Моделі часових рядів для кожної країни і різних етапів перебігу пандемії теж відрізняються. Наприклад, для моделювання хвилі, пік якої пройдений, оптимально використовувати логістичну модель. Якщо ж пік ще не пройдений і аналогічної хвилі у країні ще не спостерігалось, тоді варто використовувати лінійну регресію з різними складовими для опису сезонності та ін. Відповідно до цього у вітчизняних реаліях варто будувати окремі моделі для кожної з хвиль. В першу чергу, найбільш інтерес викликає хвиля, яка має місце в Україні у теперішній час — період невпинного зростання з семиденним періодичним інтервалом, що має місце з 6 липня 2020 р. (рис. 1).

У цій статті спробуємо побудувати адекватну модель та, за її допомогою, відповісти на такі питання щодо даних по щоденній кількості нових хворих на коронавірус за період 6.07—22.11.2020 р.:

1) чи має зростання лінійний чи нелінійний характер?

2) чи мають місце значний вплив з певним зсувом в часі аномальні дати (державні свята та псевдосвята на кшталт дат послаблення карантину, теплих днів без опадів тощо)?

3) чи має місце вплив інших видів сезонності, окрім тижневої, і, якщо має місце, то з якою періодичністю та лінійно чи нелінійно?

Проведений аналіз показав, що для моделювання періодичних часових рядів з урахуванням аномальних дат, оптимальним є застосування моделі Facebook Prophet (далі просто «Prophet»). Його основні переваги детально охарактеризовані у роботі [9].

Тому важливо побудувати модель даних по Україні для моделювання щоденного приросту кількості нових хворих на коронавірус, з урахуванням можливих часових аномалій, налаштувати її оптимальні параметри за даними періоду 6.07—22.11.2020 р., провести за нею прогнозування, наприклад на наступні 2 тижні, провести за нею аналіз закономірностей динаміки поширення цієї хвороби в Україні та випробувати її на інших країнах і зробити порівняння.

Мета дослідження — розробити інформаційну технологію аналізу та прогнозування кількості нових підтверджених випадків хвороби на коронавірус «COVID-19», викликаної інфекцією

SARS-CoV-2, на основі моделі Prophet, з урахуванням різних свят і псевдосвят, які можуть мати аномальний вплив у певному регіоні.

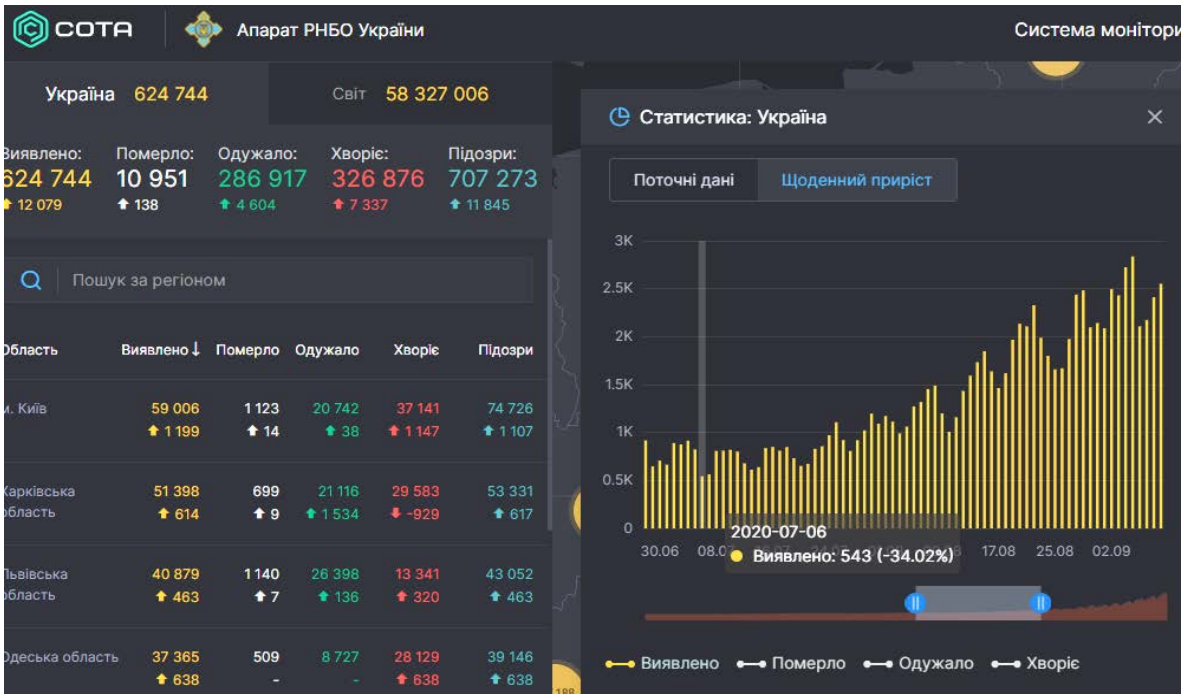


Рис. 1. Графік поточної «хвилі» щоденного приросту кількості нових хворих на веб-порталі РНБО України Системи моніторингу поширення епідемії коронавірусу (https://covid19.rnbo.gov.ua/)

Аналіз вхідних даних моделі

Для моделювання використовувались відкриті офіційні дані 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 варто використовувати відомий датасет Google-платформи «COVID-19 Open Data» (https://github.com/GoogleCloudPlatform/covid-19-open-data), де доступно багато статичної і динамічної, разом з тим щоденної, інформації (показники захворюваності на коронавірус, дати карантинних обмежень, мобільність населення за даними Google, погодні умови тощо) по більшості країн світу, але, на жаль, не вся інформація там є достовірною. Аналіз показав, що там мають місце помилки по Україні, тому у дослідженні використовували не ці дані, а дані РНБО України.

Згідно з поставленою задачею варто дослідити гіпотезу того, що свята можуть впливати на динаміку захворюваності, але із запізненням. Люди, особливо у карантинних умовах, святкують та контактують під час свят чи за інших умов, що збільшує ризик поширення вірусу. Проведений авторами кореляційний аналіз даних показав, що доцільно брати запізнення на 7 днів, але з адаптивним вікном (вікно [–1, 1] означає, що аномальними вважаються дати зі зсувом у 6…8 днів).

На етапі ідентифікації (тренування) моделі варто оптимізувати розмір адаптивного вікна навколо зсуву у 7 днів, наприклад, перебирати усі варіанти навколо нього в діапазоні від 4 до 10 днів (прирости [–3, –2, –1, 0, 1, 2, 3]) і вибирати варіант з найменшою похибкою.

Авторами досліджувався можливий вплив таких аномальних дат (свят і псевдосвят) [10]:

1. Державні свята (за даними пакету Holidays: https://github.com/dr-prodigy/python-holidays) із семиденним зсувом вперед (параметр “ds”) та адаптивним вікном (у першому наближенні беремо вікно [–3, 3], тобто від 4 до 10 днів) (рис. 2).

	ds_holidays	holiday	ds
0	2020-03-08	Міжнародний жіночий день	2020-03-15
1	2020-04-19	Пасха (Великдень)	2020-04-26
2	2020-06-07	Трійця	2020-06-14
3	2020-05-01	День праці	2020-05-08
4	2020-05-09	День перемоги	2020-05-16
5	2020-06-28	День Конституції України	2020-07-05
6	2020-08-24	День незалежності України	2020-08-31
7	2020-10-14	День захисника України	2020-10-21
8	2020-12-25	Різдво Христове (католицьке)	2021-01-01

Рис. 2. Державні свята України (за даними пакета Holidays: https://github.com/dr-prodigy/python-holidays), дата «ds» — це справжня дата, посунута на 7 днів вперед

2. Дати, коли одночасно було дуже тепло і без опадів, коли стрімко збільшувалась кількість людей у місцях відпочинку, назвемо їх «метеопаттерни» (за даними датасету «COVID-19 Open Data», зокрема, за даними National Oceanic and Atmospheric Administration (NOAA) — Національного управління з питань океану та атмосфери США (https://www.ncei.noaa.gov/)). Пропонуємо відбирати дати зі зсувом 7 днів та адаптивним вікном (рис. 3), коли кількість опадів була нульова, а середньодобова температура була більша за квантиль P95, тобто значення, вищі цього, мали місце тільки у 5 % випадків.

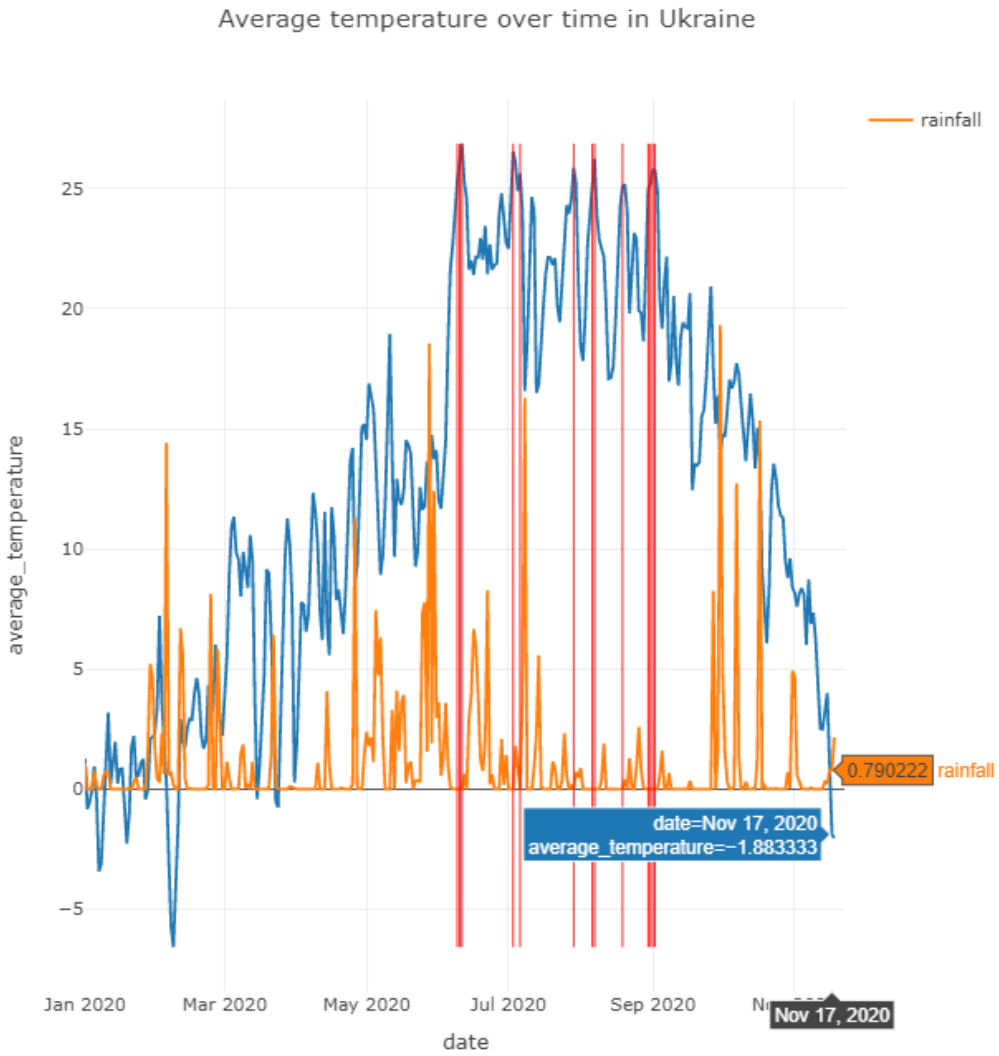


Рис. 3. Середньодобова температура в Україні (помаранчева крива), кількість опадів, мм (синя крива), дати аномальних «метопаттернів» (червоні вертикальні лінії)

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 згаданому вище датасеті «COVID-19 Open Data»), котрий відображає усі послаблення карантину, згідно з рішеннями уряду України, за 17 критеріями. Дати, коли ця сума зменшувалась, формалізовано як дати послаблення карантину (рис. 4).

4. Дні свят без зсуву на 7 днів вперед і нульовим вікном, для врахування аномально малої кількості тестувань на свята, що, на жаль, має місце в Україні.

Усі відібрані свята (вже з урахуванням зсуву у 7 днів, де він був зроблений) показані на рис. 5 вертикальними лініями.

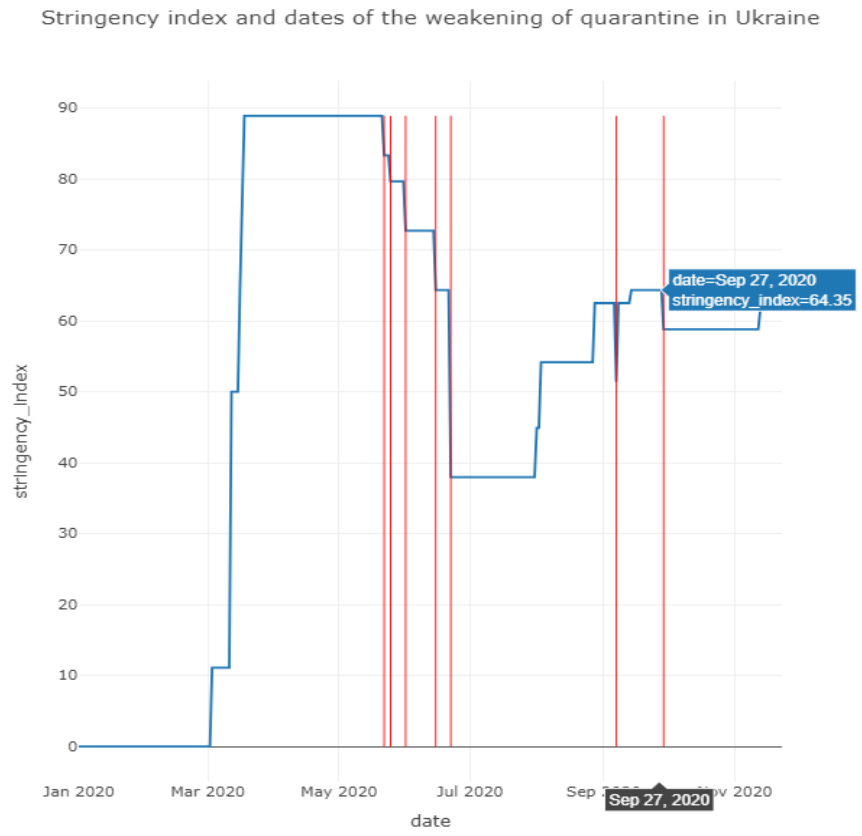


Рис. 4. Сумарний за добу Stringency-індекс України з ослабленнями та підсиленнями карантину (синя лінія) та дати аномалій-послаблень карантину (червона лінія)

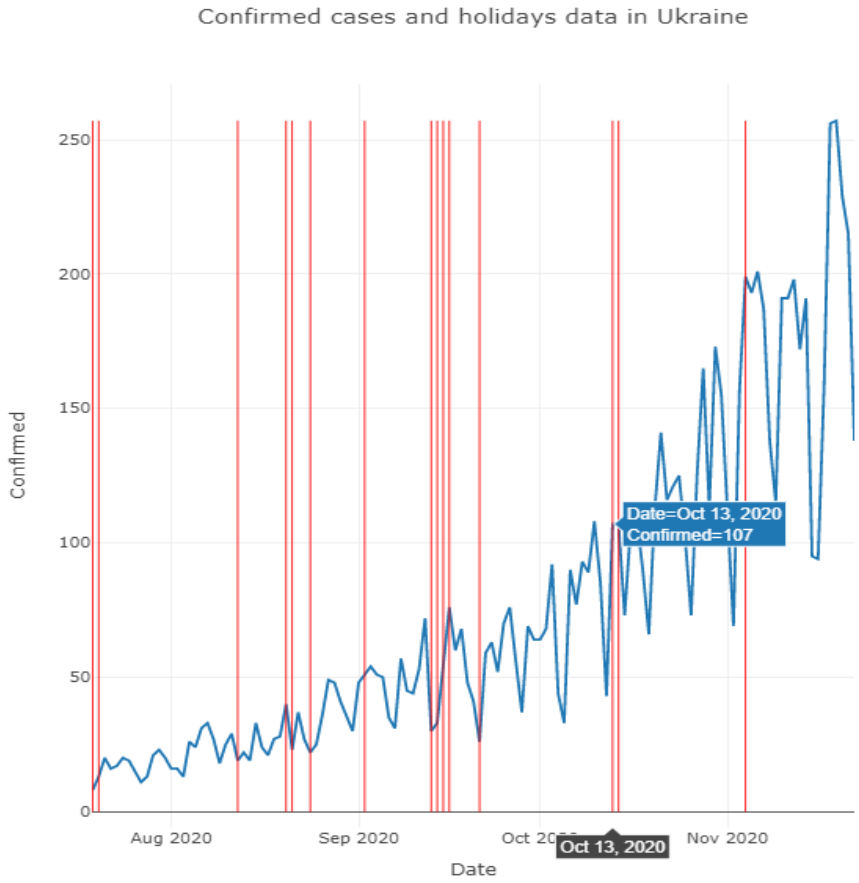


Рис. 5. Щоденні значення кількості нових випадків захворювань на коронавірус в Україні (синя крива) та дати аномалій (червоні вертикальні лінії)

Розроблення алгоритму інформаційної технології

Аналіз показав, що, параметрів Prophet, які задаються за замовчуванням, недостатньо для побудови достатньо точної моделі — часто саме цю помилку роблять інші дослідники [6]—[8]. Тому рекомендується, враховуючи складний нелінійний характер зміни даних, здійснювати налаштування таких її параметрів [10]:

– 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 (семиденної) сезонності;

– мультиплікативність чи адитивність урахування, ступінь регуляризації та кількість коефіцієнтів ряду Фур'є для опису іншої сезонності з періодом у n днів.

Розвідувальний аналіз, і не тільки для України, показав, що, окрім традиційної для цієї задачі тижневої сезонності, що диктує режим роботи лабораторій з тестування на коронавірус, часто має місце й інша специфіка динаміки усередині тижня, тобто сезонність з меншим періодом: n = 2, 3, .. 6 днів. Рекомендується пробувати ідентифікувати сезонність з парною і непарною довжиною періоду (наприклад, n = 3, 4 днів) і тоді по їх вигляду можна буде оцінити яка ж сезонність має місце насправді. Моделювання з урахуванням тижневої сезонності і сезонності з n = 3, 4 днів показало, що оптимальним є значення n = 4 дні.

Як обґрунтовано вище, для побудови моделі «хвилі», пік якої ще не досягнуто, варто використовувати лінійну регресію Prophet. Така регресія традиційно будується у вигляді кусково-лінійної апроксимації тренду між точками суттєвої зміни значень (за замовчуванням береться 30 таких точок, але модель Prophet їх кількість адаптивно оптимізує — може й усі видалити) [9].

Пропонується для моделі, призначеної для прогнозування даних на N днів у майбутнє, не менше, ніж N даних виділяти на валідаційну вибірку, щоб уникнути перенавчання моделі. Нагадуємо, що під час побудови моделей штучного інтелекту, зокрема моделей часових рядів, прийнято наявні дані розділяти на тренувальну вибірку, на якій налаштовуються параметри, і валідаційну вибірку, на якій перевіряється яка з налаштованих моделей краща. В нашому випадку, пропонуємо до валідаційної вибірки відносити дані за N останніх днів, а до тренувальної — за усі попередні дати.

Як метрику (критерій оптимальності), пропонуємо брати найменшу сумарну відносну похибку, яку ще називають метрикою WAPE (Weighted Mean Average Percentage Error), на усіх датах валідаційної вибірки.

Перебір усіх комбінацій таких параметрів — це довготривала NP-задача. Тому, для прискорення роботи програми пропонується задавати обмежену кількість варіантів можливих значень. Наприклад, значення кожного параметра вибирати тільки з 4-х варіантів, ще й у два етапи: спочатку — параметри аномальних дат, а потім з оптимальними значеннями параметрів першого етапу робити оптимізацію решти параметрів на другому етапі. Крім того, є сенс параметри регуляризації різних складових задавати взаємозалежно. Наприклад, змінювати один з них (наприклад, регуляризацію врахування свят), а інші (різні види сезонності) — обчислювати шляхом поділу його на певний коефіцієнт.

Для розв'язання поставленої задачі розроблено алгоритм інформаційної технології з опрацювання вхідних даних та налаштування і застосування моделі, який містить такі етапи (рис. 6):

Eman 1. Збирання та формалізація відкритих даних. Цей етап детально охарактеризовано вище.

Eman 2. Перший етап побудови моделі та оптимізація її параметрів, зокрема — перебір варіантів значень ширини адаптивного вікна в діапазоні [–3, …, 3] (усі цілі числа і нуль), сили впливу (масштабу) свят і псевдосвят та варіантів режиму урахування свят (мультиплікативний чи адитивний).

Eman 3. Другий етап побудови моделі та оптимізація її параметрів з вже оптимізованими на першому етапі параметрами, зокрема перебір значень таких параметрів (усі види сезонності, як зазначено вище, враховуються мультиплікативно): варіанти режиму врахування (мультиплікативний чи адитивний), ступінь регуляризації та кількість коефіцієнтів ряду Фур'є для опису тижневої (семиденного) та, окремо, чотириденного видів сезонності. Для спрощення можна вибирати однаковий варіант режиму врахування обох видів сезонності.

Eman 4. Аналіз виявлених закономірностей по структурі ідентифікованої оптимальної моделі.

Eman 5. Формування прогнозів на задану кількість N днів вперед.

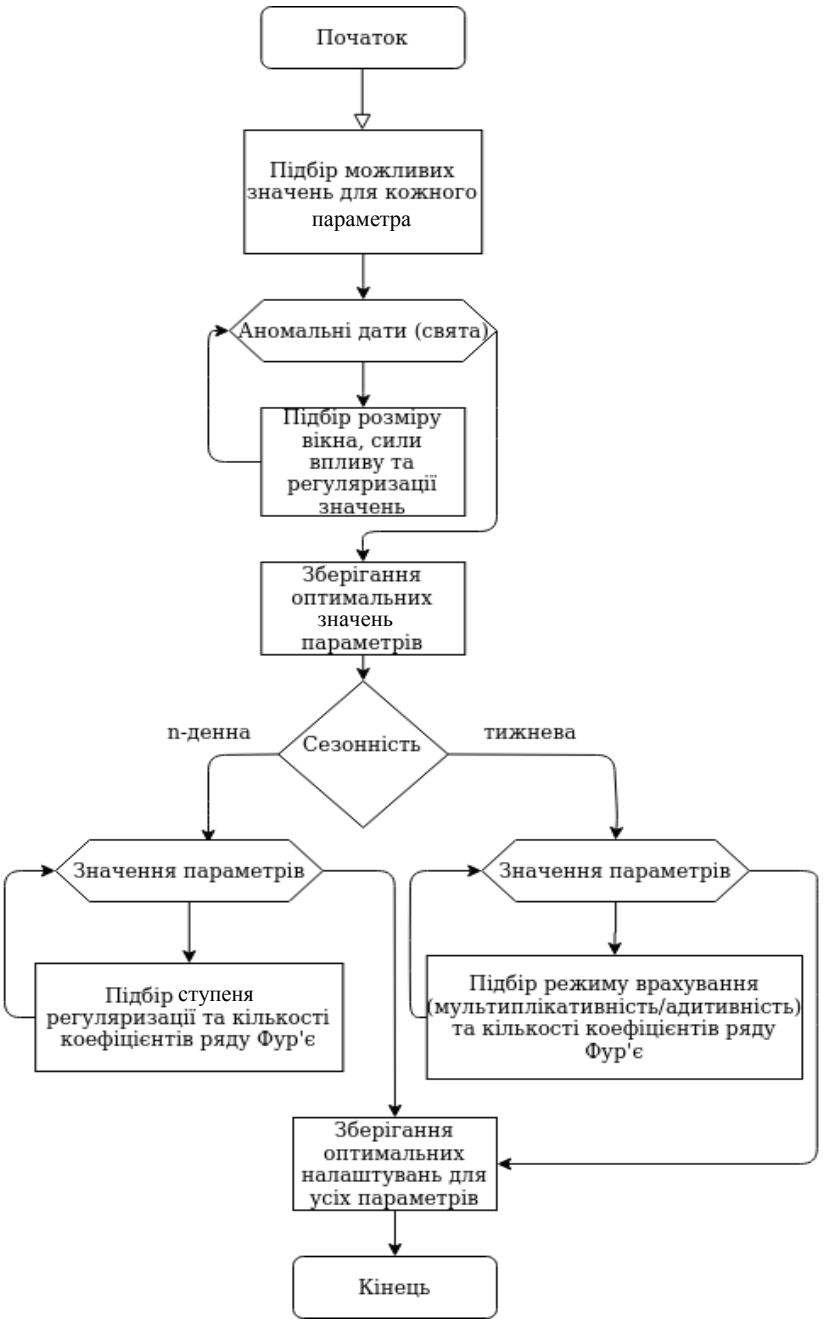


Рис. 6. Алгоритм розробленої інформаційної технології

Як виняток, може бути використана спрощена модель на основі Prophet, яка може відрізнятись від основної моделі такими спрощеннями:

– урахування тільки державних свят з семиденним зсувом та адаптивним вікном;

– урахування тільки тижневої сезонності.

Для такої спрощеної моделі достатньо лише одного етапу ідентифікації, на якому варто оптимізувати тільки такі параметри (усі складові варто одразу враховувати мультиплікативно): розмір вікна, сила впливу (масштаб) і ступінь регуляризації значень в аномальні дати (свята). Інші параметри задавати фіксованими, на основі попередніх розрахунків. Таку спрощену модель можна застосовувати одразу для досить великої кількості країн, областей та ін.

Застосування запропонованої інформаційної технології для України та аналіз результатів

Здійснено випробування запропонованої інформаційної технології для України за даними 6.07—22.11.2020 р. Дані про захворювання на коронавірус брались по API з Системи моніторингу поши-

рення епідемії коронавірусу Апарату РНБО України, решта даних — там, де рекомендовано вище.

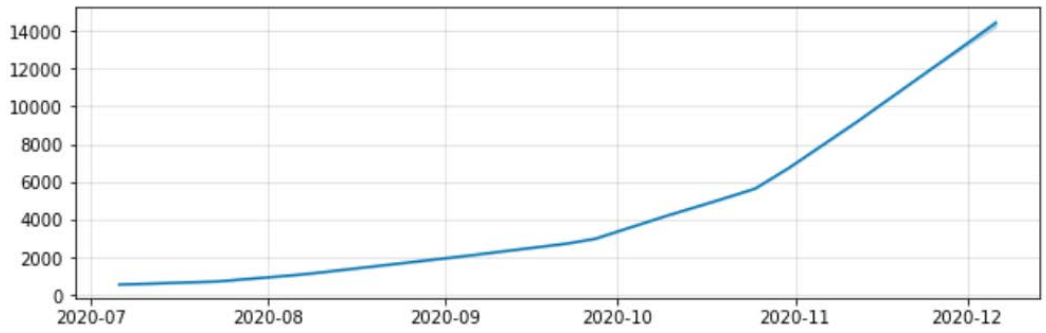
Моделювався приріст щоденних даних з урахуванням тижневої сезонності і сезонності у 4 дні (аналіз показав, що має місце певна специфіка динаміки усередині тижня і сезонність у 4 дні показала кращі результати, ніж сезонність у 3 дні).

Станом на 22.11.2020 р. оптимальною моделлю з $N = 14$ днів, тобто з прогнозом на 2 тижні вперед, є модель з такими параметрами: свята слід враховувати з вікном від 5 до 7 днів з силою впливу 2,5, регуляризацією 0,15, мультиплікативним урахуванням тижневої сезонності з регуляризацією 0,12, яка описується коефіцієнтами Фур'є порядку 8, та мультиплікативним урахуванням чотириденної сезонності з регуляризацією 0,075, яка описується коефіцієнтами Фур'є порядку 1 (рис. 7, 8) [10]. Тоді модель забезпечує сумарну відносну похибку 2,2 % за останні 14 днів і дозволяє оцінити прогноз на наступні $N = 14$ днів (табл.), за умови збереження тої самої динаміки (карантинний режим, сумарна за добу кількість тестів та ін.).

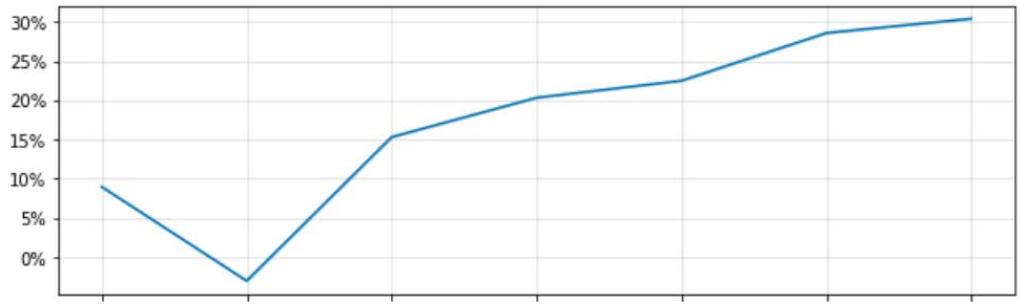
Прогноз кількості нових підтверджених випадків хворих на коронавірус в Україні за моделлю з урахуванням впливу аномальних дат з довірчим інтервалом 0,8

Дата	Нижня межа довірчого інтервалу, кількість випадків	Прогнозоване значення, кількість випадків	Верхня межа довірчого інтервалу, кількість випадків
23.11.2020	10975	11229	11464
24.11.2020	13566	13799	14033
25.11.2020	14254	14501	14728
26.11.2020	14585	14827	15047
27.11.2020	15777	16008	16257
28.11.2020	16478	16723	16967
29.11.2020	13822	14073	14324
30.11.2020	12263	12523	12775
01.12.2020	15081	15346	15598
02.12.2020	16220	16495	16779
03.12.2020	16578	16879	17176
04.12.2020	17465	17768	18055
05.12.2020	18165	18483	18826
06.12.2020	15602	15906	16230

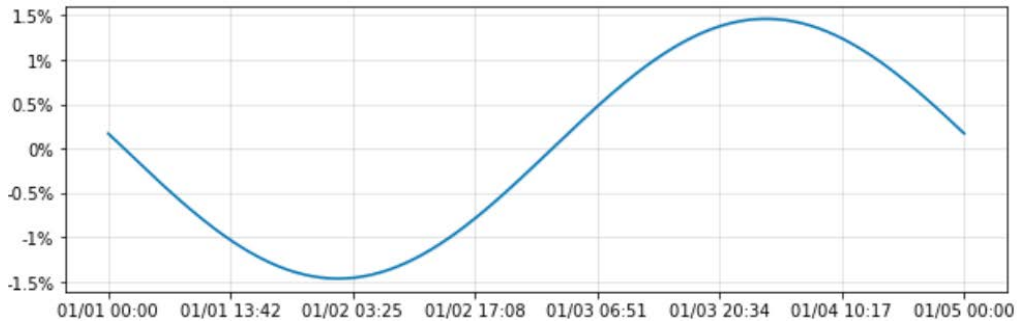
Рис. 7. Щоденна кількість нових підтверджених випадків хворих на коронавірус в Україні з 6 липня 2020 р.: чорні крапки — дані спостережень до 22.11.2020 р.; синя лінія — результат моделювання і прогнозування на 2 тижні до 6.12.2020 р. за моделлю на основі Facebook Prophet з авторським алгоритмом налаштування параметрів, з урахуванням впливу аномальних дат (відносна похибка прогнозування 2-х останніх тижнів спостережень — 2,2 %)



а)



б)



в)

Рис. 8. Складові моделі на рис. 7 для моделювання та прогнозування щоденної кількості нових підтверджених випадків хворих на коронавірус в Україні з 6 липня 2020 р.: а — основний тренд; б — тижнева сезонність; в — чотириденна сезонність

Як видно з отриманих результатів:

1. Має місце явно нелінійне зростання даних з кожним тижнем.

2. Виявлено, що, окрім тижневої сезонності, має місце ще й чотириденна сезонність, яка теж демонструє нелінійне зростання, причому, саме зростання має місце кожні 2 доби — її врахування дозволило знизити похибку з 2,38 % до 2,2 %, тобто на 7,6 %. Що означає, що має місце додаткова сукупність факторів, що прискорює вплив ще більше, ніж за кожні 7 днів.

Застосування запропонованої інформаційної технології для 70 країн світу та аналіз результатів

Здійснено випробування запропонованої інформаційної технології й для інших майже 70 країн світу, для яких є відомими державні свята у пакеті Holidays, які співавтор цієї статті професор В. Б. Мокін доповнив даними з Вікіпедії по 3-х європейських країнах (Грузія, Республіка Молдова та Албанія). З них створено новий датасет «COVID-19: Holidays of countries» (https://www.kaggle.com/vbmokin/covid19-holidays-of-countries) у платформі зі штучного інтелекту Kaggle [11]. Необхідність створення окремого датасету зумовлена тим, що технологія Prophet не допускає зсуви вікна вперед. Замість цього треба спочатку посувати саму дату, а вже потім передавати її моделі Prophet. Усі ці операції зроблені один раз для усіх 70 країн, зокрема і для України. Наявність такого датасету суттєво спрощує і прискорює застосування технології.

Для усіх 70 країн застосована спрощена модель (розрахунок з оптимізацією параметрів спро-

щеної моделі для усіх країн зайняв майже 4 години). Результат доступний у програмі-ноутбуці співавторів статті В. Б. Мокіна та А. В. Лосенка у платформі Kaggle (рис. 9, 10) [12].

Country	Country_code	Conf_real	Conf_pred	Conf_pred_h	n_h	err	err_h	lower_window	upper_window
Singapore	SG	4	0	5	13	182.759	24.1379	-2	3
Sweden	SE	7240	8423	7535	10	35.9761	14.3697	0	3
France	FR	23247	52771	22731	9	117.417	63.0913	-3	3
Spain	ES	15156	25916	21642	6	58.9786	32.7811	0	3
Israel	IL	699	447	523	38	28.9333	16.5248	-2	2
Indonesia	ID	4792	5351	4935	10	12.3761	7.16926	-1	2
Poland	PL	22464	27656	23978	9	15.8716	9.26064	0	2
Lithuania	LT	2265	2730	2567	12	38.9558	23.5408	-1	0
Iceland	IS	20	0	4	11	64.1975	44.4444	-1	2
Italy	IT	37239	45247	44040	8	12.5679	8.76904	-1	2

Рис. 9. 10 з 70 країн датасету «COVID-19: Holidays of countries» з найбільшим зменшенням відносної похибки прогнозування останніх спостережуваних 2-х тижнів за урахування свят від 30,2 % до 86,8 %

Country	Country_code	Conf_real	Conf_pred	Conf_pred_h	n_h	err	err_h	lower_window	upper_window
Russian Federation	RU	24059	23958	24045	7	2.03324	2.09732	-2	0
Ukraine	UA	14834	13911	13904	8	2.57519	2.55798	-1	0
Argentina	AR	9608	10115	9890	16	4.46349	3.61771	0	0
Philippines	PH	1637	1509	1512	8	5.75623	5.45683	-3	1
United States	US	195542	215077	214122	8	6.82466	6.80012	0	2
Estonia	EE	361	368	368	8	7.31328	7.10581	-1	3
Indonesia	ID	4792	5351	4935	10	12.3761	7.16926	-1	2
Croatia	HR	2958	3201	3198	10	7.47522	7.41363	0	0
Bulgaria	BG	3983	4410	4254	11	7.21434	7.78928	0	1
Moldova	MD	1422	1401	1413	9	8.09025	7.81601	-1	3

Рис. 10. 10 із 70 країн датасету «COVID-19: Holidays of countries» з найменшим значенням відносної похибки прогнозування останніх спостережуваних 2-х тижнів за урахування свят у спрощеній моделі

Аналіз результатів показав таке:

1. Для України вплив свят не є досить впливовим, у порівнянні, наприклад з такими країнами Швеція, Франція, Іспанія, де модель з урахуванням свят і псевдосвят має похибку майже удвічі меншу, ніж модель без урахування таких аномальних дат.

2. Щодо України спрощена модель з урахуванням аномальних дат дає похибку 2,56 %, а без їх урахування — 2,58 %. Однак, ефективніша модель дає похибку 2,2 % проти 2,48 %, тобто врахування таких часових аномалій дає зменшення похибки на на 11 %.

Висновки

В статті розроблено інформаційну технологію аналізу та прогнозування кількості нових підтверджених випадків захворювання на коронавірус «COVID-19», викликаного інфекцією SARS-CoV-2, на прикладі щодобових сумарних по Україні даних поточної «хвилі» з урахуванням різних свят і псевдосвят, які можуть мати аномальний вплив. Проведено огляд відомих моделей для врахування таких аномалій та обґрунтовано, що оптимальною для розв'язання цієї задачі є модель Facebook Prophet. Охарактеризовано наявні дані щодо можливих часових аномалій по Україні у відомому датасеті Google-платформи «COVID-19 Open Data» та запропоновано яким чином можна адаптивно враховувати такі аномалії, як: державні свята (окремо як час підвищеного ризику контактів людей та як час, коли частина лабораторій не працюють, через що, має місце аномально низька кількість проаналізованих за добу тестів), «метеопаттерни» — дати, коли за даними NOAA було дуже тепло і не було опадів та дати послаблення карантину за інформацією з «Oxford COVID-19 Government Response Tracker».

Розроблено алгоритм застосування запропонованої інформаційної технології з двоетапною ідентифікацією параметрів та окремим валідаційним датасетом для ідентифікації оптимальної структури моделі на кожному з цих етапів. Створено програмне забезпечення на Python на базі платформи Kaggle, яке застосовано, як для України, так і ще для 69 країн світу. Для прискорення роботи, по-перше, розроблено спрощену версію моделі лише з одним етапом її ідентифікації, а по-друге, створено новий датасет «COVID-19: Holidays of countries» з інформацією про свята 70 країн світу, адаптований до потреб цієї технології, розміщений у Kaggle у форматі відкритих даних.

Ці ідентифіковані моделі дали можливість відповісти на такі важливі питання щодо щодобової кількості підтверджених випадків захворювань на коронавірус в Україні з липня по листопад 2020 року:

1. Чи має зростання лінійний чи нелінійний характер? — має місце нелінійний характер, впливи свят та усі сезонні складові варто враховувати мультиплікативно.

2. Чи мають значний вплив з певним зсувом в часі аномальні дати (державні свята та псевдосвята на кшталт дат послаблення карантину, теплих днів без опадів тощо)? — Як для України, то їх урахування дає зменшення похибки на 11 %, а для інших країн світу, наприклад, Швеції, Франції, Іспанії, має місце зменшення похибки на 50 % і більше.

3. Чи має місце вплив інших видів сезонності, окрім тижневої, і, якщо так, то з якою періодичністю та лінійно чи нелінійно? — має місце нелінійна чотириденна сезонність з тригонометричним зростанням кожні 2 доби, врахування якої дозволяє зменшити похибку на 7,6 %.

Той факт, що в моделях не здійснювався повний перебір усіх можливих значень параметрів, не враховувалась явно динаміка інших факторів (наприклад, наростання 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, яка 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. Членом цієї групи є один зі співавторів статті доктор технічних наук, професор В. Б. Мокін.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

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

[2] C. L. Althaus, *(2020) Real-time modeling and projections of the COVID-19 epidemic in Switzerland, Institute of Social and Preventive Medicine*, University of Bern, Switzerland 20 April 2020, [Electronic resource]. Available: https: ispmbern.github.io/covid-19/swiss-epidemic-model .

[3] *Прогноз розвитку епідемії COVID-19 в Україні на 6–13 листопада 2020 року («Прогноз РГ-27»).* [Електронний ресурс]. Режим доступу: http://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=7129 , дата звернення: лист., 6. 2020.

[4] *Прогноз розвитку епідемії COVID-19 в Україні на 13–20 листопада 2020 року («Прогноз РГ-28»).* [Електронний ресурс]. Режим доступу: http://www.nas.gov.ua/UA/Messages/Pages/View.aspx?MessageID=7155 , дата звернення: лист., 14. 2020.

[5] V. B. Mokin, *Total Ranking of all participants of COVID19 Global Forecasting Challenges — версія ноутбука – 12.06.2020 р.* [Електронний ресурс]. Режим доступу: https://www.kaggle.com/vbmokin/all-ranking-covid19-global-forecasting-challenges , дата звернення: червень, 12. 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. ISSN No. 2455-2143.

[7] 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. [Electronic resource]. Available: https://doi.org/10.1016/j.chaos.2020.110058 .

[8] 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.

[9] В. Б. Мокін, О. В. Слободянюк, О. М. Давидюк, і Д. О. Шмундяк, «Інформаційна технологія пошуку можливих джерел підвищеного забруднення річки з використанням моделі Prophet,» *Вісник Вінницького політехнічного інституту*, № 4, с. 15-24, 2020. https://doi.org/10.31649/1997-9266-2020-151-4-15-24.

[10] *COVID in UA: Prophet with 4, 7d seasonality* — версія ноутбука — 22.11.2020 р. [Електронний ресурс]. Режим доступу: https://www.kaggle.com/vbmokin/covid-in-ua-prophet-with-4-7d-seasonality/output?scriptVersionId=47484394, дата звернення: лист., 22. 2020.

[11] *COVID-19: Holidays of countries* — версія датасета — 21.11.2020 р. [Електронний ресурс]. Режим доступу: https://www.kaggle.com/vbmokin/covid19-holidays-of-countries , дата звернення: лист., 21. 2020.

[12] *COVID-19 in 70 countries: daily Prophet forecast* — версія ноутбука — 21.11.2020 р. [Електронний ресурс]. Режим доступу: https://www.kaggle.com/vbmokin/covid-19-in-70-countries-daily-prophet-forecast?scriptVersionId=47433942, дата звернення: лист., 21. 2020.

Рекомендована кафедрою системного аналізу та інформаційних технологій ВНТУ

Стаття надійшла до редакції 23.11.2020

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, викладач кафедри системного аналізу та інформаційних технологій, e-mail: yasholt@gmail.com

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

Informational Technology of Analysis and Forecasting of Number of New Cases of Coronavirus SARS-Cov-2 in Ukraine Based on the Prophet Model

[1]Vinnytsia National Technical University

The article describes the development of information technology for analysis and forecasting of amount of new confirmed cases of the disease for coronavirus "COVID-19" caused by SARS-CoV-2 infection, based on the daily summary data of the current "wave" in Ukraine, and taking into account various holidays and pseudo-holidays. A review of the known models, which acknowledge such anomalies, was conducted and it is substantiated that considering the current short series of observational data and other conditions, the Facebook Prophet model is optimal for solving this problem. Available data on possible time anomalies in Ukraine in the well-known dataset "COVID-19 Open Data" from Google was characterized, and it is proposed how to take into consideration such anomalies as: public holidays, dates when accordingly to NOAA data weather was warm and without any precipitation, and dates of quarantine easing using information from the "Oxford COVID-19 government response tracker". An algorithm for usage of the proposed information technology was developed, which included a step of two-stage parameter identification and used a separate validation dataset to identify the optimal structure of the model at each stage. Software using Python was created and displayed on Kaggle platform, which then was applied both for Ukraine and for 69 countries around the world. To speed up the research firstly the simplified version of the model was developed with only one stage of parameter identification, and secondly that a dataset "COVID-19: Holidays of countries" was compiled, with information about the holidays of 70 countries, adapted to the needs of this technology and was saved on Kaggle as an open dataset. With the help of the identified models, a number of important conclusions were obtained regarding the understanding of the patterns of coronavirus spread both in Ukraine and in 69 other countries of the world. A model was built to calculate the number of possible new confirmed cases of coronavirus in Ukraine for the next 2 weeks with an error of 2,2 % and using this model, a forecast for the next 2 weeks was made, which was submitted to the Research Group of Mathematical Modeling of Problems Related to the SARS-CoV-2 Epidemic in Ukraine.

Keywords: information technology, SARS-CoV-2, COVID-19, time series forecasting, Prophet, artificial intelligence.

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, Lecturer 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 прогнозирование количества новых случаев болезни, вызванной коронавирусом SARS-COV-2, в Украине на основе модели Prophet

Разработана информационная технология анализа и прогнозирования количества новых подтвержденных случаев заболевания на коронавирус «COVID-19», вызванного инфекцией SARS-CoV-2, на примере ежесуточных суммарных по Украине данных текущей «волны» с учетом различных праздников и псевдо праздников, которые могут иметь аномальное влияние. Проведен обзор известных моделей для учета таких аномалий и обосновано, что при современных коротких рядах данных наблюдений и других условий оптимальной для решения этой задачи является модель Facebook Prophet. Охарактеризованы имеющиеся данные о возможных временных аномалиях по Украине в известном датасете Google-платформы «COVID-19 Open Data» и предложено каким образом можно адаптивно учитывать такие аномалии, как: государственные праздники, даты, когда по данным NOAA было очень тепло и не было осадков и даты послабления карантина по информации из «Oxford COVID-19 government response tracker». Разработан алгоритм применения предложенной информационной технологии с двухэтапной идентификацией параметров и отдельным валидационным датасетом для идентификации оптимальной структуры модели на каждом из этих этапов. Создано программное обеспечение на Python на базе платформы Kaggle, которое применено, как для Украины, так и еще для 69 стран мира. Для ускорения работы, во-первых, разработана упрощенная версия модели только с одним этапом ее идентификации, а во-вторых, создан новый датасет «COVID-19: Holidays of countries» с информацией о праздниках 70 стран мира, адаптированный к потребностям этой технологии и расположенный в Kaggle в формате открытых данных. С помощью идентифицированных моделей получен ряд важных выводов относительно понимания закономерностей распространения коронавируса как в Украине, так и в других 69 странах мира. Построена модель, которая обеспечивает моделирование количества новых подтвержденных случаев заболевания коронавирусом в Украине на 2 недели вперед с погрешностью 2,2 % и сделан прогноз на следующие 2 недели, который передан в Рабочую группу по математическому моделированию проблем, связанных с эпидемией коронавируса SARS-CoV-2 в Украине.

Ключевые слова: информационная технология, SARS-CoV-2, COVID-19, прогнозирование временных рядов, Prophet, искусственный интеллект.

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, преподаватель кафедры системного анализа и информационных технологий, e-mail: yasholt@gmail.com